Lab Invest 104 (2024) 100320

LABORATORY INVESTIGATION

journal homepage: https://laboratoryinvestigation.org/

Research Article

Union With Recursive Feature Elimination: A Feature Selection Framework to Improve the Classification Performance of Multicategory Causes of Death in Colorectal Cancer

Fei Deng[a,*], Lin Zhao[a], Ning Yu[a], Yuxiang Lin[a], Lanjing Zhang[b,c,d,e,*]

[a] School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai, China; [b] Department of Biological Sciences, Rutgers University, Newark, New Jersey; [c] Department of Pathology, Princeton Medical Center, Plainsboro, New Jersey; [d] Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey; [e] Department of Chemical Biology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, New Jersey

ARTICLE INFO

Article history:
Received 26 March 2023
Revised 5 December 2023
Accepted 20 December 2023
Available online 28 December 2023

Keywords:
colorectal cancer
feature selection
machine learning
multicategory death causes
U-RFE

ABSTRACT

Despite the use of machine learning tools, it is challenging to properly model cause-specific deaths in colorectal cancer (CRC) patients and choose appropriate treatments. Here, we propose an interesting feature selection framework, namely union with recursive feature elimination (U-RFE), to select the union feature sets that are crucial in CRC progression-specific mortality using The Cancer Genome Atlas (TCGA) dataset. Based on the union feature sets, we compared the performance of 5 classification algorithms, including logistic regression (LR), support vector machines (SVM), random forest (RF), eXtreme gradient boosting (XGBoost), and Stacking, to identify the best model for classifying 4-category deaths. In the first stage of U-RFE, LR, SVM, and RF were used as base estimators to obtain subsets containing the same number of features but not exactly the same specific features. Union analysis of the subsets was then performed to determine the final union feature set, effectively combining the advantages of different algorithms. We found that the U-RFE framework could improve various models' performance. Stacking outperformed LR, SVM, RF, and XGBoost in most scenarios. When the target feature number of the RFE was set to 50 and the union feature set contained 298 deterministic features, the Stacking model achieved F1_weighted, Recall_weighted, Precision_weighted, Accuracy, and Matthews correlation coefficient of 0.851, 0.864, 0.854, 0.864, and 0.717, respectively. The performance of the minority categories was also significantly improved. Therefore, this recursive feature eliminationebased approach of feature selection improves performances of classifying CRC deaths using clinical and omics data or those using other data with high feature redundancy and imbalance.

Introduction

Colorectal cancer (CRC) ranks second in the number of cancer deaths in the United States and is highly heterogeneous and aggressive.[1] Numerous analyses of clinical studies have shown that most deaths from CRC can be attributed to disease progression and undertreatment.[2-4] During the complex multistage progression of CRC,[5-8] local recurrence is possible in patients with cured CRC,[9,10] and the recurrence rate in patients with stage IV CRC is even as high as 80% within 3 years after surgery.[11] Therefore, accurate classification of causes of death in (COD) CRC patients is of great importance to CRC research and management.

* Corresponding authors.
E-mail addresses: lanjing.zhang@rutgers.edu (L. Zhang), 2606897447@qq.com, dengfei@sit.edu.cn (F. Deng).

0023-6837/$ - see front matter © 2023 United States & Canadian Academy of Pathology. Published by Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.labinv.2023.100320

Machine learning (ML) has evolved rapidly in the past 15 years.[12] In the medical field, ML is widely used for diagnosis and prognosis of diseases, medical imaging and signal processing, and planning and scheduling.[13-20] Prognosis refers to the use of ML to learn the relevant information about the patient's condition and use the learned model to predict the future development of CRC recurrence or death.

With the rapid development of high-throughput sequencing technologies, comprehensive information on CRC cases can be obtained at the molecular level.[21,22] However, the multiomics data are characterized by the inherent category imbalance of medical data, that is, the number of individuals with disease progression may be much smaller than that of individuals without disease progression. When classification models are applied directly to omics data, the model learns much of redundant feature information, resulting in reduced model performance.[23,24] In addition, due to imbalance in labels' distribution,[25] the learned model is further biased (in favor of) toward the majority category. This is undesirable for medical data and clinical practice. Since minority categories often contain information about diseased samples, patient health may deteriorate or even result in death if minority categories are misclassified.[26,27] This remains a great challenge for CRC prognostication at the molecular level. Thus, there is an urgent need to identify the features in CRC omic and clinical data that can achieve both high specificity and high accuracy.

