OXFORD

Briefings in Bioinformatics, 2023, **24(6)**, 1–17
https://doi.org/10.1093/bib/bbad382
Problem Solving Protocol

Filter and Wrapper Stacking Ensemble (FWSE): a robust approach for reliable biomarker discovery in high-dimensional omics data

Sugam Budhraja, Maryam Doborjeh, Balkaran Singh, Samuel Tan, Zohreh Doborjeh, Edmund Lai, Alexander Merkin, Jimmy Lee, Wilson Goh and Nikola Kasabov

Corresponding author: Sugam Budhraja, Knowledge Engineering and Discovery Research Innovation (KEDRI), School of Engineering Computer and Mathematical Sciences, Auckland University of Technology, 55 Wellesley Street East, 1010 Auckland, New Zealand. sugam.budhraja@autuni.ac.nz

Abstract

Selecting informative features, such as accurate biomarkers for disease diagnosis, prognosis and response to treatment, is an essential task in the field of bioinformatics. Medical data often contain thousands of features and identifying potential biomarkers is challenging due to small number of samples in the data, method dependence and non-reproducibility. This paper proposes a novel ensemble feature selection method, named Filter and Wrapper Stacking Ensemble (FWSE), to identify reproducible biomarkers from high-dimensional omics data. In FWSE, filter feature selection methods are run on numerous subsets of the data to eliminate irrelevant features, and then wrapper feature selection methods are applied to rank the top features. The method was validated on four high-dimensional medical datasets related to mental illnesses and cancer. The results indicate that the features selected by FWSE are stable and statistically more significant than the ones obtained by existing methods while also demonstrating biological relevance. Furthermore, FWSE is a generic method, applicable to various high-dimensional datasets in the fields of machine intelligence and bioinformatics.

Keywords: feature selection; biomarker discovery; ensemble learning; high-dimensional data; genomics; proteomics.

INTRODUCTION

A biomarker, short for 'biological marker', is a biomedical indicator that provides insight into the medical state of a patient and its measurement is reliable and reproducible [1]. Biomarkers have been widely used for disease detection, prevention, analysis of response to treatments, evaluating safety or toxicity of

Sugam Budhraja is a PhD student at Auckland University of Technology, New Zealand. His background is in machine learning and software development. His research fields include Neuroinformatics, Deep Learning, Self-Supervised Learning and Echo State Networks.
Maryam Doborjeh received her PhD in Computer Science from Auckland University of Technology, New Zealand. She is currently a senior lecturer in the School of Engineering, Computer and Mathematical Sciences at Auckland University of Technology, New Zealand. Her research fields are Neuroinformatics, Spiking Neural Networks, Machine Learning and Brain Data Analysis.
Balkaran Singh is a PhD student at Auckland University of Technology, New Zealand. His background is in computer science and applied statistics. His research fields are Optimisation in Neural Networks, Continual Learning, Meta Learning and Spiking Neural Networks.
Samuel Tan is a PhD student at Nanyang Technological University, Singapore. His background is in biological sciences and statistics. His research fields include Bioinformatics, Network Theory and Neighbourhood Optimization.
Zohreh Doborjeh received her Ph.D. in Computational Cognitive Neuroscience from Auckland University of Technology, New Zealand. She is currently a post-doctoral fellow at the Centre for Brain Research at the University of Auckland, New Zealand, and a lecturer at the School of Psychology at the University of Waikato, New Zealand. Her research fields are Neuroinformatics, Neuropsychology, Cognitive Neuroscience and Artificial Intelligence.
Edmund Lai received his PhD in Electrical Engineering from The University of Western Australia. He is currently a professor of Information Engineering in the School of Engineering, Computer and Mathematical Sciences at the Auckland University of Technology, New Zealand. His research interests are Digital Signal Processing, Computational Intelligence, Multi-Agent Dynamical Systems and Optimization.
Alexander Merkin received his PhD in Psychiatry from the Serbsky Research Centre for Social and Forensic Psychiatry, Russia. He is currently a research fellow at the Institute for Stroke and Applied Neurosciences, AUT University and a lecturer at the Department of Psychotherapy and Counselling, AUT University. His research interests include digital mental health, Artificial Intelligence, Psychogeriatrics, Psychiatry and Cognitive Neuroscience.
Jimmy Lee received his basic medical degree from the National University of Singapore. He is a psychiatrist and clinician scientist at the Institute of Mental Health, Singapore, and an Associate Professor at the Lee Kong Chian School of Medicine at the Nanyang Technological University. His research areas are in Psychiatry, Psychopharmacology, Schizophrenia and AI-based Health Technologies.
Wilson Goh received his PhD in Bioinformatics and Computational Systems Biology from the Imperial College London, United Kingdom. He is currently an assistant professor of Biomedical Informatics at Lee Kong Chian School of Medicine at Nanyang Technological University, Singapore. His research areas are Complex systems, Bioinformatics, Computational Biology, Proteomics and Genomics.
Nikola Kasabov received his PhD from the Technical University of Sofia, Bulgaria. He is the Founding Director of KEDRI and Professor of Knowledge Engineering in the School of Engineering, Computing and Mathematical Sciences at Auckland University of Technology, New Zealand. He holds Professorial positions at the University of Ulster UK, IICT Bulgarian Academy of Sciences and Dalian University, China. His research areas are Computational Intelligence, Neuroinformatics, Knowledge Discovery and Spiking Neural Networks, with more than 700 publications.
Received: February 09, 2023. **Revised:** September 18, 2023. **Accepted:** October 3, 2023

treatments, assessing risk and susceptibility in individuals, and monitoring disease progress [2]. Development of gene expression profiling techniques such as Next-Gen Sequencing [3], DNA microarray [4] and RNA sequencing (RNA-Seq) [5] have opened new avenues for data-driven discovery of genetic biomarkers for various diseases [6].

In the realm of machine learning, the problem of finding top genetic biomarkers for a disease translates to selecting top features (genes) from the data that discriminate well between the case and control groups of the disease [7]. The minimal set of top genes that maximize performance is referred to as 'signature'. The performance of the selected top genes is often measured using the accuracy of classification between case and control groups since true biomarkers capture characteristics of either the case or control group and should thus be useful in classifying between the groups. The selection of top biomarkers from high-dimensional omics data is an area of broad and current interest [8].

Feature selection on omics datasets is difficult due to high dimensionality (often tens of thousands of features) and relatively small number of samples [9]. In addition, feature selection on omics data suffers from the issues of method-dependence, meaning different top features are obtained depending on the feature selection method used; and feature instability, meaning the same feature selection algorithm does not consistently select the same top features across different perturbations of the data [10]. The non-reproducibility of same top genes across different perturbations of datasets is a serious impediment to clinical applications of the selected markers [11]. Hence, to improve the quality of the selected feature subset, both the accuracy and stability of the feature selection methods need to be analysed parallelly [12].

Differential gene expression (DGE) analysis is the most popular approach to identify top markers from omics data due to its simplicity and interpretability [13]. DEGs are identified based on the magnitude of difference between the mean expression values across different classes and the variance of the values within each class. Commonly used differential analysis tests such as Student's *t*-test, make assumptions that the underlying distribution of data is Gaussian. However, this is seldom the case for gene expression data [14]. The selection of appropriate *P*-value threshold for these tests also affects the interpretability of the results [15] and in recent times, disputes over the misuse of *P*-values have also arisen [16, 17].

More recently, traditional machine learning feature selection approaches have been applied to omics data for the discovery of potential biomarkers. Compared with statistical tests, these approaches use fewer assumptions about the distribution of data. Many of these approaches such as Recursive Feature Elimination (RFE) are able to analyse the predictive power of features as a group rather than individual strength. This leads to the selection of a set of top genes that may be poor biomarkers individually but are strong biomarkers when combined as a group. Examples of the application of traditional machine learning methods in biomarker discovery include the use of elastic net for diagnosing papillary thyroid carcinoma [18], Random Forest for tracking prostate cancer progression [19], and Lasso regression and RFE using Support Vector Machines (RFE-SVM) for identifying therapeutic targets in ferroptosis from coronary artery disease [20] and validating biomarkers for Alzheimer's Disease [21].

The performance of a given feature selection method varies a lot across different datasets. Previous studies [22–24] have shown that ensemble feature selection is an effective approach to overcome this limitation. Integrating results from multiple biomarker discovery methods increases the likelihood that the strengths of at least one method will be well-suited to the characteristics of the dataset, thereby enhancing overall performance. Ensemble feature selection techniques have been successfully employed in various contexts for biomarker discovery, such as VSOLassoBag [25], a bagging ensemble of Lasso regressors, which has been used for the diagnosis and prognosis of breast cancer. Similarly, RFE using a bagging ensemble of SVMs has been applied to the diagnosis of childhood leukemia and colon cancer [26]. Furthermore, a voting ensemble combining a correlation method, a causal inference method and a regression method using Borda count have been used to predict miRNA targets for hepatocellular carcinoma [27].

In this study, we conduct a comparative analysis of popular feature selection methods and examine their combinations using straightforward ensembling techniques. Based on the results of this analysis, we propose a novel ensemble feature selection architecture, termed FWSE (Filter and Wrapper Stacking Ensemble), pronounced 'fuse'. In addition to the ensembling analysis, we evaluate the performance of FWSE against various existing feature selection techniques for the task of biomarker discovery on four high-dimensional genomic and proteomic datasets, two of which are related to mental illnesses and two to cancer. Our study reveals that FWSE is capable of achieving the optimal combination of accuracy and stability on multiple datasets when compared with existing popular feature selection methods.

These findings suggest that our proposed ensemble feature selection architecture can serve as a potent tool for biomarker discovery in high-dimensional omics data, potentially leading to significant advancements in personalized medicine and disease prevention. We believe that our work contributes to the ongoing efforts in the field of machine learning and bioinformatics to develop robust and reliable methods for feature selection in high-dimensional data.

In the following sections, we provide an introduction to feature selection and present a comprehensive overview of the methods and datasets used in our study. We then present the results of our ensembling analysis, which led to the development of FWSE. Next, we conduct a comparative performance analysis of FWSE against established approaches in the field. Finally, we discuss the implications of our findings and suggest potential directions for future research.

METHODOLOGY

This section is structured as follows: we first introduce the concept of feature selection and its significance in the realm of machine learning and bioinformatics. We then present a detailed discussion of the feature selection methods that were compared in this study, followed by an explanation of the ensembling techniques that were employed. Subsequently, we introduce the novel FWSE architecture and provide a comprehensive overview of its design and functionality. We then describe the experimental setup that was used to analyse the performance of each feature selection method, including the metrics used to evaluate their performance. Finally, we present the datasets that were used in this study, detailing their source, composition and relevance to the task at hand.

Feature selection

Feature selection is a crucial process in machine learning that involves identifying the most significant features for predicting the outcome of a classification or regression task. By eliminating non-informative and irrelevant features, the performance

of machine learning models can be enhanced, as decisions are based solely on task-relevant features. However, feature selection, particularly in molecular data, can be challenging due to high dimensionality (large number of features) and a small number of samples. Additionally, redundancy, where different subsets of features can have the same predictive power, adds to the complexity [28, 29]. Under different perturbations of data, a feature selection algorithm might produce completely different sets of top features that may be equivalent in terms of classification performance [30]. Hence, when trying to identify biomarkers, it is essential to use a reliable method that produces consistent features that are significant across all samples [31].

