frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 26 April 2022
doi: 10.3389/fgene.2022.821832

Novel Ensemble Feature Selection Approach and Application in Repertoire Sequencing Data

*Tao He[1], Jason Min Baik[1], Chiemi Kato[1], Hai Yang[2], Zenghua Fan[3], Jason Cham[4] and Li Zhang[2,3,5]**

[1]Department of Mathematics, San Francisco State University, San Francisco, CA, United States, [2]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, United States, [3]Department of Medicine, University of California, San Francisco, San Francisco, CA, United States, [4]Department of Medicine, Scripps Green Hospital, La Jolla, CA, United States, [5]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, United States

The T and B cell repertoire make up the adaptive immune system and is mainly generated through somatic V(D)J gene recombination. Thus, the VJ gene usage may be a potential prognostic or predictive biomarker. However, analysis of the adaptive immune system is challenging due to the heterogeneity of the clonotypes that make up the repertoire. To address the heterogeneity of the T and B cell repertoire, we proposed a novel ensemble feature selection approach and customized statistical learning algorithm focusing on the VJ gene usage. We applied the proposed approach to T cell receptor sequences from recovered COVID-19 patients and healthy donors, as well as a group of lung cancer patients who received immunotherapy. Our approach identified distinct VJ genes used in the COVID-19 recovered patients comparing to the healthy donors and the VJ genes associated with the clinical response in the lung cancer patients. Simulation studies show that the ensemble feature selection approach outperformed other state-of-the-art feature selection methods based on both efficiency and accuracy. It consistently yielded higher stability and sensitivity with lower false discovery rates. When integrated with different classification methods, the ensemble feature selection approach had the best prediction accuracy. In conclusion, the proposed novel approach and the integration procedure is an effective feature selection technique to aid in correctly classifying different subtypes to better understand the signatures in the adaptive immune response associated with disease or the treatment in order to improve treatment strategies.

Keywords: feature ensemble, VJ gene usage, repertoire sequencing data, high-dimensional data, COVID-19, adaptive immune system

OPEN ACCESS

Edited by:
Zhigang Li,
University of Florida, United States

Reviewed by:
Lei Li,
Cornell University, United States
Lin Zhang,
University of Minnesota Twin Cities, United States

****Correspondence:***
Li Zhang
li.zhang@ucsf.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *24 November 2021*
Accepted: *25 March 2022*
Published: *26 April 2022*

Citation:
He T, Baik JM, Kato C, Yang H, Fan Z, Cham J and Zhang L (2022) Novel Ensemble Feature Selection Approach and Application in Repertoire Sequencing Data. Front. Genet. 13:821832. doi: 10.3389/fgene.2022.821832

INTRODUCTION

The adaptive immune system is responsible for recognizing and eliminating antigens originating from infection and disease. It recognizes antigens via an immense array of antigen-binding antibodies (B-cell receptors, BCRs) and T-cell receptors (TCRs), the immune repertoire. The interrogation of immune repertoires is highly relevant for understanding the adaptive immune response in autoimmunity, malignancy, and infection (Miho et al., 2018). Adaptive immune receptor repertoire sequencing (Rep-seq) has driven the quantitative and molecular-level profiling of immune

repertoires, thereby revealing the high-dimensional complexity of the immune receptor sequence landscape. The advancement in high-throughput next-generation sequencing (NGS) technology has allowed researchers to sequence the immune repertoire profile from a single sample of blood or tissue.

Identification of prognostic and predictive features among high-throughput sequencing data is of high clinical relevance. Because individuals share almost no exact TCR/BCR nucleotide sequencing, TCR/BCR sequencing cannot be directly compared between different patient groups on the clonal level. However, TCR and BCR are the products of somatic V(D)J gene recombination, plus the addition/subtraction of nontemplated bases at recombination junctions. Thus, individuals share V(D)J genes, which allows for direct comparison of V and J gene usage across different patients. Therefore, it will enable researchers to directly obtain statistical inferences across subjects to provide insight into TCR/BCR repertoire with clinical characteristics and outcomes.

Though preliminary analysis using Random Forest reveal has been used to identify differentially expressed VJ genes in distinct disease types such as melanoma and prostate cancer (Cham et al., 2020), it is limited by the instability of feature selection due to the small sample size and sporadic gene usages. It has been shown that selecting the right set of features for classification and/or prediction can improve the performance of supervised and unsupervised learning, reduce computational costs such as training time or required resources, and mitigate the curse of dimensionality in the case of high-dimensional input data. Computing and using feature importance scores are also necessary steps towards model interpretability. In this paper, we introduce an ensemble feature selection strategy to select the significant V and J genes that can distinguish subjects in different groups defined by clinical characteristics, clinical treatment, or outcomes. Ensemble learning combines the results from multiple approaches, instead of simply using a single method, built on the rationale of "two heads are better than one." However, it has been primarily used in the classical prediction task of machine learning and has successfully proven its effectiveness. For example, boosting (Schapire and Freund, 2013) and bagging (Breiman, 1996) (the Random Forest is a particular case of bagging) are two popular machine learning algorithms based on the ensemble idea, where aggregating multiple tree learners make the final prediction. Recently, it has become more and more popular to use pseudo-variables (e.g., permutation copies, knockoff copies) to assist variable selection, where artificial variables (independent of the response variable) will be generated (Candès et al., 2018). The advantage of introducing pseudo-variables is that they can help reduce the false-positive rate because they are designed to be inactive and provide additional information. Here, we considered implementing ensemble learning to improve the performance of feature selection based on pseudo-variables. Simulation studies were conducted to evaluate the efficiency and accuracy of the proposed procedure in addition to real data analysis by comparing our approach with current feature selection approaches.

MATERIALS AND METHODS

European COVID-19 Data

The TCR-seq data used includes a cohort of patients who recovered after COVID-19 with mild to moderate disease courses (n = 19) and a cohort of age-matched healthy donor cohort (n = 39) that tested negative for COVID-19 antibodies. The clinical characteristics of the patients and sequencing information were shown in (Schultheiß et al., 2020) (gateway.ireceptor.org; Study ID: IR-Binder-000001). The median number of unique clonotypes was 9,431 (ranging from 589 to 35065) for healthy donors. Recovered patients had a median read depth of 72,152 ranging from 21,683 to 290,424. There was a total of 708 unique VJ gene combinations across both cohorts. VJ gene usage was defined as the number of clonotypes that utilize a particular combination of V and J genes normalized by the number of unique clones. **Table 1** presents the summary statistics for the TCR sequences.

Lung Cancer Data

The 686 TCR VJ gene combinations of 50 non-small cell lung cancer (NSCLC) patients receiving durvalumab enrolled in a Phase I trial (NCT01693562, 14 September 2012) were included for analysis. The median number of unique clonotypes was 4,994 (ranging from 403 to 17,876). In order to explore the treatment effect, here we considered using $\log_2$ transformed ratio of VJ gene usage after the treatment vs. the usage at baseline, where the VJ gene usage is defined as above. The clinical characteristics of the patients and sequencing information were shown in (Naidus et al., 2021). **Table 1** presents the summary statistics for the TCR sequences.

Simulation Strategy

We use a modified version of the simulation strategy as in (Degenhardt et al., 2019). The binary outcome is simulated based on a logistic regression model

$$log\frac{Pr(Y=1)}{Pr(Y=0)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

The three base variables (x_1, x_2, and x_3) and three additional variables (x_4, x_5, and x_6) are independently sampled from a uniform distribution of (0,1). The correlated predictor variables are simulated based on

$$v_i^{(j)} = x_i + (0.01 + \frac{0.5(j-1)}{n_i - 1}) \times N(0, 0.3)$$

for $j = 1, \ldots, g_i$ and $i = 1, \ldots, 6$, where $v_i^{(j)}$ denotes the j th variable in group i and n_i is the size of group i. The correlation between the base variable x_i and $v_i^{(j)}$ decreased as j increases. The additional predictor variables that are uncorrelated with any of those base variables and each other, w_k, $k = 1, \ldots, (G - \sum_{i=1}^{6} g_i)$, are also simulated based on a uniform distribution of (0,1), where G is the total number of the genes. Please note that $x_i, i = 1, \ldots, 6$, are only used to simulate correlated variables $v_i^{(j)}$, and are not included for feature selection and classification. $v_i^{(j)}$ $j = 1, \ldots, g_i$ and $i = 1, 2, 3$, are the causal variables, while $v_i^{(j)}$ $j = 1, \ldots, g_i$ and

TABLE 1 | Summary of TCR sequences in the real datasets.