Feature selection is a data dimensionality reduction technique[28,29] that discards a large number of irrelevant and redundant features from the original feature space and retains a set of decisive/important features. It not only preserves the original feature values but also reduces the complexity of the model, improves its learning rate, and increases its performance. If feature selection is not performed, it may lead to the curse of dimensionality, which greatly reduces the efficiency of the model. As technology advances, the increasing size and complexity of genomic data have led to the development of additional feature selection methods. These methods have improved the accuracy, efficiency, and interpretability of feature selection.[30,31]

Linear analysis is the earliest feature extraction method, including Linear Discriminant Analysis (LDA).[32] Therefore, LDA can not only achieve data dimensionality reduction[33,34] but also achieve data classification.[33] Its core idea is the principle that the error within the category is the smallest and the error between the categories is the largest. The matrix eigenvalues determined by the intraclass covariance matrix and the interclass covariance matrix provide an entropy ranking, and the top eigenvalues and the corresponding eigenvectors are the solution results of the LDA algorithm.[35] Lee et al[36] preprocessed the data collected by the medical internet of things based on the LDA model to solve the problem of high computation cost caused by high data dimensions. However, the linear algorithm will inevitably lead to information loss while reducing the data dimension and at the same time lose the biological meaning of the original data, which is not conducive to analyzing the results based on the data and understanding the essential problems reflected by the data. In particular, these methods are often powerless in the face of complex nonlinear data.

In order to solve the problem of dimensionality reduction of high-dimensional nonlinear data, a series of feature extraction methods for nonlinear data, including kernelization dimensionality reduction algorithm,[37,38] popular learning algorithm,[39] data dimensionality reduction based on neural network,[40] Adaptive encoders (Autoencoder),[41] etc., have been proposed. These feature extractions based on ML techniques have made great progress in practical applications. However, nearly all of them used a single algorithm to select features for binary or multicategory classification. The ensemble feature selection methods were rarely, if at all, used for multicategory or patient-outcome classification. It is unclear whether and how we can combine multiple algorithms to help select features for the multicategory classification of patient outcomes. Therefore, we hypothesize that adding the features that are deemed important by various algorithms or removing the less important features would improve the performance metrics of ML algorithms in the multicategory classification of patient outcomes.

To test this hypothesis, we develop and evaluate a 2-stage Union with Recursive Feature Elimination (U-RFE) feature selection framework. The TCGA CRC data were used, as an example, to identify the features required for reaching the best multicategory COD classification performance. U-RFE first selects the union feature sets that are crucial in CRC progression-specific mortality. Based on the union feature sets, we compared the performances of 5 classification algorithms, including logistic regression (LR), support vector machines (SVM), random forest (RF), eXtreme gradient boosting (XGBoost), and Stacking, to identify the best model for classifying four-category deaths. Specifically, in the first stage, different ML algorithms would be used as base estimators for feature selection, generating different feature subsets and fetching the shared ones of these feature subsets. In the second stage, feature analysis was performed on feature subsets, considering the respective advantages of different ML algorithms. The intersection of the feature subsets (ie, shared features) was first used as the base set, and then, the other features contained in the subset were sorted and added to the base set sequentially. Based on this changing feature set, the LR, SVM, RF, XGBoost, and Stacking algorithms were used to classify the multicategory COD of the patients in TCGA Program dataset. Then, we evaluate the multimetric performance of each model as described before,[42] until the set with the best classification performance was found and determined as the final decisive union feature set.

Methods

The COD classification prediction model for CRC data consists of 2 main processes: data preprocessing and model tuning based on U-RFE. Data preprocessing consists of 4 steps: sample deletion, one-hot encoding processing, missing value processing, and determination of classification labels. The model tuning based on U-RFE consists of 3 steps: RFE feature selection, the U-RFE feature selection, and optimization of classification models.

Our CRC dataset was downloaded from cBioPortal.org in February 2019. All the above analyses for CRC were run on python 3.6 version. Among them, RF, LR, SVM, and grid search were from scikit-learn 0.24.2, and XGBoost was from package xgboost1.5.2. The main workflow of the analyses is shown in Figure 1.