Feature selection methods can be broadly categorized into supervised, unsupervised and semi-supervised methods [32]. Supervised feature selection is used for classification or regression tasks, selecting features that can discriminate between classes or approximate regression targets. Unsupervised feature selection, designed for clustering problems, seeks alternative criteria to define feature relevance without label information. Semi-supervised feature selection bridges the gap between supervised and unsupervised methods, leveraging limited supervision information to guide the feature selection process. In this study, we employ supervised feature selection methods for the task of biomarker discovery.

From a selection strategy perspective, feature selection methods can be classified into filter, wrapper and embedded methods [33]. Filter methods assess feature importance based on data characteristics independently of any learning algorithms. Wrapper methods use an external learning algorithm to evaluate the quality of selected features, iterating until certain criteria are met. Embedded methods offer a balance between filter and wrapper methods by incorporating feature selection into model learning. Some ensemble approaches combine multiple feature selection techniques from different categories, forming a fourth category known as 'integrated' methods. In the first part of this study, we compared some traditional feature selection methods and their simple ensembles on the LYRIKS data. These are described below.

F-statistic

The F-statistic of a feature is the result of an Analysis of Variance (ANOVA) *F*-test for that feature across target classes. Mathematically, it is computed as the ratio of the between-group variance to the within-group variance, where the groups are the target classes. The F-statistic is a powerful measure of the statistical significance of the group differences, but it assumes that the data are normally distributed and the variances of the groups are equal, which may not always be the case in real-world data.

Signal-to-noise ratio

The signal-to-noise ratio (SNR), similar to a Student's *t*-test, is a measure of the separation between the means of two classes relative to the variability within each class. It is calculated as the Euclidean distance between the class means divided by the sum of the standard deviations of each class [34]. The SNR is a robust measure of feature importance, but it assumes that the features are independent, which may not be the case in high-dimensional omics data. Equation 1 shows the formula for the calculation of SNR for the ith feature:

$$SNR_i = \frac{|M_i^{(\text{class 1})} - M_i^{(\text{class 2})}|}{Std_i^{(\text{class 1})} + Std_i^{(\text{class 2})}} \quad (1)$$

Lasso Regression

Logistic Regression is a method in which a linear decision boundary is learned to separate between classes. The coefficients in the equation that defines this boundary can be treated as feature importance. Lasso Regression is a variant of logistic regression that uses the L1 norm for penalization, leading to sparse solutions where many of the estimated coefficients are zero. This results in a smaller set of features being chosen, making Lasso Regression a powerful tool for feature selection [35]. However, Lasso Regression can be sensitive to outliers and may not perform well when the number of features is much greater than the number of samples.

Random Forest

In decision trees, nodes with the greatest decrease in impurity occur at the start, while nodes with the least decrease in impurity occur at the end. By pruning the trees below a particular node, a subset of the most important features can be created. In this study, a bagging ensemble of decision trees also known as Random Forest [36] has been used for the purposes of feature selection. Random Forest is a robust and versatile method, but it can be computationally intensive, especially with high-dimensional data.

Recursive Feature Elimination

In Recursive Feature Elimination (RFE), an external estimator is used to assign importance to the features. Based on the feature importances, a defined number of the least important features are removed. In the next step, the external estimator is used to rank the remaining features, and the least important features are removed. This process continues till the required number of features are left. In this work, RFE has been used with Linear Support Vector Machines (SVMs), Logistic Regression and Random Forests as the base models. RFE-SVM is an established feature selection method to find statistically significant features in high-dimensional gene datasets [37].

In addition to the aforementioned feature selection approaches used in traditional machine learning, the proposed FWSE is also compared with popular ensemble feature selection methods that have been used for biomarker discovery, described below.

Variable-Selection Oriented Lasso Bagging

Variable-Selection Oriented Lasso Bagging (VSOLassoBag) is an ensemble feature selection method that combines multiple Lasso regression models in a bagging ensemble. It creates multiple data subsets, filters out features based on their correlation with the outcome, and uses cross-validated Lasso regression to further rank the remaining features. Features with non-zero coefficients are considered 'selected'. The process is repeated across all data subsets, and features that are frequently selected are considered more important [25].

Multi-Criterion Fusion-based Recursive Feature Elimination

Multi-Criterion Fusion-based Recursive Feature Elimination (MCF-RFE) is an ensemble wrapper feature selection method, wherein features are recursively eliminated based on a voting ensemble of multiple feature selection methods [38, 39]. MCF-RFE leverages the strengths of multiple feature selection methods, but it can be computationally intensive and may underperform if the base feature selection methods are not well-suited to the data.

Table 1: Summary of feature selection techniques used in this work

Abbreviation	Method name	Type	Time complexity	Supervised
ANOVA	F-Statistic	Filter	Low	Yes
SNR	Signal-to-Noise Ratio	Filter	Low	Yes
Lasso	Lasso Regression	Embedded	Normal	Yes
RF	Random Forest	Embedded	Normal	Yes
RFE-SVM	Recursive Feature Elimination using Support Vector Machines	Wrapper	High	Yes
RFE-LR	Recursive Feature Elimination using Logistic Regression	Wrapper	High	Yes
RFE-RF	Recursive Feature Elimination using Random Forests	Wrapper	High	Yes
VSOLassoBag	Variable-Selection Oriented Lasso Bagging	Integrated	Normal	Yes
MCF-RFE	Multi-Criterion based Recursive Feature Elimination	Integrated	High	Yes
ESVM-RFE	Ensemble Support Vector Machine based Recursive Feature Elimination	Integrated	High	Yes
E-Borda	Voting Ensemble using Borda Count	Integrated	High	Yes

Ensemble Support Vector Machine based Recursive Feature Elimination

Ensemble Support Vector Machine based Recursive Feature Elimination (ESVM-RFE) is a variant of RFE that uses a bagging ensemble of SVM classifiers instead of a single SVM classifier to rank the features at every step of elimination [26].

Voting Ensemble using Borda Count

In Voting Ensemble using Borda Count (E-Borda), to overcome the problem of different feature selection methods being optimal for different datasets or tasks, a voting ensemble of multiple feature selection algorithms is taken [27, 40]. The aggregation of the feature rankings from the multiple algorithms is done using Borda Count. But like MCF-RFE, it can be sensitive to the choice of base feature selection methods.

By comparing FWSE with these traditional and ensemble feature selection methods, we aim to demonstrate the effectiveness and robustness of our proposed method in identifying reliable biomarkers from high-dimensional omics data. A summary of all the feature selection methods used in this study is provided in Table 1.

Feature stability

Feature stability measures the consistency of the feature rankings produced by each feature selection method, under different perturbations of the data. For this study, the Jaccard Index has been used to analyse the stability:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

$$Stability = \sum_{i=1}^{N} \sum_{j=i+1}^{N} Jaccard(R_i, R_j) \tag{3}$$

where N is the total number of rankings being compared and R_i represents the ith ranking being compared.

The objective of this study is to measure the consistency of only the top features. Hence, Jaccard Index has been used to compare the similarity between two top n feature subsets. The Jaccard Index does not account for the ranking of features within the subset, hence, the stability is calculated repeatedly over varying values of n. Metrics such as Kendall's Tau [41] and Spearman's Rho [42] were not used because they utilize ranking of all the features to compute stability, and the ranking of irrelevant features may negatively affect the final scores.

Ensemble Learning

Ensemble Learning or ensembling is a way of combining multiple models, based on the idea that a group of diverse models would perform better than a single model [43]. In the first part of this study, three popular ensembling techniques were explored to combine results from multiple feature selection methods. These are illustrated in Figure 1(A) and discussed below.

Voting

Voting is a simple ensemble approach in which results from different methods, trained on the same data, are aggregated based on majority or by taking the average. The aggregation can also be weighted, where the output of each method is multiplied by a weight before combining. The approach is illustrated in Figure 1(A). For the purposes of this study, rank aggregation [44] has been used to combine the multiple feature rankings. In rank aggregation, the feature importance results from the methods are converted into feature rankings, where a smaller rank is allotted to a feature with higher importance. Then all the feature rankings are simply added up and the resulting array is sorted to obtain the final rankings. The features with smaller values in the summed vector end up with smaller ranks after sorting and are more important.

Bagging

Bagging, short for bootstrap aggregation, is an ensemble learning method that involves training multiple instances of the same model, on different subsets of the training data. There exist many approaches for creating the subsets, which involve choosing a subset of samples at random, and may also involve choosing a subset of features at random [45–48]. However, for this study, the original approach of creating bootstrap samples has been used, which involves choosing random samples with replacement, to form the subsets [49]. This approach has been shown to decrease the variance of the model while maintaining bias [50], thus reducing overfitting. The approach is illustrated in Figure 1(A).

Stacking

Stacking is a multi-layer approach in which outputs of the methods in the previous layer act as input to the methods in the next layer. The methods in the first layer are trained on the data whereas the methods in all the succeeding layers are trained

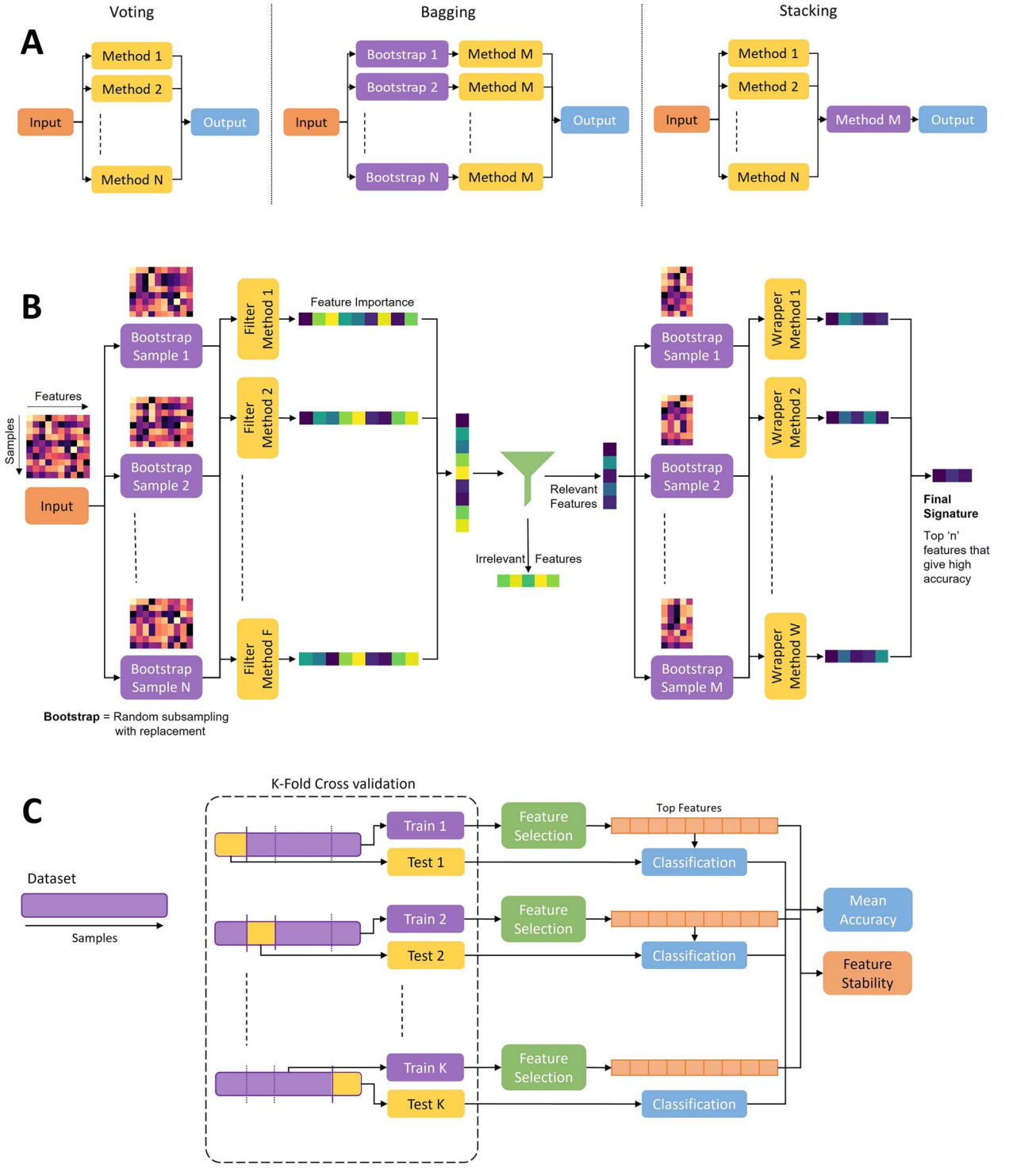

Figure 1. (**A**) The various ensemble approaches explored in this study, namely voting, bagging and stacking. (**B**) The architecture of FWSE. (**C**) Evaluation of a feature selection method on a dataset.

on the outputs of the methods in the previous layer. Figure1(A) illustrates one such stack. Stacking has been shown to outperform any of the single models used in the stack. [51].