European COVID-19 data [median (range)]	Recovered COVID patients (n = 19)		Healthy donors (n = 39)	
Number of unique clonotypes	17441 [6073,35065]		7952 [589,15271]	
Clonal counts	185758 [45066, 251020]		62429 [21683, 290424]	
VJ gene usage	1.1×10^{-3} [0, 0.816]		0.4×10^{-3} [0, 0.283]	
VJ gene usage for the selected 11 genes	0.010 [0, 0.816]		0.002 [0, 0.195]	
Lung Cancer Data (median [range])	Longer survivors (n = 17)		Shorter survivors (n = 33)	
	Baseline	Post-treatment	Baseline	Post-treatment
Number of unique clonotypes	6,144 [1,104,17876]	5,920 [403,13039]	4,708 [840, 13200]	3,737 [943,13839]
Clonal counts	206440 [1543567,3994587]	2028347 [1502019, 2718355]	2322713 [1483854,6956035]	2282314 [1348433, 7944974]
$\log_2$ (ratio of VJ gene usage)	0 [-16.31,15.60]		0 [-15.96, 16.00]	
$\log_2$ (ratio of VJ gene usage) of the selected 9 genes	0 [-13.14,12.88]		0 [-12.59,12.69]	

TABLE 2 | Simulation scenario.

Label	Sample size	# of genes	P(Y = 1)	Sparsity (# of causal genes/# of genes)	β_1	β_2	β_3
n50_G600_eta0.5	50	600	0.5	30/600	–9	6	3
n50_G1200_eta0.5	50	1,200	0.5	30/1,200	–9	6	3
n100_G600_eta0.5	100	600	0.5	30/600	–9	6	3
n100_G1200_eta0.5	100	1,200	0.5	30/1,200	–9	6	3
n50_G600_eta0.25	50	600	0.25	30/600	–14	12	–6
n50_G1200_eta0.25	50	1,200	0.25	30/1,200	–14	12	–6
n100_G600_eta0.25	100	600	0.25	30/600	–14	12	–6
n100_G1200_eta0.25	100	1,200	0.25	30/1,200	–14	12	–6

$i = 4, 5, 6$ and w_k, $k = 1, \ldots,$ $(G - \sum_{i=1}^{6} g_i)$ are the non-causal variables.

We consider several different simulation scenarios (**Table 2**) including 1) different prevalence of the binary outcome ($\eta = 0.25$ and 0.5) which is mainly determined by the coefficients in the logistic regression; 2) sparsity of causal genes (2.5%, 5%) with a different number of genes (G = 600 and 1,200); and 3) different sample sizes ($n = 50$ and 100). Under each scenario, the G genes consist of 30 causal ones $\{v_i^{(j)}, i = 1, 2, 3; j = 1, \ldots, 10\}$ as well as 30 correlated, non-causal variables $\{v_i^{(j)}, i = 4, 5, 6; j = 1, \ldots, 10\}$ and $G - 60$ uncorrelated, non-causal variables $\{w_k, k = 1, \ldots, G - 60\}$. For each of the scenarios,100 paired replicates are simulated. Each time the first one is used for feature selection and training the classifier, and the second is used for assessing stability and estimating prediction performance.

Existing Approaches of Feature Selection

Feature selection methods are often categorized into four classes: filters, wrappers, embedded, and hybrid methods. Filter methods evaluate and rank the importance of a single feature (univariate filter) or an entire subsect of features (multivariate filter) based only on their inherent characteristics, without incorporating any learning algorithm. Wrapper methods evaluate a specific subset of variables by training and testing a specific learning model (e.g., K-Nearest Neighbors (KNN) (Dudani, 1976) or Support Vector Machine (SVM) (Suykens and Vandewalle, 1999)). However, as the space of variables subset grows exponentially with the number of variables, the exhaustive search is very computationally intensive. Two alternative search schemes are commonly used to guide the search: sequential search, such as forward selection (add one at a time) or Recursive Feature Elimination (RFE, eliminate one at a time), and randomized search. Embedded methods consist of algorithms that simultaneously perform model fitting and feature selection. This approach is typically implemented using a sparsity regularizer or constraint on regression modeling, which shrinks the weight of some features to zero. Hybrid methods start with an initial feature filtering based on statistical properties, followed by a second selection based on wrapper methods. In this paper, we evaluate a variety of feature selection methods, including information gain (Kent, 1983) (univariate filter), correlation-based feature selection (Hall, 2000) (multivariate filter), SVM-RFE (Duan et al., 2005) (wrapper), Boruta (Kursa and Rudnicki, 2010) (wrapper), Vita (Malley et al., 2012) (wrapper), and LASSO (Tibshirani, 1996) (embedded). In addition, Boruta and Vita are built around Random Forest classifier and Random Forest is a bagging technique, therefore, Boruta and Vita can also be considered as feature selection approaches using bagging technique. The detailed information on those methods is

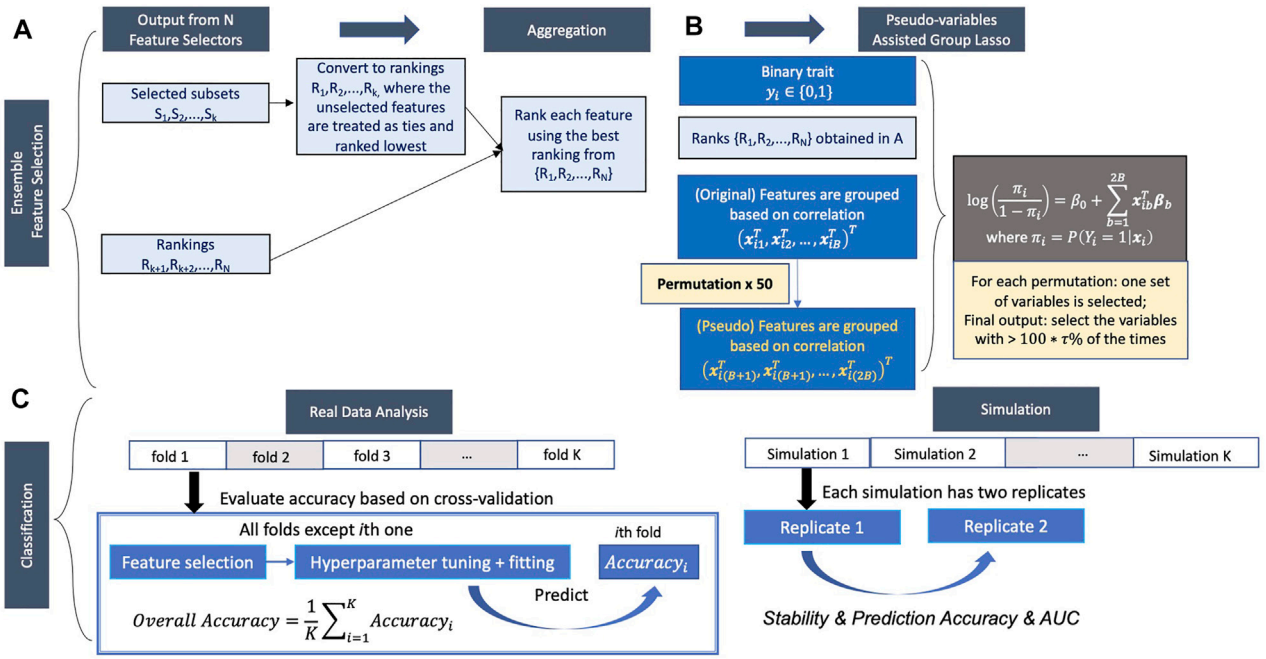

FIGURE 1 | Proposed pipeline. **(A,B)** Ensemble feature selection. **(A)** Feature selection based on different methods and aggregation of selected features. **(B)** Pseudo-variables assisted group lasso. **(C)** Classification.

provided in **Supplementary Table S1**. In addition to those listed above, there are a vast of feature selection methods existing in the literature. However, most of the time, each method selects different features, and it is difficult (almost impossible) to make a correct choice. Moreover, for small sample data (typical case for immune repertoire data), the selection based on a single feature selection method is usually not stable.