Data Preprocessing

We obtained individual-level data for CRC (Pan-Cancer Atlas) from the TCGA.[43] Additionally, TCGA data are deidentified and publicly available. Therefore, this is an exempt study using publicly available deidentified data and did not require an Institutional Review Board review. Besides, a limited removal of batch effects has been conducted.[44]

The CRC data contain the following 3 parts: data_clinical_patient, data_clinical_sample, and data_RNA_Seq_v2_mRNA_median_all_sample_Zs-cores (RNA). All RNA-seq (mRNA) data were dichotomized based on the z-scores (greater than the

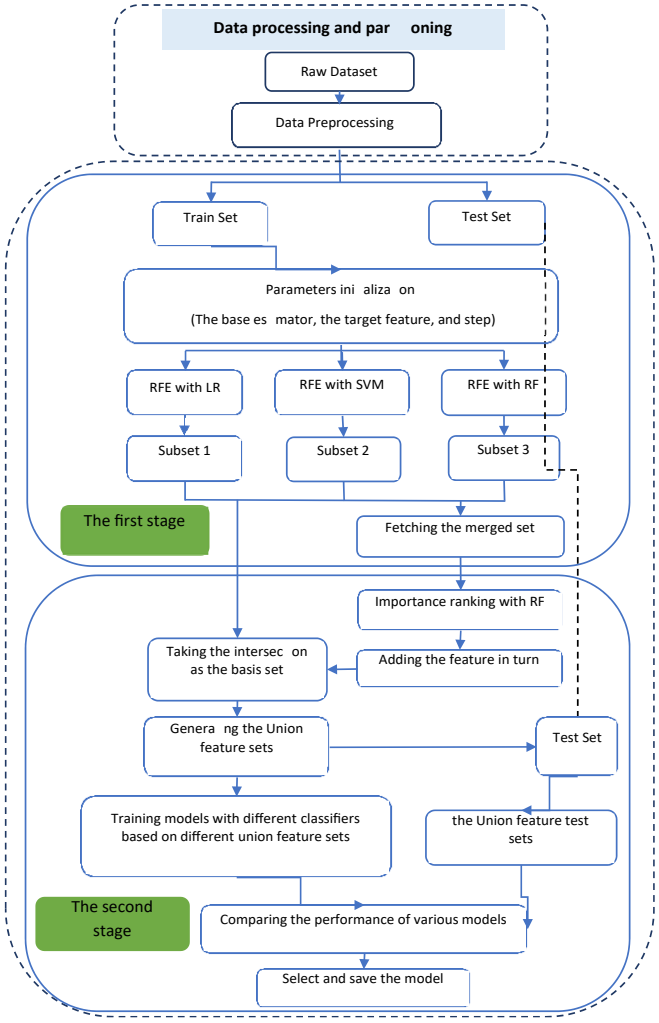

Figure 1.
Flow chart of the Union with Recursive Feature Elimination (U-RFE) framework.

median considered "1"; otherwise considered "0") so that no normality test would be required. These 3 data parts are linked by the PATIENT_ID, which was randomly generated and not linked to any protected health information.

In the data cleaning process, as shown in Table 1, the samples TCGA_F5_6810, TCGA_5M_AAT5, and TCGA_5M_AATA were first removed because they did not have corresponding tags. The samples TCGA_AF_2689 and TCGA_AA_3558 were removed because they did not exist in the RNA dataset.

For data_clinical_patient and data_clinical_sample, all features were dichotomized using a one-hot encoding approach as described before.[42] The features with missing values are filled with a median value of the feature when the missing data present less than 10% of the samples (n ¼ 14 (2.38%) for AJCC_PATHOLOGIC_TUMOR_STAGE, 7 (1.19%) for PATH_M_STAGE, 3 (0.51%) for ICD_O_3_SITE, respectively). The features with missing values are replaced with "NA" (not available) when the missing data constitute not less than 10% of the samples (eg, n ¼ 101, 17.15% for radiation therapy, and n ¼ 274, 46.52% for weight, respectively). The patients (n ¼ 2) who had no available data on sex or age are removed. The weight is classified into 3 categories, including weight_above_average (weight > median of the weight), weight_below_average (weight median of the weight), and weight_NA (missing data) before being dichotomized using a one-hot encoding approach. A small number of missing values also occur in the RNA dataset, but due to their relatively high data dimensionality (20,531 features), all features containing missing values were removed directly. The processed CRC data contained 589 samples and 17,719 features.