Filter and Wrapper Stacking Ensemble

Based on the results of the ensembling experiments (shown in Results section), a new ensemble architecture is proposed to achieve an optimal combination of accuracy and stability. This architecture is illustrated in Figure 1(B). Initially, multiple bootstrap samples are generated from the dataset. Filter feature selection methods are then applied to these samples, and their outputs are combined using rank aggregation. The 'pruning factor,' a parameter ranging from 0 to 1, governs the elimination of the least significant features. In the scope of this study, this factor is set at 0.5, leading to the removal of 50% of features in all analyses presented. Subsequently, new bootstrap samples are generated using the pruned feature set. Wrapper methods are then applied to further rank the remaining features. The final feature ranking is obtained through aggregation of the results from these wrapper methods. A comprehensive analysis exploring

Table 2: Summary of datasets used in the study

Dataset	Profiling technique	# of features	# of samples	Clinical factor	Case	Control	Reference
LYRIKS	Microarray	34928	84	Total	56 (66.7%)	28 (33.3%)	Lee *et al.* [53]
				Age	22.1	22.5	
				Female sex	14 (25.0%)	7 (25.0%)	
Bipolar	RNASeq	20581	480	Total	240 (50%)	240 (50%)	Krebs *et al.* [54]
				Age	50.3	43.4	https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE124326&format=file&file=GSE124326_count_matrix.txt.gz
				Female sex	131 (54.6%)	119 (49.6%)	
LUAD	RNASeq	20531	569	Total	510 (89.6%)	59 (10.4%)	Liu *et al.* [55]
				Age	65.3	66.0	https://cbioportal-datahub.s3.amazonaws.com/luad_tcga_pan_can_atlas_2018.tar.gz
				Female sex	274 (53.9%)	32 (56.1%)	
PDAC	Proteome	11662	215	Total	140 (65.1%)	75 (34.9%)	Cao *et al.* [56]
				Age	64.3	64.5	https://www.linkedomics.org/data_download/CPTAC-PDAC/
				Female sex	66 (47.1%)	33 (49.2%)	

the impact of varying the pruning factor on model accuracy and stability is provided in Appendix B.

Study design

This study consists of two parts, the first part compares basic feature selection techniques and analyses the effect of ensembling these techniques. Previous studies [11] have shown that ensembling is an effective approach to deal with highly correlated features. Based on results from this part, FWSE is proposed as a novel ensemble feature selection algorithm. The second part of this study compares the performance of the proposed FWSE method with the traditional feature selection techniques and also popular ensemble feature selection approaches.

To evaluate the performance of every feature selection method, the method is first run on the training data to obtain a set of top features. Then different classifiers are trained using the features selected by the methods. The accuracy of these classifiers is evaluated on the test set. Since the data contain a small number of samples, K-fold cross validation is used to analyse model performance [52]. In K-fold cross validation, the whole dataset is split into K equal parts (called folds). In each iteration, one of the folds is treated as the test set and the other K-1 folds are treated as the training set. This process is repeated K times such that each one of the K folds gets to be a test set once.

Since, in each fold, the training sets are different, the feature selection algorithms end up producing different top features each time. To evaluate the reproducibility of the features selected by these algorithms, the consistency of the top features across the folds is analysed. This process is illustrated in Figure 1(C).

Datasets

In this study, we applied a number of feature selection methods to identify biomarkers in four datasets, the Longitudinal Youth at Risk Study (LYRIKS) [53], Bipolar disorder [54] Lung Adenocarcinoma [55] and Pancreatic Ductal Adenocarcinoma [56].

The LYRIKS dataset was recorded at the Institute for Mental Health (IMH), Singapore and contains 34 928 gene expression values for 84 participants (56 Ultra-High Risk [57] and 28 control). Ultra-High Risk (UHR) is a criterion that aims to identify people at-risk of developing psychosis in the future. Participants were assessed to be UHR using the Comprehensive Assessment of At Risk Mental States (CAARMS) [58]. RNA was extracted from peripheral blood and gene expression profiles were assessed on Illumina HumanHT-12 v4 Expression BeadChip arrays. Goh *et al.* [59] previously reported a 12-gene signature on this dataset that is 90% accurate in identifying people at UHR.

The Bipolar data were recorded by the University Medical Center Utrecht, Netherlands, and contains 20 581 gene expression values of 480 participants (240 bipolar and 240 control). Bipolar Disorder (BD) is a complex mental disorder characterized by mood instability, and has high level of heritablility [60]. The participants were diagnosed to be Bipolar using the Structured Clinical Interview for DSM-IV (SCID) [61] test. Peripheral whole blood was used for RNA-seq to derive the gene expression values.

The Lung Adenocarcinoma (LUAD) dataset was taken from The Cancer Genome Atlas (TCGA) PanCancer Atlas study [55]. It contains 20531 gene expression values of 569 participants (510 cancerous and 59 control). Lung adenocarcinoma is a common form of non-small cell lung cancer and is the most common type of lung cancer among non-smokers. It is also the second most common cause of cancer-related deaths worldwide [62]. The gene expression was obtained using the Illumina HiSeq platform.

The Pancreatic Ductal Adenocarcinoma (PDAC) dataset comprises of eight types of omics data sourced from 140 pancreatic tumor tissues, 67 paired normal adjacent tissues (NATs) and 9 normal pancreatic duct tissues. The data was collected in accordance with the Clinical Proteomic Tumor Analysis Consortium (CPTAC) guidelines. PDAC is a highly aggressive cancer with a dismal 5-year survival rate below 10%. Often diagnosed at advanced stages, it poses a significant challenge for effective treatment and is projected to become a leading cause of cancer death by 2030 [63]. In this study, we used the proteomic expression in which 11 662 proteins were quantified.

A summary of the datasets used in this study is shown in Table 2.

RESULTS

Analysis of feature selection ensembles

First, six feature selection methods (summarized in Table 1) were run unmodified (also referred to as vanilla) on the LYRIKS dataset. Here, 5-fold cross validation was used to evaluate the

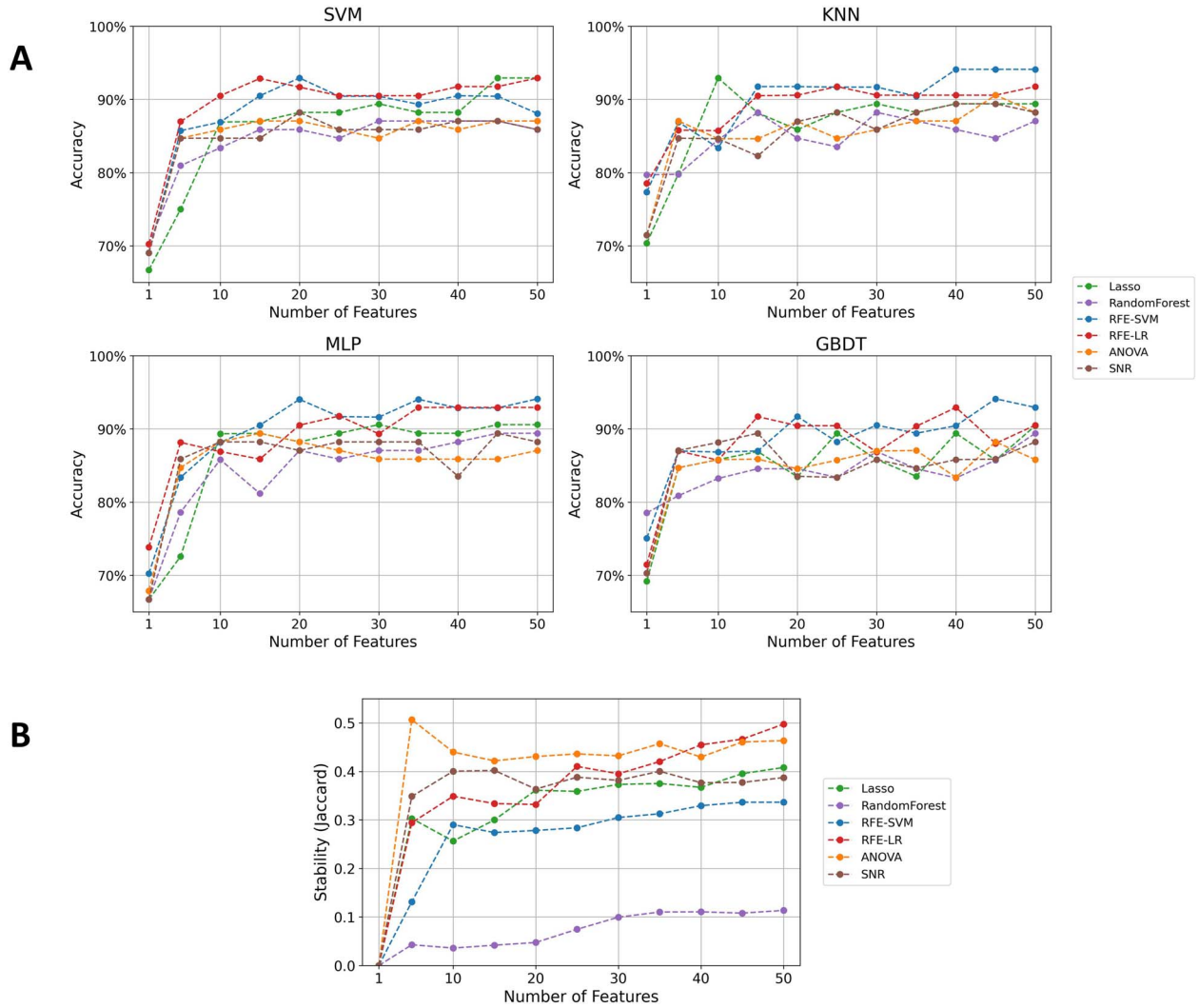

Figure 2. (**A**) Comparison of the classification accuracies of different classifiers using the top features selected by various feature selection algorithms. The accuracy represents the average accuracy of the classifier across the K folds of cross-validation. (**B**) Comparison of stability of the features selected by the various feature selection algorithms across the K folds. The stability score is calculated using Jaccard Index (defined in Equation 3).

performance of the methods. In each fold, the feature selection methods were run on the training set and the identified features were used to train four different classifiers viz. SVM [64], K-Nearest Neighbours (KNN) [65], Multi-layer Perceptron (MLP) [66, 67] and Gradient-Boosted Decision Tress (GBDT) [68], for classification of UHR and control groups. Four different types of classifiers were used to ensure the selected features perform well across any type of classification method. The accuracy of these classifiers was evaluated on the test set. Figure 2(A) shows the accuracy of these classifiers with increasing number of selected features.