A Novel Ensemble Feature Selection

Ensemble feature selection by combing the outputs from different feature selection methods can solve the problem mentioned above. Here, we propose a new ensemble feature selection procedure based on pseudo-variables, which has two major steps: 1) aggregating the feature selection results from multiple feature selectors (**Figure 1A**) and 2) fitting a group lasso model on the candidate feature set with a new permutation-assisted tuning strategy (**Figure 1B**). In the first step, we further expand to highly correlated features to generate a candidate set of features.

Aggregating the results from different feature selection approaches is a critical step in ensemble learning. The outputs of the different approaches can be various, either the subsets of selected features, the rankings of all features, or both. We consider the following general scheme to obtain the candidate feature set depending on the types of outputs (**Figure 1A**). Suppose the feature selection approach returns the subset output. In that case, the selected features will be first converted to ranking, where the selected features are treated as ties (unless there is order in output) and ranked highest (tied for the first place), and the unselected features are also treated as ties but rank lowest using the total number of features. The highest rank across the approaches (i.e., the best position that the feature achieved) is used to generate the aggregated ranking across all feature selection approaches for each feature. For example, if one feature ranks first and 10th in two approaches, first will be recorded as the aggregated ranking for this feature.

Now we introduce a group lasso model (Meier et al., 2008) on the candidate feature set (**Figure 1B**). Denote the total number of features included in the candidate feature set after aggregation and expansion is p. Assume observations after the aggregation and expansion are $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$, where $\boldsymbol{x}_i$ is of p-dimensional vector and $y_i \in \{0, 1\}$ is a binary outcome. Without loss of generality, assume the p features are quantitative variables, but the method can be applied for categorical variables or a mixed type. By using the correlation structure of the p variables, we can define blocks $1, 2, \ldots, B$ such that within each block, the absolute value of pairwise correlation is all greater than a self-correlation threshold parameter ρ_T. Assume bth block includes L_b variables and $\sum_{b=1}^{B} L_b = p$. The p-dimensional vector $\boldsymbol{x}_i$ can be rewritten as $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T, \ldots, \boldsymbol{x}_{iB}^T)^T$ with $\boldsymbol{x}_{ib}$ of dimension L_b, $b = 1, \ldots, B$. We model the relationship between the binary Y_i and features $\boldsymbol{x}_i$ using the following logistic regression model

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{b=1}^{B} \boldsymbol{x}_{ib}^T \boldsymbol{\beta}_b \triangleq \gamma_{\boldsymbol{\beta}}(\boldsymbol{x}_i),$$

with

$$\pi_i = P(Y_i = 1|\boldsymbol{x}_i),$$

where β_0 is the intercept, and $\boldsymbol{\beta}_b \in R^{L_b}$ is the parameter vector for bth block. Denote the complete parameter vector by $\boldsymbol{\beta} =$

$(\beta_0, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_B^T)^T \in R^{p+1}$. In the repertoire-sequencing data, our focus is to identify the crucial groups of VJ genes associated with the binary outcome, i.e., which $\boldsymbol{\beta}_b \neq 0$. The main challenge here is that the dimension of the VJ genes of repertoire data is usually very high (~1,000) compared to the sample size (20–50). Despite the high dimensionality, we assume that only a small number of VJ gene groups impact the phenotype (i.e., a sparse model). Hence, the group-lasso-type methods (Meier et al., 2008) fit the scenario well because of their ability to shrink some of the coefficient vectors to precisely zero. The important VJ gene sets with notable effects on the phenotype will stand out. The estimation of the complete parameter vector $\boldsymbol{\beta}$ is given by minimizing the following objective function

$$Q_\lambda(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \lambda \sum_{b=1}^{G} s_b \|\boldsymbol{\beta}\boldsymbol{b}_{\|2},$$

where $l(\boldsymbol{\beta})$ is the log-likelihood function, i.e.,

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \gamma_\beta(\boldsymbol{x}_i) - \log\left[1 + \exp\left\{\gamma_\beta(\boldsymbol{x}_i)\right\}\right],$$

And λ is the tuning parameter that controls the amount of shrinkage (larger lambda shrinks more to zero). The s_b is used to rescale the penalty to each group and its default setting in group lasso methods was $\sqrt{L_b}$. To put a small penalty on top-ranked feature sets, we propose using the product of the minimum rank among different feature selectors and $\sqrt{L_b}$.

The selection of the tuning parameter λ of the group lasso model is typically performed by maximizing the cross-validation error (an estimate of prediction accuracy). However, the cross-validation error has a considerable variation when the sample size is small and potentially leads to less reliable conclusions (Varoquaux, 2018). The objective of this study is more about selecting the important features than improving the prediction accuracy. Therefore, we propose to use pseudo-variables (Candès et al., 2018) to facilitate the group-lasso tuning parameter selection. Let $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T = (X_1, X_2, \ldots, X_p)$ be the original input features (VJ gene usage) matrix. The pseudo-variables matrix X^π is generated through permutation, i.e., $X^\pi = (\boldsymbol{x}_{\pi(1)}, \boldsymbol{x}_{\pi(2)}, \ldots, \boldsymbol{x}_{\pi(n)})^T$, where $\{\pi(1), \pi(2), \ldots \pi(n)\}$ is a permutation of {1,2,. . .,n}. Combining the original and permutated design matrixes, we define the augmented design matrix as $X^A = (X, X^\pi)$ of n rows and 2 p columns, where the first p columns are original input features and second p columns are pseudo features (but preserve the correlation of the structure of original features). The optimization problem corresponding to group lasso fitting after augmentation becomes

$$\boldsymbol{\beta}^A(\lambda) = \text{argmin}_{\beta^A} \; -l\left(\boldsymbol{\beta}^A\right) + \lambda \sum_{b=1}^{2B} s_b \left\|\boldsymbol{\beta}_{A,b\|2}^{\boldsymbol{T}}\right.$$

where $\boldsymbol{\beta}^A = (\beta_{A,0}, \boldsymbol{\beta}_{A,1}^T, \ldots, \boldsymbol{\beta}_{A,B}^T, \boldsymbol{\beta}_{A,B+1}^T, \ldots, \boldsymbol{\beta}_{A,2B}^T)^T \in R^{2p+1}$ is the regression coefficients vector including the intercept, B sets of original VJ genes and B sets of pseudo-VJ genes. Given a tuning parameter λ, the estimated $\hat{\boldsymbol{\beta}}^A(\lambda)$ can be obtained by solving a convex optimization problem. As λ increases, more blocks of coefficient vectors $\boldsymbol{\beta}_{A,b}$ shrink to zero (i.e., fewer groups remain in the model), and the most important group shrinks lastly. For each (either original or pseudo) set, define $R_b = \sup\{\lambda: \boldsymbol{\beta}_{A,b}^T \neq 0\}$, $b = 1, \ldots, 2B$, which can be viewed as an importance measure of the feature set. The larger R_j is, the more important the set is. Then define $T_\pi = \max_{(B+1)\leq b\leq 2B} R_b$, which can be utilized to separate the active features from the inactive artificial ones. Based on the value "benchmark" T_π, the selection for each permutation can be made with $\hat{S}_\pi = \{b: R_b > T_\pi, b = 1, \ldots, B\}$, i.e., selecting the original sets which are more important than the artificial ones. Repeat this process for K times and report the feature sets selected more than a certain percentage threshold τ (e.g. 50%).