Table 1
Excluded samples and reasons

Excluded samples	Reason for exclusion
TCGA_F5_6810 TCGA_5M_AAT5 TCGA_5M_AATA	Category label does not exist
TCGA_AF_2689 TCGA_AA_3558	Not present in the RNA data

Two features from data_clinical_patient, OS_STATUS (OS, Overall Survival Status) and PFS_STATUS (PFS, Progression Free Status), were then used together to determine the COD classification labels. Among them, OS_STATUS has 2 values as follows: 0 means alive (LIVING) and 1 means dead (DECEASED), and PFS_STATUS also has 2 values as follows: 0 means no progress status (CENSORED) and 1 means disease progress status (PROGRESSION). Therefore, the category labels are formulated as shown in Table 2. The corresponding label names for each category are as follows: alive without progression (AWNP), alive with progression (AWP), death without progression (DWNP), and death with progression (DWP). The sample sizes for each category were 406 (AWNP, 68.9%), 64 (AWP, 10.9%), 35 (DWNP, 5.9%), and 84 (DWP, 14.3%), respectively.

Model Tuning Based on Union with Recursive Feature Elimination

Recursive Feature Elimination Feature Selection

Therefore, RFE is the main representative of wraparound feature selection, which introduces classification algorithms into the feature selection process to eliminate redundancy between features and output the optimal combination of features.[45] It obtains the importance of each feature in the current training set by means of a base estimator and then removes the low-importance features from it to obtain a new subset of features. The core idea of RFE is to repeat this recursive process on a new subset of features until the number of selected features is reached.

When RFE was used for feature selection, the base estimator, the number of features selected (n_feature_to_select), and the feature removal step must be determined. Based on the training set, an optimized base estimator was used to obtain the importance of each feature (feature parameters). The number of features removed during each recursion was controlled according to the feature removal step until a feature set was obtained that matched the number of features selected. Even if the number of features selected was consistent, the final selection of features may vary depending on the base estimator, the parameters of the base estimator, or the feature removal step.

The Union with Recursive Feature Elimination Feature Selection

To select more representative, important, and comprehensive features, the U-RFE feature selection framework was proposed in this study.

Normally, only classifiers with the feature_importances_ or coef_ attribute can be used as base estimators of U-RFE. By contrast, LR, SVM, and RF classifiers all have feature_importances_ or coef_ attributes, but use their own different strategies to

Table 2
Formulate classification labels

OS_STATUS	PFS_STATUS	Meaning	Classification label
0	0	Alive with No Progression, AWNP	1
0	1	Alive with Progression, AWP	2
1	0	Dead with No Progression, DWNP	3
1	1	Dead with Progression, DWP	4

compute feature importance and have their own advantages in the classification process. For example, the regularization parameter in the LR classifier prevents overfitting of the model and solves the problem of possible colinearity between high-dimensional features;[46,47] the SVM classifier is suitable for small sample sizes;[48,49] and the RF is not easily affected by categories imbalance.[50] Therefore, in this article, the LR, SVM, and RF classifiers were used as the base estimators to select their respective feature subsets, and then, feature subset fusion was used to combine the advantages of each base estimator.

Due to the high dimensionality of the CRC data (17,719 features) and the high redundancy among the features, the setting of the step size during feature removal is crucial for the final set of retained features. RF has the ability to rank the importance of features and is often used as a tool for feature removal. Therefore, before designing the U-RFE algorithm, we used RF to observe the correspondence between the classification effect of the model and the gradual decrease in the number of features. The whole analysis process is repeated 10 times, and the average value is taken to obtain the curves of accuracy and F1_weighting with the change of the number of features, as shown in Figure 2. From the curves, it can be seen that although the overall classification effect is not satisfactory, the correspondence between the number of features and the classification effect can still be seen: when the number of features is high, the classification performance does not change much and the classification effect is poor. As the number of features decreases, the classification performance gradually improves. However, when the number of features is reduced to nearly 50, the classification performance declines sharply again. It is noteworthy that after repeating the above process 10 times, there were still differences in the individual feature sets generated by the same base estimator. The differences are more pronounced for the low-ranking features whose importance appears to vary by the experiment runs.

According to this rule of change, in the subsequent RFE process using different base estimators, we first set the feature removal step size to 5,000, which can efficiently and quickly remove the features that have a small impact on the classification effect, but when the number of features is reduced to nearly 200, the removal step size is reset to 10. The subsequent use of different base estimators to select the base features according to the setting of the removal step size can make the feature removal process efficient and effective. Following this method of setting the removal step and using different base estimators several times to select the base features, the feature removal process can effectively retain the sensitive features and is relatively efficient. Then, when fusing the feature sets obtained by each base estimator, it is possible to maximize the strengths and avoid the weaknesses to improve the classification effect.

The 3 base estimators (LR, SVM, and RF) need to be parameters tuned in each round of RFE feature recursive elimination. After 10 repeats, 30 feature subsets with the same number of features were obtained for feature union.