The highest accuracies have been achieved by the wrapper methods RFE-SVM and RFE-LR. Lasso performs the next best while Random Forest gives lowest accuracies over the top 20 features. For the filter methods ANOVA and SNR, after the first five features, the accuracy remains nearly constant with increasing number of features.

The stability of these traditional algorithms with increasing number of features is shown in Figure 2(B). It is evident that the filter methods ANOVA and SNR are more stable relative to other methods. ANOVA achieves high top 5 feature consistency. RFE-LR achieves higher stability than filter methods after the top 40 features, whereas its companion wrapper method RFE-SVM, achieves much lower stability comparatively. Random Forest performs the worst in terms of stability because of the intrinsic randomness involved in the algorithm [69].

In the following, we explored the effect of the three ensemble techniques (presented in Figure 1A) for feature selection.

Bagging

The first ensemble technique explored in the study is bagging. In bagging, in each fold of the cross-validation, 10 bootstrap samples were created, each bootstrap sample containing a subset of available samples in the training set, chosen at random with replacement. The feature selection algorithms were run on these bootstrap samples, and the feature rankings generated by each algorithm were aggregated to obtain one final ranking per algorithm.

The mean accuracies of the feature selection algorithms across the four classifiers viz. SVM, KNN, MLP and GBDT, with and without bagging, are compared in Figure 3(A). The mean accuracy does not change much with bagging, across all types of feature selection methods. Figure 3(B) records the stability of the features selected by the algorithms. Significant improvement in stability

can be seen for algorithms that involve some randomness in initialization like Random Forests. The stability does not improve for filter methods SNR and ANOVA partially because there are no hyperparameters whose values affect the calculation of feature importance.

Voting

In Voting, rankings from different algorithms are combined using rank aggregation to obtain one ensemble ranking. To analyse the effect of voting, the vanilla rankings of SNR and ANOVA were combined into a filter ensemble, RFE-SVM and RFE-LR into a wrapper ensemble, Lasso and Random Forest into an embedded ensemble and all the six algorithms in one ensemble referred to as 'All'. Figure 4(A) and (B) show the mean accuracies and stability scores of these ensembles compared with the vanilla algorithms used to create the ensemble.

From the results, it can be seen that in all cases, the accuracy and stability of an ensemble ranking are nearly an average of the vanilla rankings used to create them. For example, the stability results of the filter ensemble lie midway between the stability of ANOVA and SNR and similarly, the mean accuracy of the filter ensemble lies between the accuracy of ANOVA and SNR.

Stacking

Stacking is an ensembling technique in which the output of one algorithm is provided as input to another algorithm. In the feature selection context, this means the features selected by the first algorithm are passed to the second algorithm to further select a smaller subset of important features. In the analysis, one feature selection technique was used from each category, viz. ANOVA from filter, RFE-SVM from wrapper and Lasso from embedded and basic stacking ensembles were explored that can be created using two out of the three algorithms. The first algorithm was used to select the top 50% of the features and the second algorithm ranked the remaining features to achieve the final ranking.

Figure 5(A) shows the mean accuracy of the stacking ensembles compared with the vanilla algorithms used in them. The only stack that significantly outperforms its vanilla algorithms is the ANOVA+RFE-SVM stack, where ANOVA is the first algorithm that is used to select the top 50% of features and RFE-SVM is the second and final algorithm used to rank the selected top 50% features. RFE-SVM is the closest to the stack in terms of mean accuracy, but as Figure 5(B) illustrates, the ANOVA+RFE-SVM stack achieves higher stability than vanilla RFE-SVM.

Based on the insights gathered from these ensembling analyses, the FWSE algorithm was designed. For details about the FWSE algorithm please refer to the Methodology section.

Comparative analysis of FWSE and established approaches

Case study #1: LYRIKS data

The LYRIKS dataset, which deals with UHR individuals, presents a unique landscape for biomarker discovery. For this dataset, in FWSE, we employed ANOVA and SNR as filter methods, and RFE-SVM and RFE-LR as wrapper methods, creating 10 bootstrap subsets at both stages, each subset equal in size to the original dataset.

As depicted in Figure 6(A), FWSE outperforms traditional feature selection algorithms in terms of accuracy. The filter methods in FWSE effectively eliminate irrelevant features, while the wrapper methods rank the remaining relevant features to identify a minimal subset that maximizes accuracy. The stability of FWSE is also comparable to the most stable traditional algorithms, demonstrating the robustness of our ensemble approach. The usage of bagging and multiple filter/wrapper methods greatly improves the stability of the proposed ensemble compared with a simple one filter and one wrapper method stack. ANOVA and RFE-LR are closest to FWSE's stability but FWSE heavily outperforms them in accuracy, achieving $\approx$ 93% mean accuracy in top 15 features and $\approx$ 95% mean accuracy in top 30 features. RFE-SVM is the closest to FWSE in terms of accuracy, but the ensemble significantly outperforms RFE-SVM in stability.

When compared to ensemble biomarker discovery methods (Figure 6B), FWSE exhibits superior accuracy and stability on the LYRIKS dataset. While E-Borda comes close in terms of accuracy, FWSE achieves much higher stability, demonstrating its robustness in the face of data perturbations. Similarly, ESVM-RFE and VSOLassoBag initially give better stability than FWSE when the number of selected features is less than 30, but FWSE significantly outperforms both in terms of accuracy ($\approx$ 8% on average).

Case study #2: Bipolar data

The Bipolar dataset, with its focus on a complex mental disorder, presents a challenging testbed for our proposed FWSE method due to low separability between the target classes. For this dataset, ANOVA and SNR were employed as filter methods and RFE-RF was employed as the wrapper method in FWSE.

As shown in Figure 6(C), FWSE outperforms traditional methods in terms of both accuracy and stability. RFE-RF comes close in terms of accuracy, but FWSE demonstrates higher stability, especially when the number of selected genes is less than 20.

When compared with ensemble biomarker discovery methods (Figure 6D), FWSE again outperforms the competition in terms of both accuracy and stability, demonstrating its robustness across different datasets and conditions. MCF-RFE is closest to FWSE in terms of accuracy and VSOLassoBag is closest in terms of stability but when accuracy and stability are considered together, FWSE significantly outperforms all the ensemble biomarker discovery methods. This demonstrates the robustness of FWSE, even in the face of data perturbations on a dataset with low separability, making it a promising tool for biomarker discovery in mental health research.

Case study #3: LUAD data

The LUAD dataset, which focuses on Lung Adenocarcinoma, presents a different set of challenges due to very high separability. Based on the performance of the traditional feature selection methods, in FWSE, ANOVA and SNR were employed as the filter methods and RFE-SVM and RFE-LR were employed as the wrapper methods.

As shown in Figure 6(E), FWSE achieves the highest accuracy among traditional feature selection methods, although its stability is lower than filter methods ANOVA and SNR. This is likely due to the high separability of the LUAD dataset, which contains many groups of genes that can achieve similar high levels of separability. FWSE does outperform other wrapper and embedded methods in terms of stability. RFE-LR has slightly lower accuracy and stability than FWSE. Lasso comes close in terms of accuracy only around the top 20 features mark but has lower stability than FWSE throughout.

When compared with ensemble biomarker discovery methods, FWSE outperforms all other approaches in terms of both accuracy and stability (Figure 6F). MCF-RFE, while slightly lower in accuracy and slightly higher in stability around top 15–20 features, eventually falls behind in both accuracy and stability. E-Borda

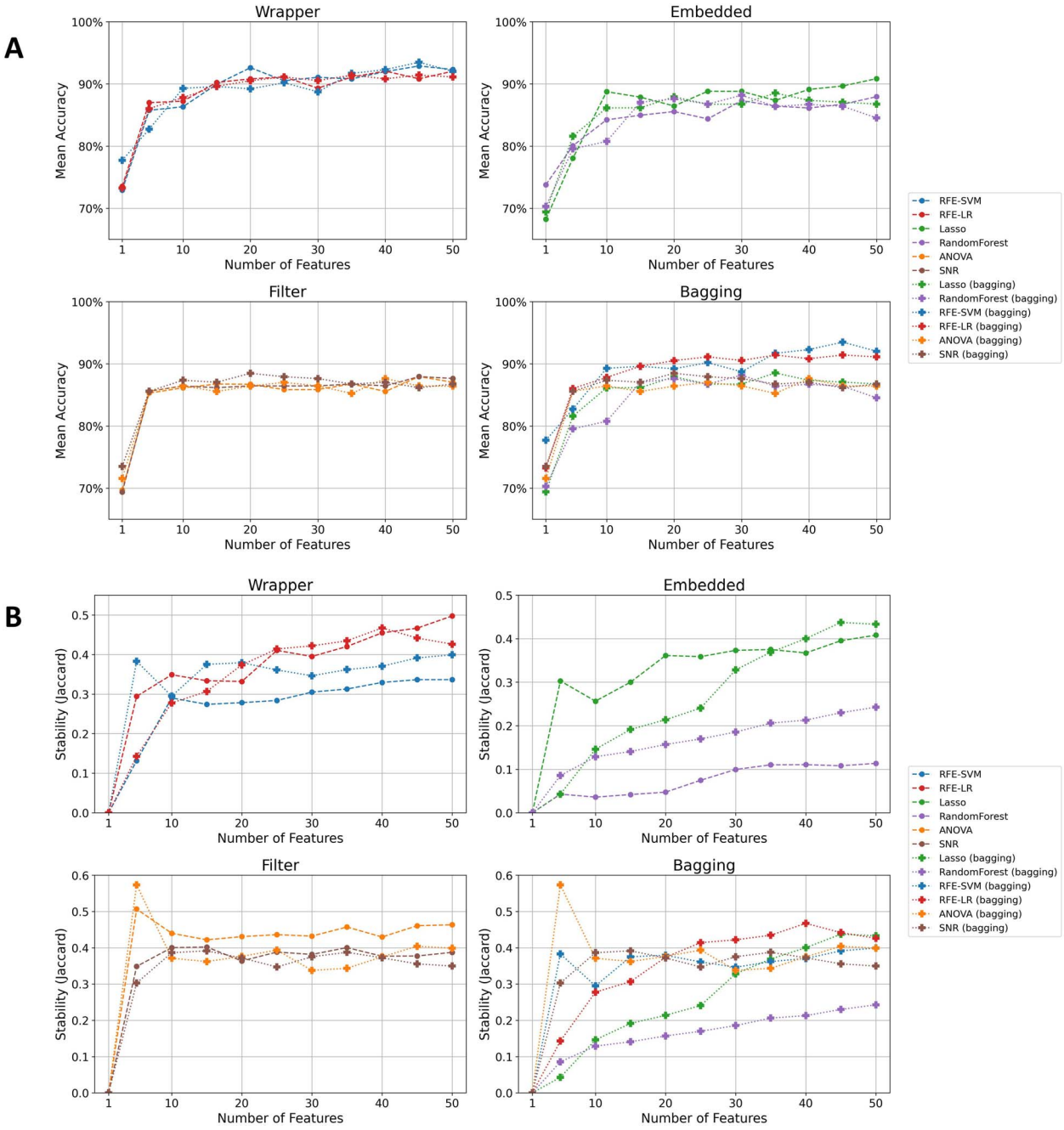

Figure 3. Comparison of feature selection methods and their bagging variants. (**A**) The accuracy of all wrapper, embedded and filter methods displays negligible changes after implementing bagging. (**B**) Bagging ensembles substantially enhance stability, especially for methods where initialization involves an element of randomness.

also comes close to FWSE after the top 35 features, but FWSE maintains higher accuracy and stability.