Integrated Feature Selection and Classification Pipeline

We then feed the selected features (based on six existing methods and the novel ensemble feature selection approach) into eight different classifiers, including SVM with linear (SVM linear), polynomial (SVM poly) and radius kernels (SVM rad) (Amari and Wu, 1999), K-nearest neighbors (KNN) (Dudani, 1976), Random Forest (Breiman, 2001), extreme gradient boosting (XGB) (Chen et al., 2015), ridge (Le Cessie and Van Houwelingen, 1992) and LASSO (Tibshirani, 1996).

Performance Evaluation

For simulation studies, we assess and compare the performance of the different variable selection approaches by using the following measures: false discovery rate (FDR), sensitivity, stability, F-1 score (Hua et al., 2009), and empirical power (**Figure 1C**). For each method, within each replicate, FDR is calculated as the ratio of the number of false-positive results, i.e., the total number of non-causal variables ($v_i^{(j)}$ $j = 1, \ldots, g_i$ and $i = 4, 5, 6$ and w_k, $k = 1, \ldots, (G - \sum_{i=1}^{6} g_i)$) selected to the total number of variables selected. In contrast, sensitivity is defined as the proportion of correctly identified causal variables ($v_i^{(j)}$ $j = 1, \ldots, g_i$ and $i = 1, 2, 3$) among all causal variables per replicate and method. F-1 score is calculated as 2*(precision*sensitivity)/(precision + sensitivity), a balance (the harmonic mean) of precision and sensitivity, where precision = 1-FDR (Hua et al., 2009). For each pair of replicates, the Jaccard's index is calculated as the ratio of the length of the intersection and the length of the union of the two sets of selected variables (He and Yu, 2010). The average across all pairs is used to quantify the stability of variable selection for the particular method (Kalousis et al., 2007). The empirical power of each causal variable ($v_i^{(j)}$ $j = 1, \ldots, g_i$ and $i = 1, 2, 3$) is calculated as the frequency of correct selections among all replicates (Dash and Liu, 1997). The prediction accuracy and area under the curve (AUC) are assessed on the paired replicate to evaluate the performance of the classification (Huang and Ling, 2005). The parameters we used were listed in **Table 3**.

In the real data analysis, we use 5-fold cross-validation to evaluate the prediction accuracy and AUC (**Figure 1C**). Feature selection, model fitting, and parameter tuning are performed using the four folds of the data, and prediction accuracy and AUC are evaluated by averaging the values on the held-out fold data. However, because the causal variables in real data are unknown, we can't assess FDR, sensitivity, F-1 score, and empirical power of

TABLE 3 | Parameters used for feature selection methods.

Approach	R Package	Parameter	Description	Value
Information gain	FSelector	k	Select top k features	0.05 × G
CFS	FSelector	default		
Boruta	Boruta	Final Decision	Three possible values as the final decision	Confirmed /tentative
Vita	Vita	k	Cross-validation fold	5
		pvalue threshold	Selection criteria of pvalue	0
SVM_RFE	mSVM-RFE	k	Cross-validation fold	5
		selection criteria	Top features	0.05 × G
LASSO	glmnet	lamba	Tunning parameter grid values	$10^{(-10,-9.9,\ldots,0,\ldots,9.9,10)}$
Ensemble		ρ_T	Minimum pairwise correlation within block	0.75
		K	Total number of permutations	50
		τ	Threshold of selection percentage	0.5

G is the total number of the features.

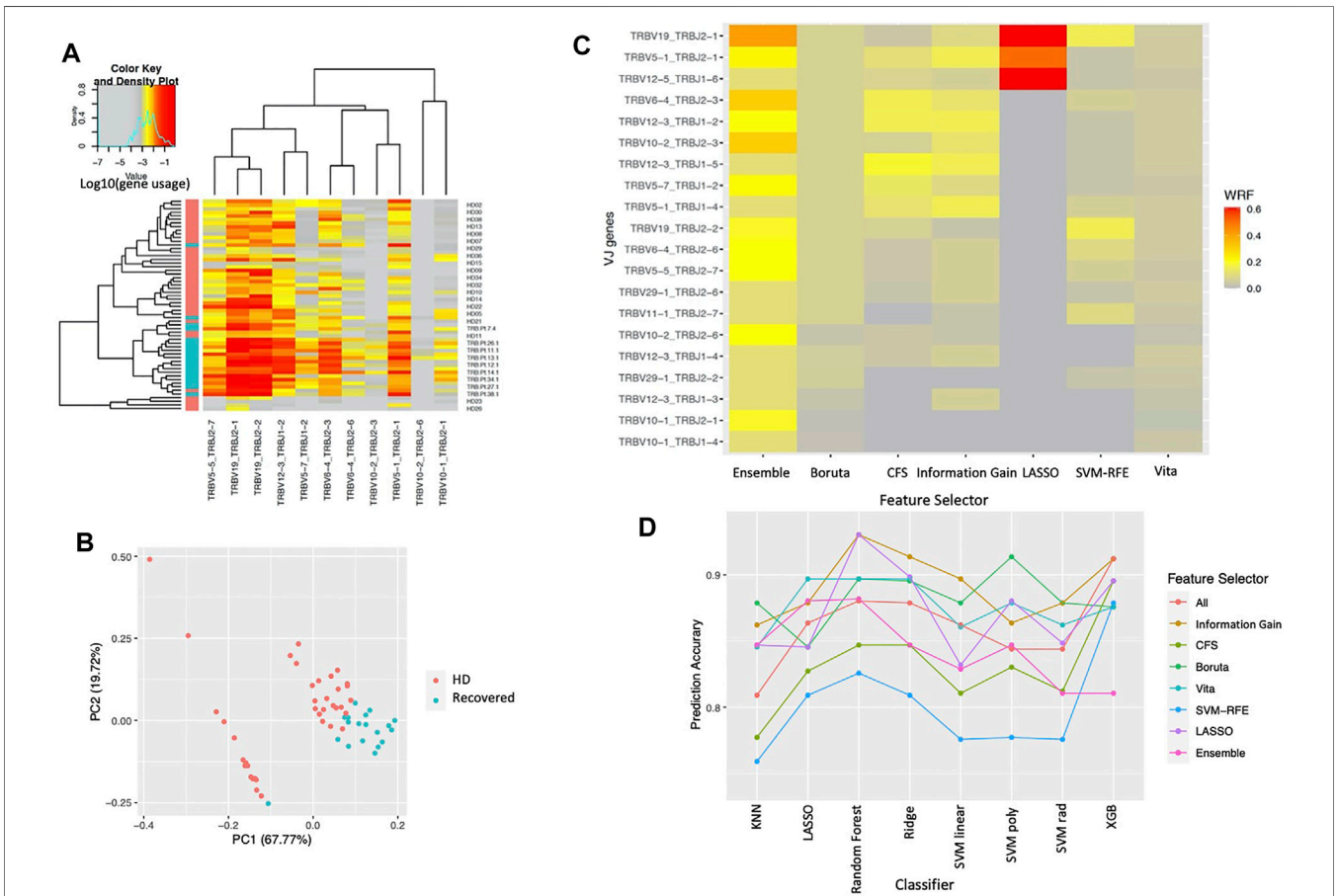

FIGURE 2 | The results for COVID-19 dataset. **(A)** TCR VJ gene usage heatmap. VJ gene usage was assessed for each sample. Each column represents an individual single VJ gene combination, and each row represents an individual subject with red bar and green bar in left side of the heat map representing healthy donors (HD) and COVID-19 recovered patients (Recovered), respectively. The heat represents the VJ gene usage, with red to grey representing increased to decreased gene usage, respectively. **(B)** Principal components analysis based on the selected VJ genes. Each dot represents a single subject with red and green representing healthy donors (HD) and COVID-19 recovered patients (Recovered), respectively. **(C)** Normalized selection frequency of the top selected VJ genes by each feature selection approach. Each row represents an individual single VJ gene combination, and each column represents the feature selection approaches. **(D)** Prediction accuracy based on the selected VJ genes by each feature selection approach for different classification methods. Each line presents the prediction accuracy for the corresponding feature selection approach based on the corresponding classification method (x-axis).