The error of a ML model arises from the combination of biases and variances. In general, increasing the number of features reduces the bias, and decreasing the number of features reduces the

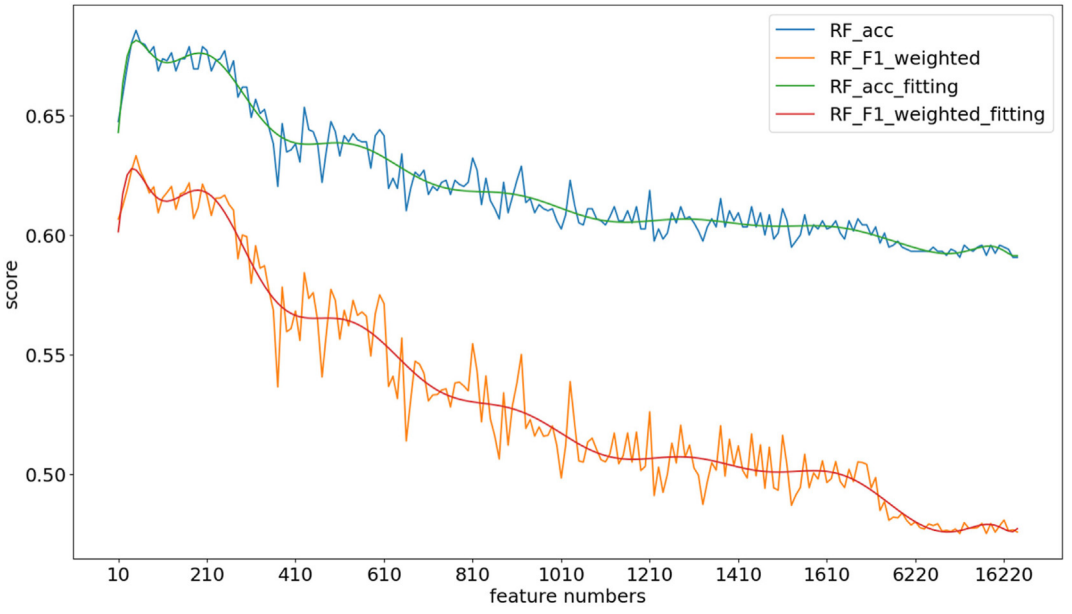

Figure 2.
Classification performance of different numbers of features using RF for feature selection.

variance. However, reducing bias increases variance and vice versa (bias-variance trade-off). Therefore, the number of features in the final feature subset should neither be as high as it could be nor as low as it could be.

For the high-dimensional TCGA CRC dataset, to take into account the strengths of each of the 3 base estimators, a two-step process of coarse and fine-tuning is performed to determine the optimal target feature number for the subset of features used for union.

In the coarse tuning stage, the target feature numbers for RFE feature selection are set to the following representative numbers: 50, 100, 200, 800, 1400, 2000, and 2600. For each target feature number, 3 base estimators (LR, SVM, and RF) will be used to obtain the corresponding feature subset. The 3 feature subsets with the same target feature number were merged to obtain the corresponding merged sets. Before deciding to use the RF classification algorithm, we also tried other classification algorithms such as LR and SVM. The classification effect of each model was compared using effective evaluation metrics. The RF algorithm had a better overall classification effect than other algorithms (data not shown) and thus was finally chosen. Based on the 7 merged sets with different target feature numbers, the RF classification algorithm was used to classify COD on each merged set, and the optimal target feature number for the subset was roughly determined based on the classification effect.

In the fine-tuning stage, the optimal target feature number in the subsets was analyzed in more detail: the target feature number in the subsets was gradually optimized in steps of 10 or 50 (the step size is set to 50 when the number of features is greater than 100 and to 10 when the number of features is less than 100). The intersection of 30 subsets with the same target feature number based on 10 repeats was taken as the basis for each round and then features outside the intersection and within the merged set were gradually added to the union set according to the feature priority obtained by the RF algorithm (features belonging to 2 subsets have priority over features belonging to one subset). Our preliminary data using a small set of tuning parameters show that XGBoost and Stacking both performed very well. Thus, we directly included them in the second stage. For the data corresponding to different feature union sets, LR, SVM, RF, XGBoost, and Stacking classifiers were selected for classification prediction, and suitable evaluation metrics were used for model evaluation. Finally, the set of feature union sets that reaches the highest Accuracy or Recall_weighed as the tie-breaker for the sets with the same yet highest accuracy was selected as the final feature selection set.