Case study #4: PDAC data

The PDAC dataset also exhibits good separability, like the LUAD dataset. For PDAC, in FWSE, ANOVA and SNR were employed as the filter methods, and RFE-RF and RFE-LR (with L1 norm) were employed as the wrapper methods. Just like in previous case studies, 10 bootstrap subsets were created at each stage, each subset equal in size to the original dataset.

As depicted in Figure 6(G), FWSE stands out as the most accurate among traditional machine learning feature selection methods. While Lasso displays slightly higher stability in the top 20–30 features, FWSE surpasses it considerably in terms of accuracy. The stability of SNR and ANOVA on this dataset further confirm this trend that in datasets with high separability, filter methods tend to achieve the highest stability. However, FWSE outperforms these methods in accuracy, making it a superior choice when prioritizing accuracy.

When pitted against ensemble feature selection methods, FWSE again outperforms in terms of accuracy and stability.

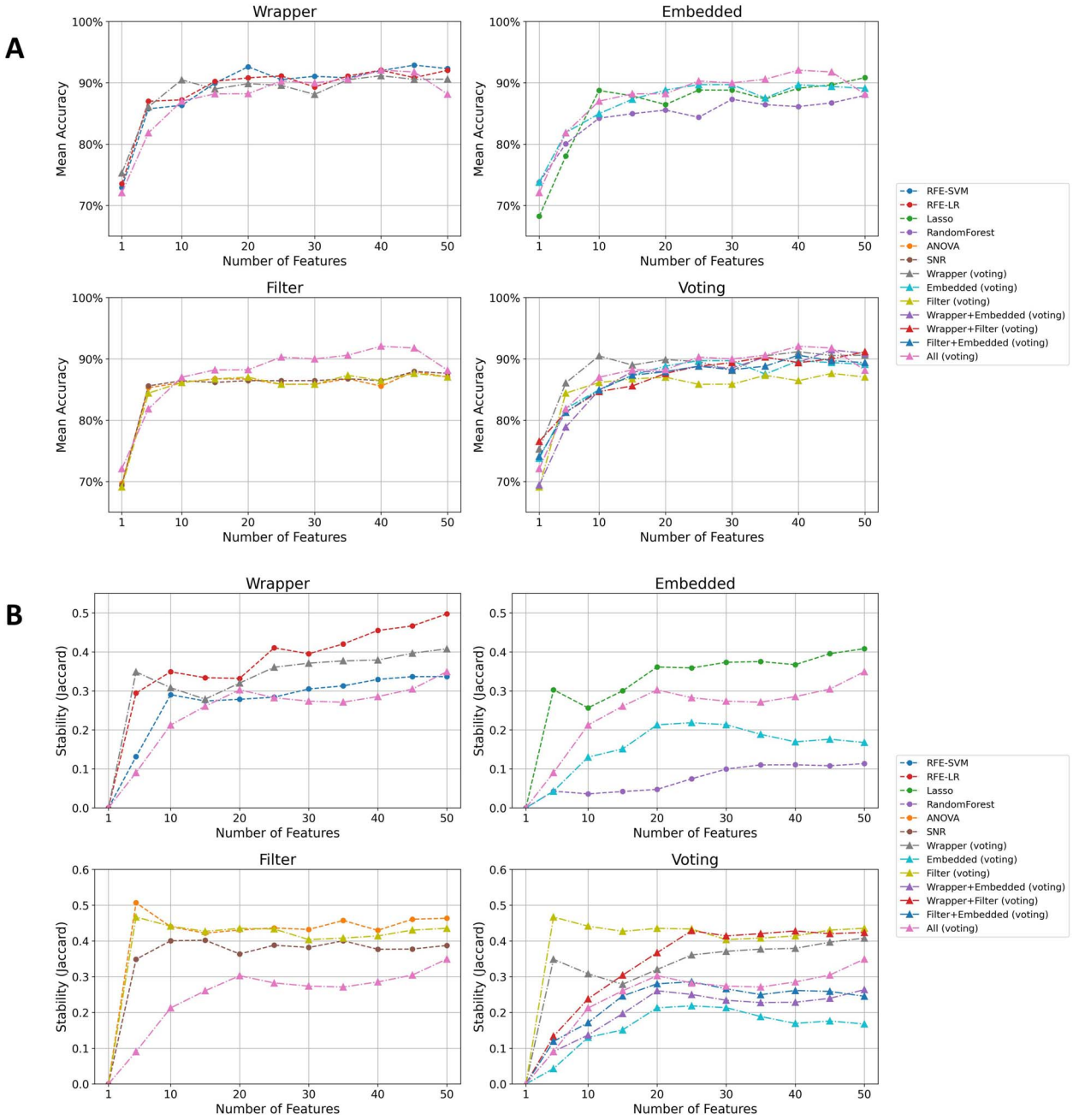

Figure 4. Comparison of feature selection methods to their voting ensembles. (**A**) The accuracy of voting ensembles tends to reflect the mean performance of the individual methods included in the ensemble. (**B**) Similarly, the stability of voting ensembles typically approximates the average stability of the individual algorithms within the ensemble.

E-borda demonstrates comparable accuracy but falls significantly short in stability. VSOLassoBag exhibits equivalent stability in the top 15–30 features but does not match FWSE's accuracy. ESVM-RFE shows subpar performance in both accuracy and stability.

These results underscore the versatility of FWSE across different datasets and conditions. An interesting trend to note is the relationship between class separability and the stability of FWSE. In the Bipolar dataset, where class separability is the lowest, FWSE surpasses all other methods in terms of stability, demonstrating its resilience in challenging conditions. In the LYRIKS dataset, FWSE's stability is on par with the top-performing approach, further attesting to its robustness. However, in the LUAD and PDAC datasets, where class separability is very high, the stability of FWSE is slightly lower than that of filter methods. This is potentially because, in high separability scenarios, many groups of genes can achieve the same level of accuracy. Nonetheless, the high accuracy and overall performance of FWSE across all datasets underscores its potential as a reliable and versatile tool for biomarker discovery from high-dimensional omics data.

Biological significance

Beyond demonstrating enhanced accuracy and stability, we further evaluated the biological relevance of the biomarkers identified by FWSE.

The genes selected on LUAD, including ALDH18A1, CSTF2, MYO7A, SMYD5, SRPK1, C1orf63, GMPPA, ZNF207 and ABCC3,

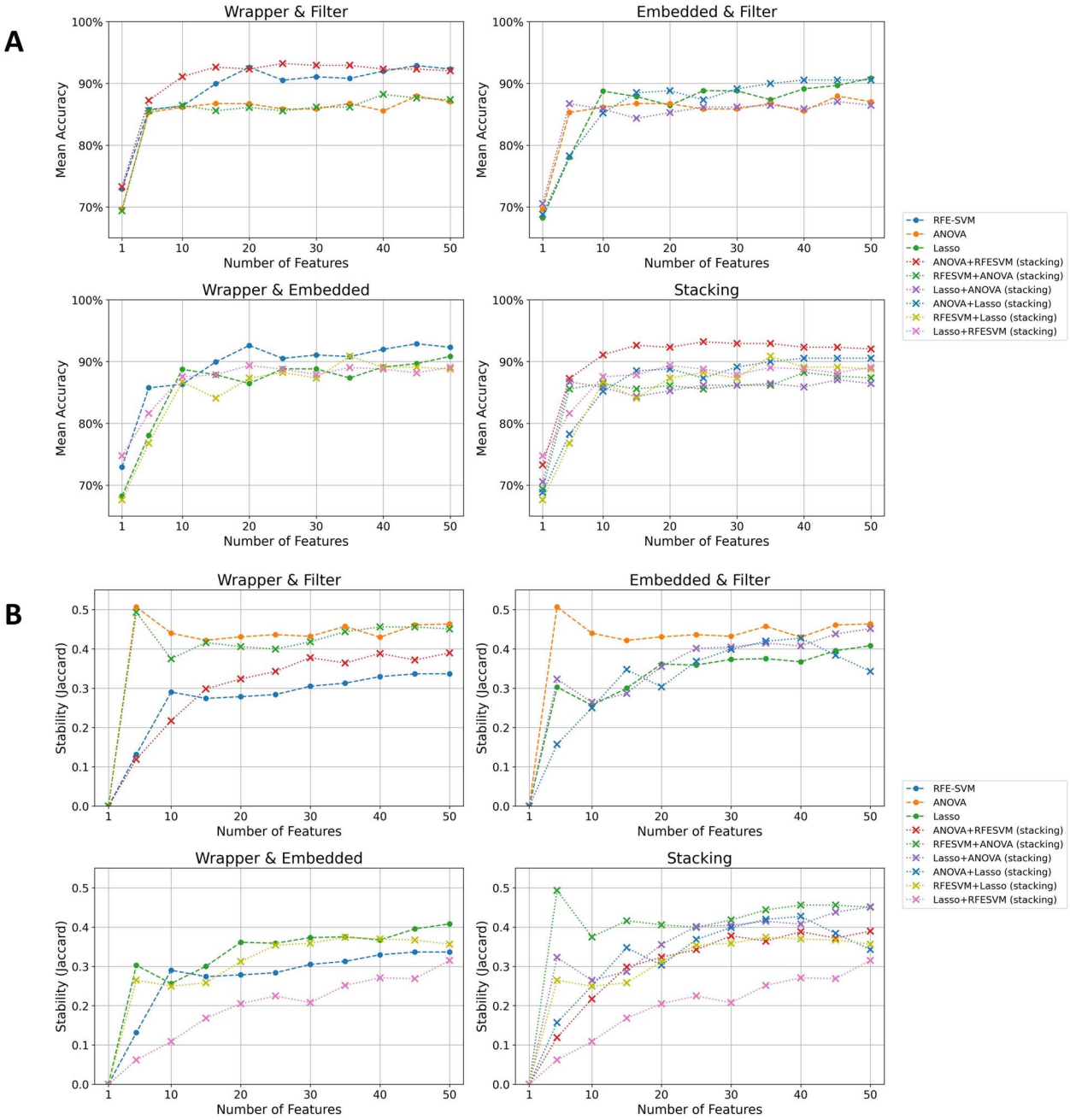

Figure 5. Comparison of feature selection methods and their stacking combinations. (**A**) The combination of Filter and Wrapper methods (ANOVA+RFE-SVM stack) demonstrates enhanced accuracy relative to their individual performances. (**B**) Furthermore, the ANOVA+RFE-SVM stack exhibits superior stability compared with the standalone Wrapper method RFE-SVM.

have demonstrated associations with lung cancer and other cancers in the literature. For instance, ALDH18A1 has shown overexpression in lung cancer [70]. CSTF2 has been identified as an independent prognostic factor in non-small cell lung cancer (NSCLC) and its suppression has been linked to inhibited lung cancer cell growth [71]. MYO7A, although primarily studied in the context of melanoma, has demonstrated roles in cell growth and migration, suggesting potential significance in lung cancer [72]. SMYD5 and SRPK1, when depleted, have been associated with increased tumor growth [73, 74]. C1orf63 has been found overexpressed in several cancers including lung cancer [75]. Mutations in GMPPA have shown a significant association with patient mortality in LUAD [76]. SERINC2 plays a critical role in LUAD, with SERINC2 knockdown shown to inhibit proliferation, migration and invasion in this cancer type [77]. ABCC3 has been identified as a marker for multiple drug resistance and predictor for poor clinical outcome in NSCLC, indicating its critical role in lung cancer pathology [78].