the feature selection approaches. The relationship of features selected in different feature selection methods was investigated, and the most frequently selected features in each fold among all methods for both datasets were also evaluated. Considering each feature selection output varies in the number of features selected, we used a weighted relative frequency (WRF) to measure the relative frequency that a feature is selected across five different folds. Specifically, $WRF_j = \sum_{f=1}^{5} I\{X_j \in S_f\}/n_f$ for j th feature, where S_f is the set of selected features using all but f th fold data and n_f is the total number in S_f. $I\{X_j \in S_f\}$ is an indicator function which takes value 1 if feature X_j is one of the selected features and takes value 0 if not. For example, if the five selections are $\{X_1, X_2\}$, $\{X_2, X_4, X_5\}$, $\{X_1, X_5\}$, $\{X_1\}$ and $\{X_2, X_5, X_7\}$, then $WRF_1 = 1/2 + 0/3 + 1/2 + 1/1 + 0/3 = 2$.

All the analyses were performed by R (https://www.r-project.org).

RESULTS

Ensemble Feature Selection Approach Efficiently Selected Key Features on Real Data Analysis

We classified clonotypes into 708 VJ gene combinations and assessed whether VJ gene usage within the T-cell repertoire differed between the two cohorts in the COVID-19 dataset. Our proposed ensemble method shows the gene usage of 11 VJ genes that were all significantly higher in the COVID-19 recovered patients compared to their usage in healthy donors (**Figure 2A**). The 11 genes are selected at least twice in the 5-folds cross-validation procedure (**Table 1**), including a more significant increase in the TCRV5-1/J2-1, V5-5/J2-7, V6-4/J2-3, V12-3/J1-2, V19/J2-1, and V19/J2-2 gene usages in COVID-19 recovered patients, which were reported in the original paper (Schultheiß et al., 2020). The principal components analysis (PCA) based on the 11 selected VJ genes shows that the two cohorts can be segregated mainly by these 11 VJ usages (**Figure 2B**). In addition, we compared the selection frequency across the different feature selection approaches by cross-validation. We did 5-fold cross-validation within each approach and calculated the weighted relative frequency for each gene. **Figure 2C** shows the heatmap of the top 20 selected genes ordered by their WRFs based on the ensemble method, for different feature selection methods. It can be observed that lasso can only identify the top signals with strong signals (hence those variables have large WRFs), and the proposed feature ensemble method can aggregate the top signals identified by the existing approaches. **Figure 2D** presents the prediction accuracy of the selected genes based on the different feature selection methods for different classifiers. We found that all feature selection approaches (including the feature ensemble method) have very similar prediction accuracy in terms of classification performance comparing to the results without feature selection.

In addition, we considered to identify the VJ genes in the lung cancer dataset based on the patients' overall survival time. There are 17 longer survivors and 33 short survivors, which were defined based on longer or shorter than the median overall survival (20.3 months), respectively. Because the lung cancer patients received durvalumab, we selected the VJ genes based on their usage changes from baseline to the post-treatment, which was defined as the ratio of VJ gene usages from post-treatment vs. the usages from baseline. We identified 9 genes: TRBV5-3/J1-1, TRBV1/J2-7, TRBV1/J1-5, TRBV20-1/J1-4, TRBV7-4/J2-3, TRBV11-1/J2-6, TRBV7-7/J2-2, TRBV1/J1-1, and TRBV5-7/J1-6 when long survivors compared to short survivors (**Supplementary Figure S1**).

Ensemble Feature Selection Approach Consistently Outperformed on Simulation Studies

In general, the ensemble feature selection approach consistently outperforms the other state-of-the-art feature selection methods in terms of both stability and accuracy. It possesses consistent higher stability and sensitivity but lower FDR independent of the sample size choices, the sparsity of the causal genes, and the prevalence of the outcomes (**Figure 3**). As expected, a larger sample size increases the stability, sensitivity and F-1 score, but almost didn't change FDR. Interestingly, a higher outcome prevalence results in lower stability for most methods except lasso and CFS, higher FDR and lower sensitivity. More genes in the pool introduce less stability, slightly more FDR and almost no change in sensitivity. Together with LASSO, the ensemble method is relatively robust to the choices of sample size, the sparsity of the causal genes, and the prevalence of the outcomes in terms of F-1 score, sensitivity and FDR. However, the performance of LASSO is always much worse than the ensemble method. In addition, the proposed ensemble approach always maintains the largest power (close to 1) in all simulation scenarios while some approaches could have as low as less than 50% of power (**Figure 4**). And the power that the ensemble approach can achieve is robust to the number of causal variables in the simulation, unlike the traditional approaches, the power is significantly impacted by the number of the causal variables.

Overall, the ensemble feature selection also improves the classification performance. The ensemble feature selection approach has the best prediction accuracy when integrating with LASSO, Random Forest, ridge, and SVM classification methods. While combing with KNN, the ensemble feature selection approach occasionally is not as good as information gain performs, but most of the time is worse. When working with xgb, the ensemble feature selection approach has competitive performance compared to LASSO (**Figure 5**). The ensemble feature selection approach has the highest AUC except when interpreting with KNN and Random Forest, and it has competitive performance compared to information gain (**Supplementary Figure S2**). Similarly, a larger sample size or a smaller number of genes increases the prediction accuracy while a higher outcome prevalence results in lower prediction accuracy (**Figure 5**). However, the influence on AUC is relatively small for the ensemble method (**Supplementary Figure S2**).