Optimization of Classification Models

In the U-RFE process, each of the classifiers involved requires the tuning of model parameters and hyperparameters. Although the hyperparameters to be optimized vary between models (as shown in Table 3), all can be optimized using a grid search for the main hyperparameters.

However, CRC data are a highly imbalanced dataset. If accuracy was simply used as a performance metric, the trained classifier would be biased toward the majority category, resulting in lower recognition rates for the minority categories.[51-54] Therefore, in addition to a hierarchical 10-fold cross-validation strategy, performance metrics Precision_weighted, Recall_weighted, F1_weighted, and Matthews correlation coefficient (MCC) as shown in (1e4) were used to help determine the model parameters.[55-57] F1_weighted is a harmonized average of precision and recall, and the weights were set separately according to the percentage of each classification in the training set, which can better overcome the problem of model bias toward the majority categories. MCC also provides a more reliable performance measure than accuracy in the case of unbalanced data or different category sizes.

$$\text{Precision\_weighted} = \frac{\sum_{i=1}^{n} \text{Precision}\$\text{Num}_i}{\sum_{i=1}^{n} \text{Num}_i} \quad (1)$$

$$\text{Recall\_weighted} = \frac{\sum_{i=1}^{n} \text{Recall}\$\text{Num}_i}{\sum_{i=1}^{n} \text{Num}_i} \quad (2)$$

$$\text{F1\_weighted} = \frac{\sum_{i=1}^{n} \text{F1}\$\text{Num}_i}{\sum_{i=1}^{n} \text{Num}_i} \quad (3)$$

$$\text{MCC} = \frac{\sum_{i}^{n} \text{TP}_i \sum_{i}^{n} \text{TN}_i - \sum_{i}^{n} \text{FP}_i \sum_{i}^{n} \text{FN}_i}{\sqrt{\left(\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FP}_i\right)\left(\sum_{i=1}^{n} \text{TP}_i + \sum_{i=1}^{n} \text{FN}_i\right)\left(\sum_{i=1}^{n} \text{TN}_i + \sum_{i=1}^{n} \text{FP}_i\right)\left(\sum_{i=1}^{n} \text{TN}_i + \sum_{i=1}^{n} \text{FN}_i\right)}} \quad (4)$$

Among the various classifiers employed for selecting the model and determining its parameters, the stacked classifier is unique in that it is an ensemble classifier. Often, several individual classifiers were combined to achieve significantly better performance metrics than a single classifier. When using a Stacking classifier for the CRC dataset, 2 factors need to be considered: the classification performance of the individual classifiers should not be too poor and there should be a wide variety among the individual classifiers.[58,59] Therefore, to better combine the advantages of different classifiers, the classifiers in the first layer of Stacking were set to LR, SVM, RF, and XGBoost. However, the features used in the second layer of the Stacking model are a combination of the predicted labels of the different classifiers obtained in the first layer, and the complexity of the features is greatly reduced. Therefore, the second layer classifier is usually a simple classifier for classification prediction.[60] Considering the overall complexity and efficiency of the model, decision tree (DT) is used here as the classifier in the second layer of Stacking. Our preliminary study on a few different stacking combinations shows that DT seems like the best choice (data not shown).

Table 3
Hyperparameter optimization range for different classification models

Classifiers	Optimized parameters	Hyperparameter optimization range	Increment
RF	n_estimators	[20, 210]	10
	min_samples_split	[2, 20]	2
	min_samples_leaf	[2, 20]	2
	max_depth	[2, 20]	2
SVM	kernel	[rbf, linear, sigmoid]	e
	C	[0.1, 3]	0.1
LR	solver	[liblinear, newton-cg, lbfgs, sag, saga]	e
	C	[0.1, 3]	0.1
XGBoost	n_estimators	[20, 110]	10
	learning_rate	[0.1, 1]	0.1
	max_depth	[2, 20]	2
DT	min_samples_split	[2, 20]	1
	min_samples_leaf	[2, 20]	1
	max_depth	[2, 20]	1

Automation of the Union With Recursive Feature Elimination Process

We developed a Python package to automate and simplify the U-RFE process for wide and convenient use of the methodology. The package provides the end user with 2 options, including the default range and increment of targeted feature numbers (10 to 200 features, with 10 as the increment) or those defined by the end user. Based on the range and increment of targeted feature numbers, the package will automate the process and present the end use with tabulated performance metrics of LR, SVM, RF, Stacking, and XGBoost, using various union feature sets. After the selected feature set is determined by reviewing the tabulated algorithm performances, the end user can simply customize the package with the final number of the features without a specified increment of 0 (ie, a preset number of features) and save the tuned algorithm and selected features for future application.