On PDAC, FWSE identified some proteins that could be key potential biomarkers such as S100A14, MISP, SFN, SULF1, SAMD9 and SERPINB5. S100A14 is not only an indicator of PDAC progression but also contributes to gemcitabine resistance, making it a potential therapeutic target [79]. MISP's upregulation is linked to poor patient outcomes and is instrumental in immune system alterations in PDAC [80]. SFN has been validated as a stromal marker with prognostic significance, specifically affecting both

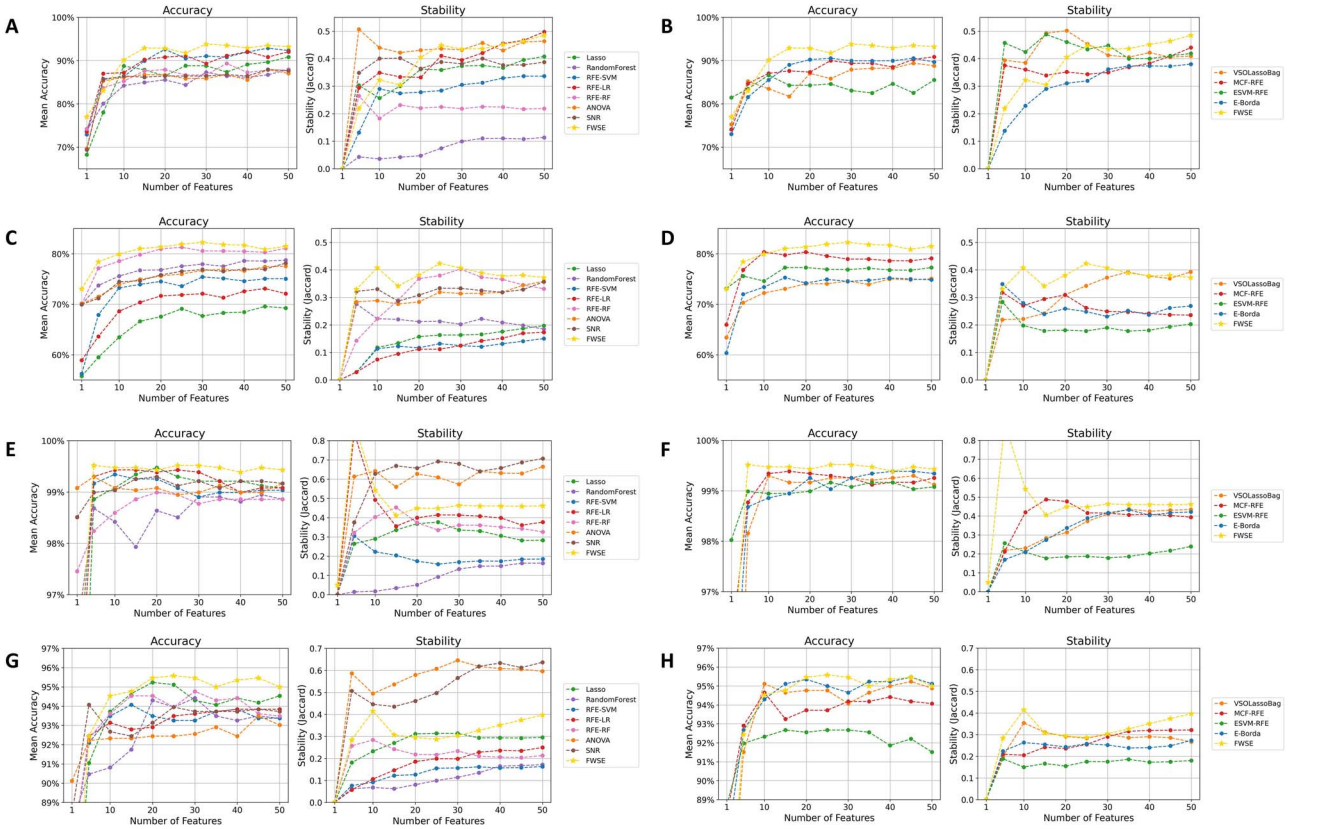

Figure 6. Comparison of mean accuracy and stability of FWSE to other feature selection algorithms. (**A**) Comparison against traditional feature selection methods on the LYRIKS data. (**B**) Comparison against ensemble feature selection methods on the LYRIKS data. (**C**) Comparison against traditional feature selection methods on the Bipolar data. (**D**) Comparison against ensemble feature selection methods on the Bipolar data. (**E**) Comparison against traditional feature selection methods on the LUAD data. (**F**) Comparison against ensemble feature selection methods on the LUAD data. (**G**) Comparison against traditional feature selection methods on the PDAC data. (**H**) Comparison against ensemble feature selection methods on the PDAC data.

overall and disease-free survival [81]. SERPINB5 stands out for its capacity to differentiate PDAC from pancreatitis, owing to its promoter hypomethylation [82]. SAMD9 and SULF1, although not as extensively studied, also show associations with PDAC pathology and are candidates for future investigations [83]. IGF2BP3 plays a role in PDAC malignancy by affecting cell invasiveness and modulating miRNA-mRNA interactions [84]. The comprehensive list of potential biomarkers recognized utilizing our novel FWSE method is provided in Appendix A.

Further strengthening our findings, the Gene Ontology (GO) enrichment analysis of the biomarker signatures identified using FWSE revealed compelling biological significance with *P*-values less than 0.01 considered significant. The LYRIKS signature was enriched in GO terms related to integral components of the membrane, plasma membrane, oxygen binding, heme binding and endoplasmic reticulum membrane. The plasma membrane and endoplasmic reticulum are integral to the immune response, facilitating the recognition of antigens and the production of immune-related proteins. This aligns with the widely reported observation of immune dysfunction in individuals at UHR for psychosis [85].

In the context of bipolar disorder, our FWSE methodology pinpointed genes enriched in GO terms such as carbohydrate binding, inflammatory response, apoptotic process, and cellular response to lipopolysaccharide. These terms are closely tied to immune function as well. For instance, carbohydrate binding is involved in cell-cell recognition, a crucial aspect of immune response, while inflammatory response and cellular response to lipopolysaccharide are directly linked to immune activation. This is consistent with the growing body of evidence suggesting a role for immune dysfunction in the pathophysiology of bipolar disorder [86].

The biomarker signature identified in the LUAD dataset from TCGA exhibited enrichment in GO terms related to the endoplasmic reticulum membrane, ER to Golgi vesicle-mediated transport, mitochondrion, and ATP binding. The ER is involved in protein folding and transport, lipid metabolism and calcium homeostasis, disruptions in which can lead to ER stress, a condition implicated in various diseases, including cancer [87]. The ER also plays a role in vesicle-mediated transport to the Golgi apparatus, a pathway crucial for protein secretion [88]. Mitochondria, known for their role in energy production through ATP synthesis, also play key roles in apoptosis and reactive oxygen species (ROS) production, critical processes in cancer development [89].

Lastly, the biomarker signature on the PDAC dataset was enriched in GO terms associated with actin binding, calcium ion binding, extracellular space and cytosol. Actin binding is relevant for cellular structure and motility, potentially contributing to cancer cell invasiveness. Calcium ion binding is involved in various cellular processes, including signal transduction pathways that could be altered in cancerous cells. The extracellular space is key for cell-to-cell communication, often dysregulated in cancer, and the cytosol is involved in numerous metabolic and signaling pathways. These terms complement previous findings on GO terms related to extracellular structure and binding properties being associated with PDAC [83].

In summary, the identified biomarker signature's enrichment in these GO terms aligns with existing literature on UHR, Bipolar, LUAD, and PDAC, further validating the biological relevance of the identified biomarkers. Altogether, these findings underscore the potential of our FWSE method in identifying biologically relevant biomarkers across diverse disease contexts and emphasize the importance of considering the collective action of these biomarkers in disease pathology.

DISCUSSION

In this study, we introduce a novel method to address a critical scientific challenge associated with high-dimensional medical data, often referred to as 'the curse of dimensionality.' More specifically, our focus is on the discovery of reliable and stable biomarkers. While statistical tests have been routinely employed to identify differentially expressed genes as biomarkers due to their simplicity, they operate under the assumption of feature independence. This is often not the case in reality, as genes function in interconnected networks and pathways [90]. Recent studies have demonstrated that ensemble methods can effectively overcome the inherent limitations of various feature selection methods, thereby achieving a more reliable consensus [91].

The first phase of the study delves into an exploration of various feature selection techniques (filter, embedded and wrapper) for biomarker selection from high-dimensional omics data. We also investigate combinations of these methods using popular ensemble techniques (voting, bagging and stacking). Our findings underscore that filter methods, which utilize univariate or multivariate statistics independent of the classifier used, yield stable features that are less prone to overfitting [92]. In contrast, wrapper and embedded methods, while often delivering higher accuracy when using the same classifiers that were used for feature selection, can overfit due to their complex nature, as evidenced by lower stability in some cases. Our study further illustrates that ensemble approaches can enhance the consistency (using bagging and voting) and performance (using stacking) of the selected features by leveraging the diversity of the feature selection algorithms used.

The second phase of the study presents the novel FWSE method. FWSE synergistically combines filter and wrapper methods to create a signature that not only delivers high accuracy but is also stable, reproducible and biologically significant. The use of bootstrapping enhances the stability of our approach and mitigates overfitting. The stacking strategy allows us to harness the strengths of both filter and wrapper methods. The filter methods effectively eliminate non-differentially expressed genes across groups, while the wrapper methods evaluate the remaining features as a collective set. This results in a set of genes that provide high group separability, even though each gene may be a weak biomarker individually. Our robust signature identified for the LYRIKS data outperforms previous works in predicting UHR criteria [59], and the genes and proteins identified for LUAD and PDAC are associated with the respective cancers.

Despite FWSE's strong performance in accuracy and stability, it is computationally intensive. The use of time-consuming wrapper feature selection techniques like RFE, when repeated on multiple bootstrap samples, slows the process even further. Future work will explore the incorporation of more constraints into the FWSE architecture to improve computational efficiency and biological relevance, thereby further enhancing its utility in the realm of biomarker discovery.

In summary, this study represents a meaningful contribution to the field of biomarker discovery, by systematically analysing the effect of ensemble feature selection methods and providing a novel and robust approach to identifying stable and biologically relevant biomarkers in high-dimensional medical data. The implications of this work are broad, with potential applications in disease diagnosis, prognosis and therapeutic development.