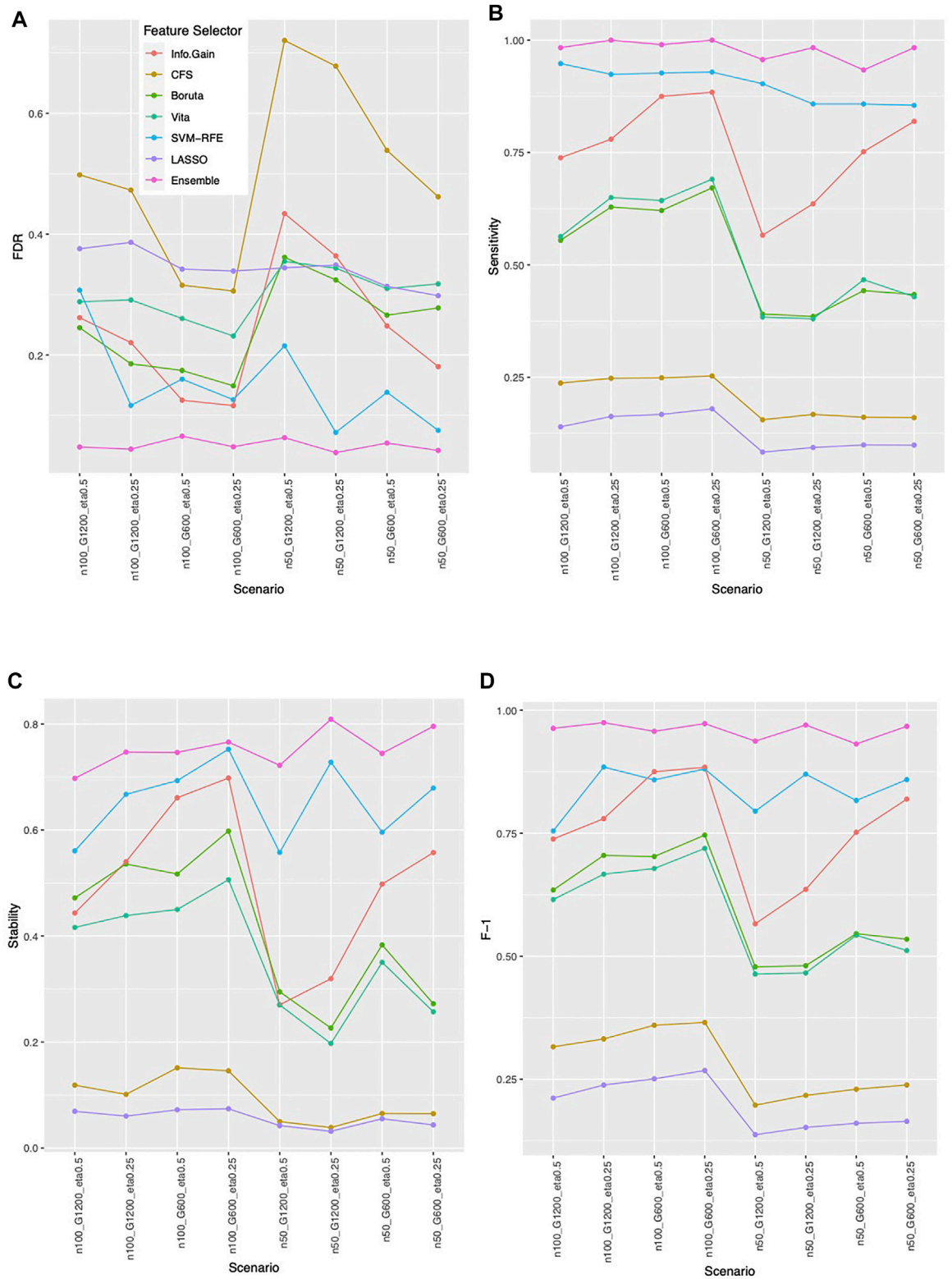

FIGURE 3 | Feature selection performance based on simulation. **(A)** FDR **(B)** Sensitivity **(C)** Stability **(D)** F-1. In each panel, x-axis stands for different simulation scenario listed in **Table 2**. For example, n50_G600_eta0.5 stands for sample size is 50 with 600 candidate genes and the probability of the outcome is 0.5. Each colored curve stands for different feature selection approaches.

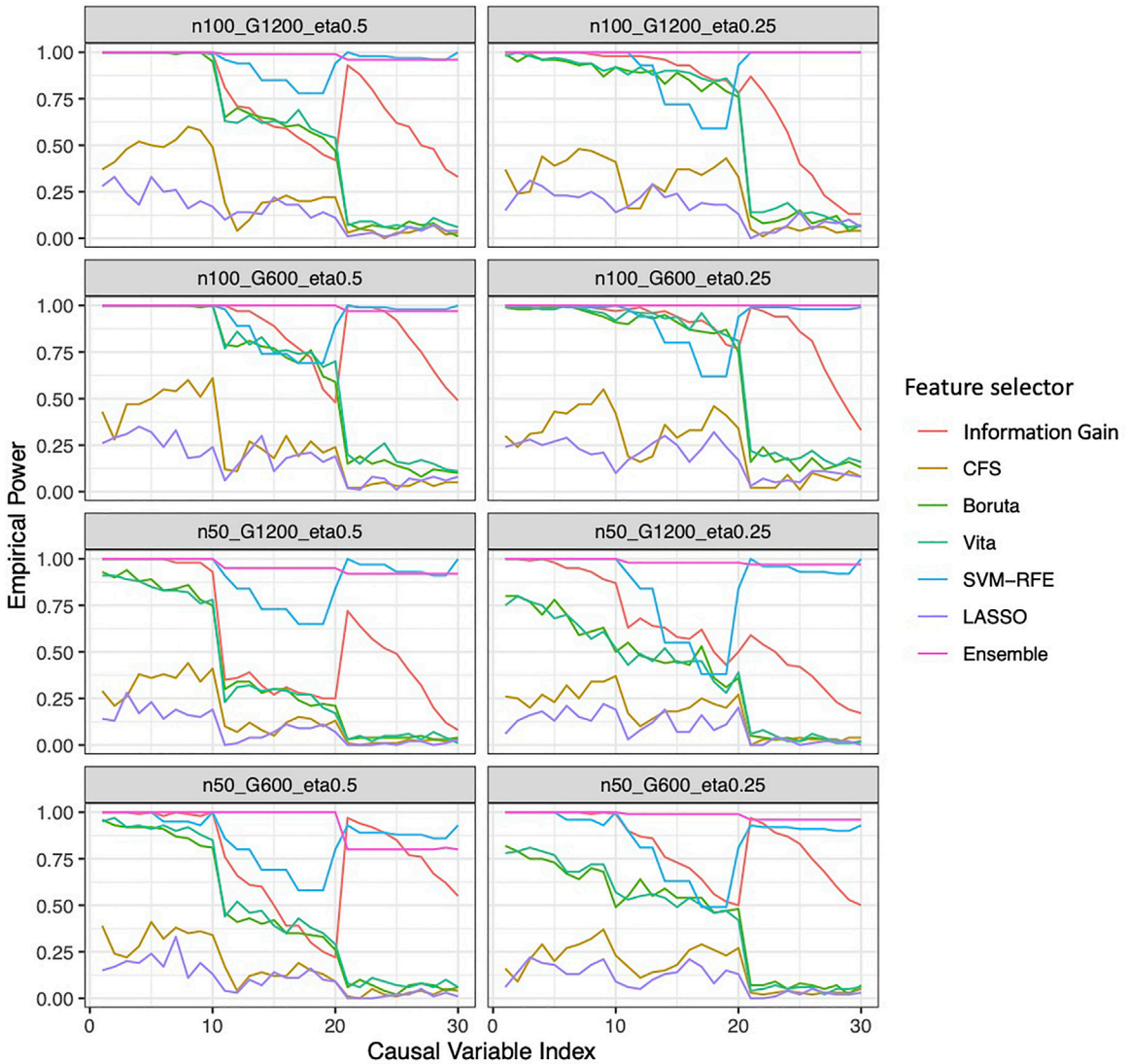

FIGURE 4 | Empirical power of the feature selection approaches based on simulation. Each panel represents different simulation scenario listed in **Table 2**. For example, n50_G600_eta0.5 stands for sample size is 50 with 600 candidate genes and the probability of the outcome is 0.5. In each panel, x-axis stands for the number of causal variables.

DISCUSSION

We formulated a novel ensemble feature selection approach with a customized statistical learning algorithm focused on VJ gene usage in repertoire-sequencing data. Using the proposed approach and algorithm, we identify the VJ genes with significantly different usage in COVID-19 recovered patients and healthy donors. Wang et al. analyzed the TCR repertoire in patients with COVID-19 using single-cell sequencing and found that the frequencies of TRAV4, TRAJ2-7, TRBV7-9, and TRBJ2-3 were significantly higher compared to healthy patients (Wang et al., 2021). We found that the TCR beta chains TCRV5-1/J2-1, V5-5/J2-7, V6-4/J2-3, V12-3/J1-2, V19/J2-1, and V19/J2-2 had higher frequencies among patients with COVID-19. Overall, identifying these VJ genes could reflect a specific antigen milieu leading to the selection of a distinct combination of VJ genes. Further correlation of these unique VJ gene profiles with clinical outcomes can potentially aid in the development of sorely needed prognostic tools for patients infected with COVID-19. Additionally, among the 9 VJ gene usages identified in lung cancer patients treated with Durvalumab, one of the identified V segments, TRBV20-1 has been previously shown to be differentially expressed in cancer tissue compared to healthy tissue (Wang et al., 2019). Furthermore, TRBV20-1 usage has been associated with improved response and survival in lung cancer patients treated with anti-PD1 therapy such as Durvalumab (Dong et al., 2021). Thus, the other identified VJ gene segment pairs above should be explored as potential additional features of the TCR repertoire associated with improved clinical response.