Results

Coarse Tuning for Feature Selection

Based on the RFE strategy, the base estimators were LR, SVM, and RF, respectively, and the selected target feature numbers were set to 50, 100, 200, 800, 1400, 2000, and 2600 to obtain the corresponding feature subsets. The subsets of features with the same target feature number were combined, and then, the RF classifier was trained based on the training merge set according to the features contained therein. The RF classifier was used to classify and predict the corresponding test merge set samples. The Accuracy, Precision_weighted, Recall_weighted, F1_weighted, and MCC metrics were compared, and the results are shown in Figure 3.

The results in Figure 3 show that the individual feature subsets obtained when the selected target feature numbers were set to 50, 100, 200, 800, 1400, 2000, and 2600, respectively, differed, but the RF classification models corresponding to their concatenated sets obtained almost equal Accuracy and Recall_weighted on the test merge sets composed of the same samples, whereas the Precision_weighted, F1_weighted, and MCC differed significantly, but several curves still show a trend of decreasing with increasing number of feature values. When the selected target feature number is greater than 200, the classification performance indicator drops sharply.

The Accuracy, Precision_weighted, Recall_weighted, F1_weighted, and MCC metrics obtained without feature selection were 0.689, 0.482, 0.689, 0.567, and 0.011 for 4-category COD, respectively, and were almost useless for the identification of a few minority categories. However, the results in Figure 3 show that the overall effect can be effectively improved after feature selection, and the performance metrics corresponding to an increase in the target feature number corresponding to the base estimator show a decreasing trend; therefore, the target feature number was set within 200, and the classification effect was relatively better.

Fine-Tuning for Feature Selection

The first stage of the analysis determined that a target feature number of less than 200 was preferable for the selection of the RFE base estimator. Further feature selection was performed on this basis.

Consider the following 2 points: 1, the initial selection stage exhibits a decreasing trend in the overall performance metrics as the number of features increases, and 2, after obtaining 3 feature subsets with the number of features equal to the target feature

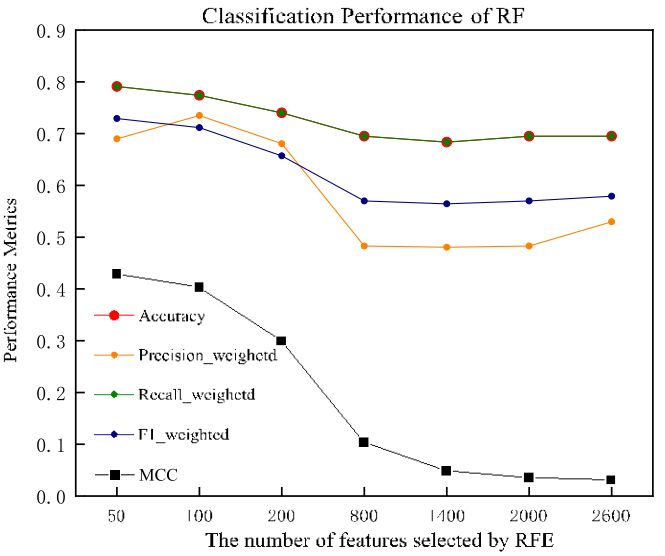

Figure 3.
Model performance metrics for the selected merge features in the first stage.