Appendix A: List of Identified Biomarkers

This section shows the complete list of potential biomarkers identified for diagnosis of Ultra-High Risk (UHR) and bipolar using the novel ensemble feature selection method, FWSE, developed in this study. The top 20 markers are listed in Table 3 in order of importance. The use of each of the biomarker sets for diagnosis, prognosis and treatment design is subject to a copyright by the authors.

Table 3: Potential biomarkers identified for the different datasets using FWSE

Dataset	Potential Biomarkers
LYRIKS	LDLRAD1, CYP8B1, CDH11, TMEM225, HS.170946, LOC100134138, HS.537754, C2CD3, LRRTM2, OR4A15, LOC100127888, LOC100129002, LOC100131961, PLA2G5, LOC389118, LOC100134413, LOC641964, SVOPL, LOC100133959, SNORD113-6, OR56B4, LOC653113, SRD5A2, HS.146184, CNTNAP5
Bipolar	TSPAN2, TAGLN2, FAR2, CXCL8, PFKFB2, LINC01765, CD300A, TAGLN2P1, MIR23AHG, FLT3, PIGB, SLC31A2, IVNS1ABP, AKAP12, CNTNAP3, LINC00877, YIPF4, LILRA4, RFX2, SYTL3, PGM5, CFAP45, MAK, DNASE1L3, CASP10
LUAD	ALDH18A1, SRPK1, SMYD5, KIAA0907, MYO7A, CSTF2, SERINC2, ZNF207, ABCC3, GMPPA, C1orf63, C1orf131, UBFD1, DLG3, P4HB, GYG2, SKIV2L, TXNDC5, PVRL4, NEK6
PDAC	C19orf33, GLA, S100A14, MISP, SCEL, IGF2BP3, SDCBP2, GALNT7, HEPH, SFN, SULF1, SAMD9, SERPINB5, PGM2L1, LMO7, MDK, REG4, STON1, HK2, GSDMB, ARPC1B, MYO1E, SDR16C5, S100A16, ACTN1

Appendix B: Sensitivity Analysis of Pruning Factor

In this appendix, we aim to analyse the sensitivity of the pruning factor, which plays an important role in the feature elimination process in our FWSE method. Although a pruning factor of 0.5 was used in the study, this section explores how variations in this parameter influence the performance metrics of accuracy and stability.

The pruning factor specifies the proportion of features that are discarded during the initial filter-based feature selection phase of the FWSE algorithm. It is expressed as a fraction and takes on values within the range of 0–1. To conduct this analysis, we consider multiple values for the pruning factor 0, 0.2, 0.33, 0.5, 0.66, 0.8 and 1.

We evaluate the impact of changing the pruning factor on two distinct datasets: LYRIKS, which contains microarray gene expression data, and PDAC, which has proteomics data and is characterized by high separability. For each dataset and pruning factor setting, we calculate the mean accuracy and mean stability

Table 4: Impact of varying pruning factor on accuracy and stability, averaged across top 50 features

Pruning	LYRIKS		PDAC	
Factor	Accuracy	Stability	Accuracy	Stability
0	88.41	0.325	94.88	0.367
0.2	88.19	0.339	94.77	0.355
0.33	89.17	0.340	94.65	0.356
0.5	90.29	0.346	94.53	0.349
0.66	89.20	0.340	94.11	0.351
0.8	87.77	0.322	93.97	0.338
1	88.08	0.371	93.09	0.577

that FWSE achieves across the top 50 features. The results of this analysis are presented in Table 4.

On LYRIKS, the accuracy of FWSE appears to peak at the pruning factor of 0.5, achieving a mean accuracy of 90.29%. The accuracy tends to decrease as the pruning factor deviates from this optimal value in either direction. A similar trend can be observed with FWSE's stability, except when the pruning factor is set to 1. This suggests that a pruning factor of 0.5 provides a balanced feature selection that maximizes classification accuracy for this particular dataset.

For PDAC, the accuracy tends to increase as the pruning factor decreases. When the pruning factor is set to 0, the features are ranked only using wrapper methods. Hence, we see higher accuracy on PDAC. Whereas when the pruning factor is set to 1, the features are ranked purely using filter methods. Hence, higher stability can be observed on both datasets.

In summary, this sensitivity analysis reveals that the pruning factor has a nuanced impact on the performance of the FWSE method. It implies that the optimal setting may depend on the specific characteristics of the dataset and the priorities of the study (e.g., maximizing accuracy vs. stability). Future work will include developing methods to automatically determine the most appropriate pruning factor for different types of datasets.

Key Points

- The curse of dimensionality is a major challenge when dealing with high-dimensional medical data and finding reliable biomarkers. Ensembling techniques, as demonstrated in this study, can improve the consistency and performance of feature selection algorithms, overcoming the limitations of individual methods.
- Different ensemble techniques offer unique advantages. Voting ensemble provides an average accuracy across the methods applied, while bagging ensemble enhances stability for methods that involve randomness in their initialization. Stacking of filter and wrapper feature selection methods, on the other hand, leads to a notable improvement in accuracy.
- The proposed FWSE method stacks multiple bagging ensembles of filter and wrapper methods to create a robust biomarker signature. This novel approach outperforms traditional algorithms in terms of accuracy and stability, although it is computationally expensive.
- The biomarkers identified by FWSE are not only statistically significant but also biologically relevant, enhancing the practical utility of the method in disease diagnosis, prognosis, and therapeutic development. Furthermore, the stability of the FWSE method underscores its potential as a reliable tool for biomarker discovery.

AUTHOR CONTRIBUTIONS STATEMENT

S.B. spearheaded the development of the methodology and experimental setup, conducted the experiments, analyzed the results and drafted the initial manuscript. M.D. secured the project funding in New Zealand, contributed to the experimental design, reviewed the experimental results and participated in manuscript writing. B.S. assisted in the pre-processing of the Bipolar data and provided critical review of the manuscript. S.T. played a key role in pre-processing and cleaning the LYRIKS datasets, and offered valuable insights into the biological significance of the results. Z.D. contributed to the interpretation of experimental results and provided manuscript review. E.L. reviewed the experimental results and the manuscript, providing valuable feedback. A.M. aided in the analysis of the selected markers and reviewed the biological significance of the markers. W.G. secured the project funding in Singapore, and contributed to data pre-processing and interpretation of the results. J.L. provided insights into the LYRIKS data collection and critically reviewed the results. N.K. secured the project funding in New Zealand, contributed to the design of the experiments, and reviewed the experimental results and the manuscript. All authors have read and agreed to the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their valuable suggestions.

FUNDING

This research is supported by the MBIE Catalyst: Strategic–New Zealand-Singapore Data Science Research Program and the National Research Foundation, Singapore, under its Industry Alignment Fund–Pre-positioning (IAF-PP) Funding Initiative. The LYRIKS study was supported by the National Research Foundation Singapore under the National Medical Research Council Translational and Clinical Research Flagship Program (NMRC/TCR/003/2008). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

COMPETING INTERESTS

The authors have no competing interests as defined by Oxford University Press, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

DATA AVAILABILITY

This study employed four distinct datasets. The LYRIKS dataset, owned by the Institute for Mental Health, Singapore, is not publicly available due to privacy considerations and the absence of

participant consent for public data sharing. Researchers interested in accessing this dataset for scientific purposes may reach out directly to the Institute for Mental Health, Singapore, to explore potential data access arrangements.

The Bipolar dataset, on the other hand, is publicly accessible and can be downloaded from the following link: [Link]. The Lung Adenocarcinoma (LUAD) dataset, part of The Cancer Genome Atlas (TCGA) PanCancer Atlas study, can be downloaded from the following link: [Link]. The Pancreatic Ductal Adenocarcinoma (PDAC) dataset is publicly available at the following link: [Link]

We encourage researchers to utilize these resources in accordance with the respective data use agreements and ethical guidelines.

CODE AVAILABILITY

The complete source code for the Filter and Wrapper Stacking Ensemble (FWSE) method, as implemented in this study, is openly available on GitHub: [Link]. We encourage the use of this code for academic and research purposes.

REFERENCES

1. Strimbu K, Tavel JA. What are biomarkers? *Curr Opin HIV AIDS* 2010;**5**(6):463.
2. Cagney DN, Sul J, Huang RY, *et al.* The FDA NIH biomarkers, endpoints, and other tools (best) resource in neuro-oncology. *Neuro Oncol* 2018;**20**(9):1162–72.
3. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet* 2008;**24**(3):133–41.
4. Heller MJ. DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng* 2002;**4**(1):129–53.
5. Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**(1):57–63.
6. Mohammadi M, Noghabi HS, Hodtani GA, Mashhadi HR. Robust and stable gene selection via maximum–minimum correntropy criterion. *Genomics* 2016;**107**(2–3):83–7.
7. Dessì N, Pascariello E, Pes B. A comparative analysis of biomarker selection techniques. *Biomed Res Int* 2013;**2013**:1–10.
8. Pollack JR, Perou CM, Alizadeh AA, *et al.* Genome-wide analysis of dna copy-number changes using cDNA microarrays. *Nat Genet* 1999;**23**(1):41–6.
9. Loscalzo S, Yu L, Ding C. Consensus group stable feature selection. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paris, France: Association for Computing Machinery, pp. 567–76, 2009.
10. Ioannidis JPA, *et al.* Microarrays and molecular research: noise discovery? *Lancet* 2005;**365**(9458):454–4, 5.
11. He Z, Weichuan Y. Stable feature selection for biomarker discovery. *Comput Biol Chem* 2010;**34**(4):215–25.
12. Goh WWB, Wong L. Evaluating feature-selection stability in next-generation proteomics. *J Bioinform Comput Biol* 2016;**14**(05):1650029.
13. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform* 2013;**14**(1):1–18.
14. Lyons-Weiler J, Patel S, Bhattacharya S. A classification-based machine learning approach for the analysis of genome-wide expression data. *Genome Res* 2003;**13**(3):503–12.
15. Dalman MR, Deeter A, Nimishakavi G, Duan Z-H. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics*. Springer, 2012;**13**(2):1–4.
16. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. Nature Publishing Group UK London, 2019;**567**(7748):305–307.
17. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle p value generates irreproducible results. *Nat Methods* 2015;**12**(3):179–85.
18. Park KS, Kim SH, Jung Hun O, Kim SY. Highly accurate diagnosis of papillary thyroid carcinomas based on personalized pathways coupled with machine learning. *Brief Bioinform* 2021;**22**(4):bbaa336.
19. Toth R, Schiffmann H, Hube-Magg C, *et al.* Random forest-based modelling to detect biomarkers for prostate cancer progression. *Clin Epigenetics* 2019;**11**:1–15.
20. Xun W, Qin K, Iroegbu CD, *et al.* Genetic analysis of potential biomarkers and therapeutic targets in ferroptosis from coronary artery disease. *J Cell Mol Med* 2022;**26**(8):2177–90.
21. Liu Z, Li H, Pan S. Discovery and validation of key biomarkers based on immune infiltrates in Alzheimer's disease. *Front Genet* 2021;**12**:658323.
22. Brahim AB, Limam M. Robust ensemble feature selection for high dimensional data sets. In: *In 2013 International Conference on High Performance Computing & Simulation (HPCS)*. Helsinki, Finland: IEEE, 2013, 151–7.
23. IJzendoorn DGP VAN, Szuhai K, Briaire-de Bruijn IH, *et al.* Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLoS Comput Biol* 2019;**15**(2):e1006826.
24. Seijo-Pardo B, Porto-Díaz I, Bolón-Canedo V, Alonso-Betanzos A. Ensemble feature selection: homogeneous and heterogeneous approaches. *Knowl-Based Syst* 2017;**118**:124–39.
25. Liang J, Wang C, Zhang D, *et al.* Vsolassobag: a variable-selection oriented lasso bagging algorithm for biomarker discovery in omic-based translational research. *J Genet Genomics* 2023.
26. Anaissi A, Goyal M, Catchpoole DR, *et al.* Ensemble feature learning of genomic data using support vector machine. *PloS One* 2016;**11**(6):e0157330.
27. Shi Y-H, Wen T-F, Xiao D-S, *et al.* Predicting miRNA targets for hepatocellular carcinoma with an integrated method. *Transl Cancer Res* 2020;**9**(3):1752.
28. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;**23**(19):2507–17.
29. Yeung KY, Bumgarner RE, Raftery AE. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* 2005;**21**(10):2394–402.
30. Li Y, Si J, Zhou G, *et al.* FREL: a stable feature selection algorithm. *IEEE Trans Neural Netw Learn Syst* 2014;**26**(7):1388–402.
31. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci* 2002;**99**(10):6562–6.
32. Li J, Cheng K, Wang S, *et al.* Feature selection: a data perspective. *ACM Comput Surv (CSUR)* 2017;**50**(6):1–45.
33. Zhang X, Jonassen I, Goksøyr A. Machine learning approaches for biomarker discovery using gene expression data. *Exon Publ* 2021;53–64.
34. Kasabov N. Global, local and personalised modeling and pattern discovery in bioinformatics: an integrated approach. *Pattern Recogn Lett* 2007;**28**(6):673–85.
35. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Methodol* 1996;**58**(1):267–88.
36. Breiman L. Random forests. *Mach Learn* 2001;**45**(1):5–32.

37. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn* 2002;**46**(1):389–422.
38. Jing D, Jin W, Cai Z, *et al.* A new feature evaluation algorithm and its application to fault of high-speed railway. In: *International Conference on Intelligent Transportation*. Singapore, Singapore: Springer, 2016, 1–14.
39. Khaire UM, Dhanalakshmi R. Stability of feature selection algorithm: a review. *J King Saud Univ-Comput Inform Sci* 2019.
40. Drotár P, Gazda M, Vokorokos L. Ensemble feature selection using election methods and ranker clustering. *Inform Sci* 2019;**480**:365–80.
41. Kendall MG. A new measure of rank correlation. *Biometrika* 1938;**30**(1/2):81–93.
42. Myers JL, Well AD, Lorch Jr RF. *Research Design and Statistical Analysis*. New York, NY, USA: Routledge, 2013.
43. Opitz D, Maclin R. Popular ensemble methods: an empirical study. *J Artif Intell Res* 1999;**11**:169–98.
44. Dwork C, Kumar R, Naor M, Sivakumar D. Rank aggregation methods for the web. In: *Proceedings of the 10th international conference on World Wide Web*. Hong Kong, Hong Kong: Association for Computing Machinery, pp. 613–22, 2001.
45. Breiman L. Bagging predictors. *Mach Learn* 1996;**24**(2):123–40.
46. Breiman L. Pasting small votes for classification in large databases and on-line. *Mach Learn* 1999;**36**(1):85–103.
47. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell* 1998;**20**(8):832–44.
48. Louppe G, Geurts P. Ensembles on random patches. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Bristol, UK: Springer, 2012, 346–61.
49. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York, NY, USA: CRC press, 1994.
50. Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants. *Mach Learn* 1999;**36**(1):105–39.
51. Wolpert DH. Stacked generalization. *Neural Netw* 1992;**5**(2):241–59.
52. McLachlan GJ, Do K-A, Ambroise C. *Analyzing Microarray Gene Expression Data*. Hoboken, New Jersey, USA: John Wiley & Sons, 2005.
53. Lee J, Rekhi G, Mitter N, *et al.* The Longitudinal Youth at Risk Study (LYRICS)-an Asian UHR perspective. *Schizophr Res* 2013;**151**(1–3):279–83.
54. Krebs CE, Ori APS, Vreeker A, *et al.* Whole blood transcriptome analysis in bipolar disorder reveals strong lithium effect. *Psychol Med* 2020;**50**(15):2575–86.
55. Liu J, Lichtenberg T, Hoadley KA, *et al.* An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 2018;**173**(2):400–416.e11.
56. Cao L, Huang C, Zhou DC, *et al.* Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell* 2021;**184**(19):5031–5052.e26.
57. Yung AR, Phillips LJ, Yuen HP, *et al.* Psychosis prediction: 12-month follow up of a high-risk ("prodromal") group. *Schizophr Res* 2003;**60**(1):21–32.
58. Yung AR, Yung AR, Yuen HP, *et al.* Mapping the onset of psychosis: the comprehensive assessment of at-risk mental states. *Aust N Z J Psychiatr* 2005;**39**(11–12):964–71.
59. Goh WWB, Sng JC-G, Yee JY, *et al.* Can peripheral blood-derived gene expressions characterize individuals at ultra-high risk for psychosis? *Comput Psychiatr* 2017;**1**:168–83.
60. Grande I, Berk M, Birmaher B, Vieta E. Bipolar disorder. *The Lancet* 2016;**387**(10027):1561–72.
61. First MB, Gibbon M. The Structured Clinical Interview for DSM-IV axis I disorders (SCID-I) and the Structured Clinical iInterview for DSM-IV axis II disorders (SCID-II). In: Hilsenroth MJ, Segal DL. (eds.), *Handbook of Psychological Assessment*. John Wiley & Sons, Inc., Vol. **2**, 2004, 134–43.
62. Sung H, Ferlay J, Siegel RL, *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021;**71**(3):209–49.
63. Quante AS, Ming C, Rottmann M, *et al.* Projections of cancer incidence and cancer-related deaths in Germany by 2020 and 2030. *Cancer Med* 2016;**5**(9):2649–56.
64. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**(3):273–97.
65. Fix E, Hodges JL. Discriminatory analysis. Nonparametric discrimination: consistency properties. *Int Stat Rev* 1989;**57**(3):238–47.
66. Rosenblatt F. Principles of neurodynamics. Perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.
67. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. Technical report, California University San Diego La Jolla Institute for Cognitive Science, 1985.
68. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Analy* 2002;**38**(4):367–78.
69. Wang H, Yang F, Luo Z. An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinform* 2016;**17**(1):1–18.
70. Ye Z, Zhang H, Kong F, *et al.* Comprehensive analysis of alteration landscape and its clinical significance of mitochondrial energy metabolism pathway-related genes in lung cancers. *Oxid Med Cell Longev* 2021;**2021**.
71. Aragaki M, Takahashi K, Akiyama H, *et al.* Characterization of a cleavage stimulation factor, 3′ pre-RNA, subunit 2, 64 kDa (CSTF2) as a therapeutic target for lung cancercstf2 activation in lung cancer. *Clin Cancer Res* 2011;**17**(18):5889–900.
72. Liu Y, Wei X, Guan L, *et al.* Unconventional myosin VIIA promotes melanoma progression. *J Cell Sci* 2018;**131**(4):jcs209924.
73. Kidder BL, He R, Wangsa D, *et al.* SMYD5 controls heterochromatin and chromosome integrity during embryonic stem cell differentiationsmyd5 regulates genome stability. *Cancer Res* 2017;**77**(23):6729–45.
74. Liu H, Xuefei H, Zhu Y, *et al.* Up-regulation of SRPK1 in non-small cell lung cancer promotes the growth and migration of cancer cells. *Tumor Biology* 2016;**37**:7287–93.
75. Hong C-Q, Zhang F, You Y-J, *et al.* Elevated C1orf63 expression is correlated with CDK10 and predicts better outcome for advanced breast cancers: a retrospective study. *BMC Cancer* 2015;**15**(1):1–12.
76. Cho H-J, Lee S, Ji YG, Lee DH. Association of specific gene mutations derived from machine learning with survival in lung adenocarcinoma. *PloS One* 2018;**13**(11):e0207204.
77. Zeng Y, Xiao D, He H, *et al.* SERINC2-knockdown inhibits proliferation, migration and invasion in lung adenocarcinoma. *Oncol Lett* 2018;**16**(5):5916–22.
78. Zhao Y, Hailing L, Yan A, *et al.* ABCC3 as a marker for multidrug resistance in non-small cell lung cancer. *Sci Rep* 2013;**3**(1):1–6.
79. Zhu H, Gao W, Li X, *et al.* S100a14 promotes progression and gemcitabine resistance in pancreatic cancer. *Pancreatology* 2021;**21**(3):589–98.

80. Huang X, Zhao L, Jin Y, *et al.* Up-regulated MISP is associated with poor prognosis and immune infiltration in pancreatic ductal adenocarcinoma. *Front Oncol* 2022;**12**:827051.
81. Robin F, Angenard G, Cano L, *et al.* Molecular profiling of stroma highlights stratifin as a novel biomarker of poor prognosis in pancreatic ductal adenocarcinoma. *Br J Cancer* 2020;**123**(1): 72–80.
82. Mardin WA, Ntalos D, Mees ST, *et al.* SERPINB5 promoter hypomethylation differentiates pancreatic ductal adenocarcinoma from pancreatitis. *Pancreas* 2016;**45**(5):743–7.
83. Tan M, Ove B, Muckadell S DE, Joergensen MT. Gene expression network analysis of precursor lesions in familial pancreatic cancer. *J Pancreat Cancer* 2020;**6**(1):73–84.
84. Ennajdaoui H, Howard JM, Sterne-Weiler T, *et al.* IGF2BP3 modulates the interaction of invasion-associated transcripts with RISC. *Cell Rep* 2016;**15**(9):1876–83.
85. Radhakrishnan R, Kaser M, Guloksuz S. The link between the immune system, environment, and psychosis. *Schizophr Bull* 2017;**43**(4):693–7.
86. Rosenblat JD, McIntyre RS. Bipolar disorder and immune dysfunction: epidemiological findings, proposed pathophysiology and clinical implications. *Brain Sci* 2017;**7**(11): 144.
87. Hai H, Tian M, Ding C, Shengqing Y. The C/EBP Homologous Protein (CHOP) transcription factor functions in endoplasmic reticulum stress-induced apoptosis and microbial infection. *Front Immunol* 2019;**9**:3083.
88. Lee JE, Cathey PI, Haoxi W, *et al.* Endoplasmic reticulum contact sites regulate the dynamics of membraneless organelles. *Science* 2020;**367**(6477):eaay7108.
89. Ghemrawi R, Khair M. Endoplasmic reticulum stress and unfolded protein response in neurodegenerative diseases. *Int J Mol Sci* 2020;**21**(17):6127.
90. Dix A, Vlaic S, Guthke R, Linde J. Use of systems biology to decipher host–pathogen interaction networks and predict biomarkers. *Clin Microbiol Infect* 2016;**22**(7):600–6.
91. Abeel T, Helleputte T, Van de Peer Y, *et al.* Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* 2010;**26**(3):392–8.
92. Hastie T, Tibshirani R, Friedman JH, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Vol. **2**. New York, NY, USA: Springer, 2009.