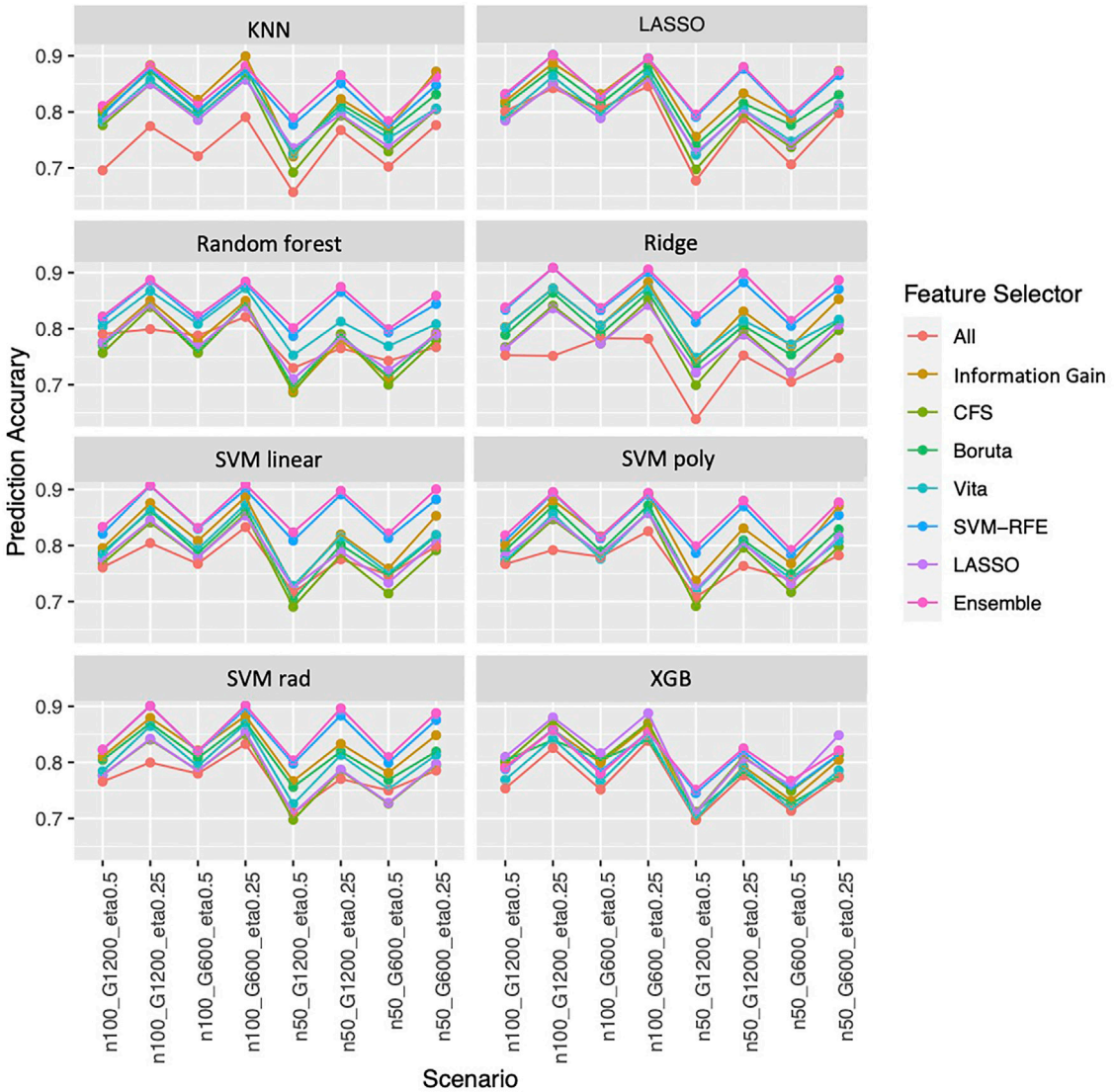

FIGURE 5 | Prediction accuracy based on simulation. Panels present the prediction accuracy for the corresponding classification approach as indicated. In each panel, x-axis stands for different simulation scenario listed in **Table 2**. For example, n50_G600_eta0.5 stands for sample size is 50 with 600 candidate genes and the probability of the outcome is 0.5. Each colored curve stands for different feature selection approaches. The eight classification approaches are: SVM with linear (SVM linear), polynomial (SVM poly) and radius kernels (SVM rad) (Amari and Wu, 1999), K-nearest neighbors (KNN) (Dudani, 1976), Random Forest (Breiman, 2001), extreme gradient boosting (XGB) (Chen et al., 2015), Ridge (Le Cessie, Van Houwelingen) and LASSO (Tibshirani, 1996).

In the real data analysis, we found that none of the feature selection approaches (including the feature ensemble method) have substantial improvements in terms of classification performance comparing to the results without feature selection. This is not surprising, since feature selection is not guaranteed to improve the prediction accuracy. However, feature selection is able to reduce the dimensinsionality and complexity of the predictive models, which eventually leads to a faster model trainning time and convergency. Our ensemble method, though is not driven by the prediction accuracy as some other feature selectors (e.g., SVM-RFE, Boruta, Vita) is still able to exhibit a competitive performance in spite of a small sample size and even with highly correlated features included.

In addition, we carry out intensive simulation studies in different scenarios. We found that the ensemble feature selection approach surpasses the other commonly used feature selection methods based on efficiency and accuracy. When integrating with varying types of classification methods, in most cases, the ensemble feature selection approach has the best prediction performance. These results indicate that the ensemble feature selection approach not only identifies the most stable, highest sensitive features with low false discovery rates but also greatly improves the prediction performance. Sample size, sparsity of causal genes, and the prevalence of the outcomes influence the performance but are relatively small for the ensemble approach.

In the similation studies shown above, the base learners used in the first phase were information gain, SVM-RFE, Vita and Boruta. We have conducted additional simulation studies, where the different base learners were used. The results shown that our proposed ensemble method is neither sensitive to the number nor to the choice of base learners (**Supplementary Table S2**). In addition, we found that the proposed method is also relatively robust to the choices of those parameters based on the simulation studies (**Supplementary Table S3**). Note that the large threshold ρ_T will introduce singleton groups and small threshold will introduce large groups, which may impact the variable selection results. Very large groups, unless very strong signal, will less likely to be included in the selection, because the proposed group lasso model penalizes based on the group size. In an extreme case of all groups are singleton, it is reduced to a regular lasso model, where only one variable might got picked among highly correlated variables. Therefore, a moderate threshold is recommended. In our case, we set the correlation threshold of 0.75 to keep the highly correlated variables in the same block while the uncorrelated variables in singleton groups.

Though the repertoire-sequencing data was used to illustrate this approach, the proposed approach can be applied to any other feature selection work. In a real application, we could perform stratified feature selection within the strata defined by the important covariates or consider the covariates as additional features. Although the ensemble feature selection was currently applied to a binary outcome, it can also be extended to different types of outcomes (continuous, multi-level categorical outcomes, and time-to-event outcomes) by changing the log-likelihood in the objective function in optimization. Not surprisingly, the proposed ensemble method, which aggregates the output from multiple feature selectors, takes longer than a single feature selection method. However, it has been demonstrated in the simulation studies that the feature selection performance can be significantly boosted. Moreover, we want to point out that the proposed method can be applied as an independent feature selection method by setting the rank matrix with all elements equal to a constant. In addition, by using parallel computing, the computational time can be dramatically decreased. Furthermore, learning in a small n large p case is always challenging due to the minimal information observed in the high dimensional space. However, our simulation shows that the performance of both feature selection and classification is still appealing.

In conclusion, the proposed novel approach and integrated procedure can help us pursue an effective feature selection technique to aid in correctly prioritizing the important features and classifying different subtypes.

DATA AVAILABILITY STATEMENT

R codes are available on GitHub (https://github.com/mlizhangx/Ensemble-Feature-Selection). The COVID data is available via gateway.ireceptor.org; Study ID: IR-Binder-000001. The lung cancer data underlying the findings described in this article may be obtained in accordance with AstraZeneca's data sharing policy described at https://astrazenecagrouptrials.pharmacm.com/ST/Submission/Disclosure.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TH and LZ concepted and designed the experiment. HY, ZF and LZ acquired the data. TH, JB and LZ performed the data analysis and simulation studies. All authors participated in the interpretation of study results, and in the drafting and approval of the final version of the manuscript.

FUNDING

JC is supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health, through grant number UL1TR002550 and linked award KL2TR002552. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. TH and LZ are partially supported by National Cancer Institute, National Institutes of Health, through grant number R21CA264381. HY and LZ are partially supported by UCSF Prostate Cancer Program 2021 Pilot Research Award. TH is partially supported by National Science Foundation through grant number DMS-2137983.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.821832/full#supplementary-material

REFERENCES

Amari, S., and Wu, S. (1999). Improving Support Vector Machine Classifiers by Modifying Kernel Functions. *Neural Networks.* 12 (6), 783–789. doi:10.1016/s0893-6080(99)00032-5

Breiman, L. (1996). Bagging Predictors. *Mach Learn.* 24 (2), 123–140. doi:10.1007/bf00058655

Breiman, L. (2001). Random Forests. *Machine Learn.* 45 (1), 5–32. doi:10.1023/a:1010933404324

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. *J. R. Stat. Soc. B.* 80 (3), 551–577. doi:10.1111/rssb.12265

Cham, J., Zhang, L., Kwek, S., Paciorek, A., He, T., Fong, G., et al. (2020). Combination Immunotherapy Induces Distinct T-Cell Repertoire Responses when Administered to Patients with Different Malignancies. *J. Immunother. Cancer.* 8 (1), e000368. doi:10.1136/jitc-2019-000368

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., and Cho, H. (2015). Xgboost: Extreme Gradient Boosting. *R. Package Version.* 1 (4), 1–4.

Dash, M., and Liu, H. (1997). Feature Selection for Classification. *Intell. Data Anal.* 1 (1-4), 131–156. doi:10.1016/s1088-467x(97)00008-5

Degenhardt, F., Seifert, S., and Szymczak, S. (2019). Evaluation of Variable Selection Methods for Random Forests and Omics Data Sets. *Brief. Bioinformatics.* 20 (2), 492–503. doi:10.1093/bib/bbx124

Dong, N., Moreno-Manuel, A., Calabuig-Fariñas, S., Gallach, S., Zhang, F., Blasco, A., et al. (2021). Characterization of Circulating T Cell Receptor Repertoire Provides Information about Clinical Outcome after PD-1 Blockade in Advanced Non-small Cell Lung Cancer Patients. *Cancers.* 13, 2950. doi:10.3390/cancers13122950

Duan, K.-B., Rajapakse, J. C., Wang, H., and Azuaje, F. (2005). Multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data. *IEEE Trans.on Nanobioscience.* 4 (3), 228–234. doi:10.1109/tnb.2005.853657

Dudani, S. A. (1976). The Distance-Weighted K-Nearest-Neighbor Rule. *IEEE Trans. Syst. Man. Cybern.* SMC-6 (4), 325–327. doi:10.1109/tsmc.1976.5408784

Hall, M. A. (2000). *Correlation-based Feature Selection of Discrete and Numeric Class Machine Learning*. (Working paper 00/08). Hamilton, New Zealand: Department of Computer Science, University of Waikato.

He, Z., and Yu, W. (2010). Stable Feature Selection for Biomarker Discovery. *Comput. Biol. Chem.* 34 (4), 215–225. doi:10.1016/j.compbiolchem.2010.07.002

Hua, J., Tembe, W. D., and Dougherty, E. R. (2009). Performance of Feature-Selection Methods in the Classification of High-Dimension Data. *Pattern Recognition.* 42 (3), 409–424. doi:10.1016/j.patcog.2008.08.001

Huang, J., and Ling, C. X. (2005). Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* 17 (3), 299–310. doi:10.1109/tkde.2005.50

Kalousis, A., Prados, J., and Hilario, M. (2007). Stability of Feature Selection Algorithms: a Study on High-Dimensional Spaces. *Knowl Inf. Syst.* 12 (1), 95–116. doi:10.1007/s10115-006-0040-8

Kent, J. T. (1983). Information Gain and a General Measure of Correlation. *Biometrika.* 70 (1), 163–173. doi:10.1093/biomet/70.1.163

Kursa, M. B., and Rudnicki, W. R. (2010). Feature Selection with theBorutaPackage. *J. Stat. Soft.* 36 (11), 1–3. doi:10.18637/jss.v036.i11

Le Cessie, S., and Van Houwelingen, J. C. (1992). Ridge Estimators in Logistic Regression. *J. R. Stat. Soc. Ser. C (Applied Statistics).* 19, 191. doi:10.2307/2347628

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., and Ziegler, A. (2012). Probability Machines. *Methods Inf. Med.* 51 (01), 74–81. doi:10.3414/me00-01-0052

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The Group Lasso for Logistic Regression. *J. R. Stat. Soc. Ser. B (Statistical Methodology).* 70 (1), 53–71. doi:10.1111/j.1467-9868.2007.00627.x

Miho, E., Yermanos, A., Weber, C. R., Berger, C. T., Reddy, S. T., and Greiff, V. (2018). Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. *Front. Immunol.* 9, 224. doi:10.3389/fimmu.2018.00224

Naidus, E., Bouquet, J., Oh, D. Y., Looney, T. J., Yang, H., Fong, L., et al. (2021). Early Changes in the Circulating T Cells Are Associated with Clinical Outcomes after PD-L1 Blockade by Durvalumab in Advanced NSCLC Patients. *Cancer Immunol. Immunother.* 70 (7), 2095–2102. doi:10.1007/s00262-020-02833-z

Schapire, R. E., and Freund, Y. (2013). Boosting: Foundations and Algorithms. *Kybernetes.* 42, 164–166. doi:10.1108/03684921311295547

Schultheiß, C., Paschold, L., Simnica, D., Mohme, M., Willscher, E., von Wenserski, L., et al. (2020). Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease. *Immunity.* 53 (Issue 2), 442–455. e4. doi:10.1016/j.immuni.2020.06.024

Suykens, J. A. K., and Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* 9 (3), 293–300. doi:10.1023/a:1018628609742

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological).* 58 (1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Varoquaux, G. (2018). Cross-validation Failure: Small Sample Sizes lead to Large Error Bars. *Neuroimage.* 180, 68–77. doi:10.1016/j.neuroimage.2017.06.061

Wang, P., Jin, X., Zhou, W., Luo, M., Xu, Z., Xu, C., et al. (2021). Comprehensive Analysis of TCR Repertoire in COVID-19 Using Single Cell Sequencing. *Genomics.* 113 (Issue 2), 456–462. doi:10.1016/j.ygeno.2020.12.036

Wang, X., Zhang, B., Yang, Y., Zhu, J., Cheng, S., Mao, Y., et al. (2019). Characterization of Distinct T Cell Receptor Repertoires in Tumor and Distant Non-tumor Tissues from Lung Cancer Patients. *Genomics, Proteomics & Bioinformatics.* 17 (Issue 3), 287–296. doi:10.1016/j.gpb.2018.10.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.