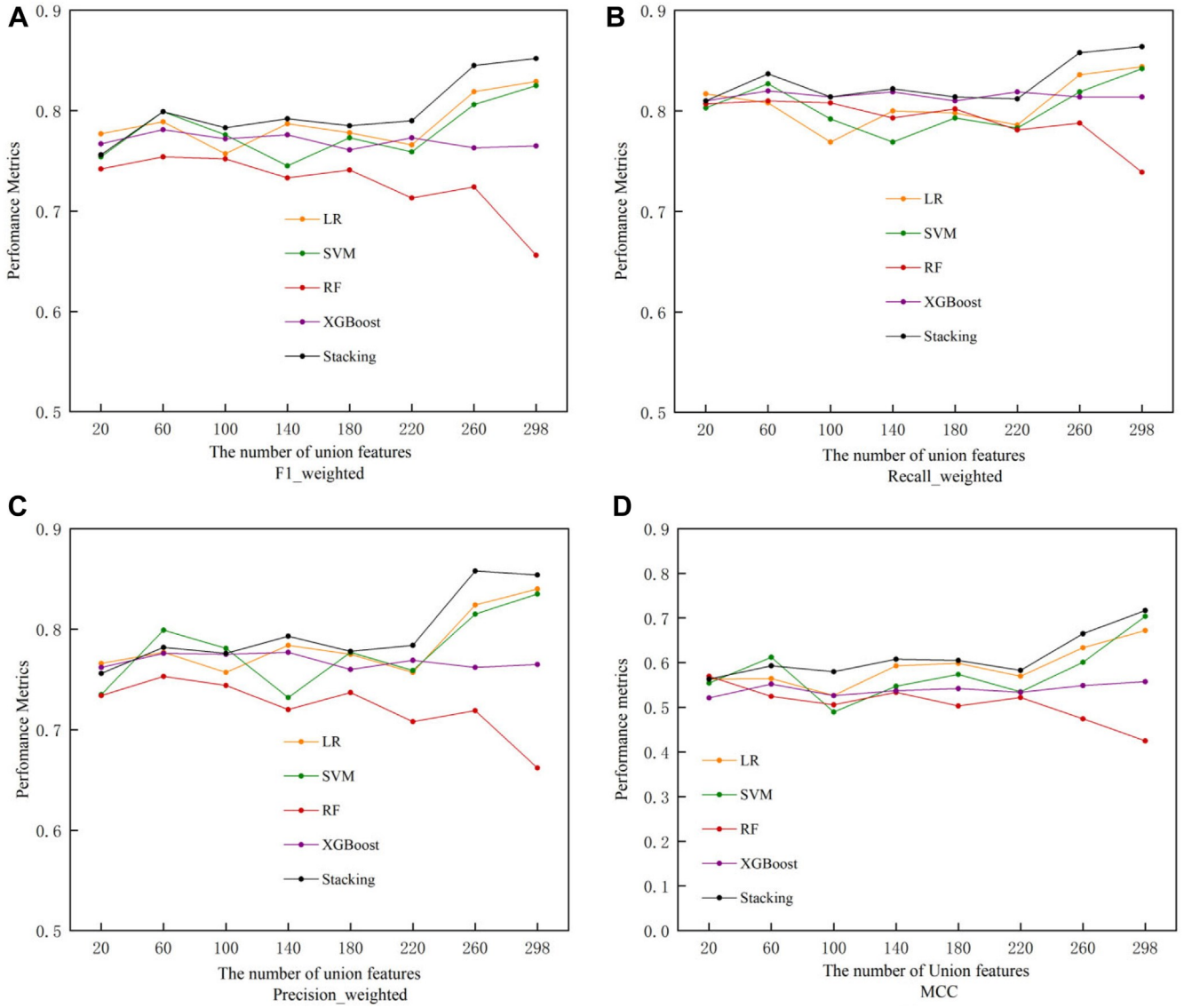

Figure 4.
The classification performance metrics of different classifiers for union features when the target feature number was 50.

number, the further union of the 3 feature subsets needs to be completed. The final number of optimal features obtained will be greater than the target feature number. Therefore, the target feature number selected by the RFE-based estimator in the selection stage was set to 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, and 200, respectively.

For each target feature number, each of the 3 RFE-based estimators selects a different subset of features. This process is repeated 10 times to obtain 30 subsets of features. The intersection of these 30 subsets was first taken to obtain the most important common features as the base set for feature union. For example, in the case of a target feature number of 50, all 30 feature subsets had 50 features, but after taking the intersection, there are only 20 features in common, whereas there are 278 features that were outside the intersection but within the union.

To determine whether these 278 features should be included in the final union feature set, we created 2-tier subset groups. The higher-ranked group includes common features belonging to different subsets obtained with different types of estimators, whereas the lower-ranked group includes common features belonging to subsets obtained with the same type of estimator. We then rank the features within each group according to the importance given by the RF algorithm. We finally add these ranked features sequentially to the base set by their group tier (the higher-ranked group first) and then within-group ranking order. Using these union sets, we choose LR, SVM, RF, XGBoost, and Stacking classifiers to perform classification prediction and evaluate their performances. Figure 4 shows that the Accuracy, Precision_weighted, Recall_weighted, F1_weighted, and MCC metrics exhibit some smoothness when the target feature number is smaller than 200, after which the performance varies between classifiers, with some rising while others falling.

The same analysis was carried out when the target feature number was set to the other aforementioned values. The trend of the performance metrics was shown in Figure 5 for a target feature number of 70 and in Figure 6 for a target feature number of 150. The trends in the performance metrics remain similar.

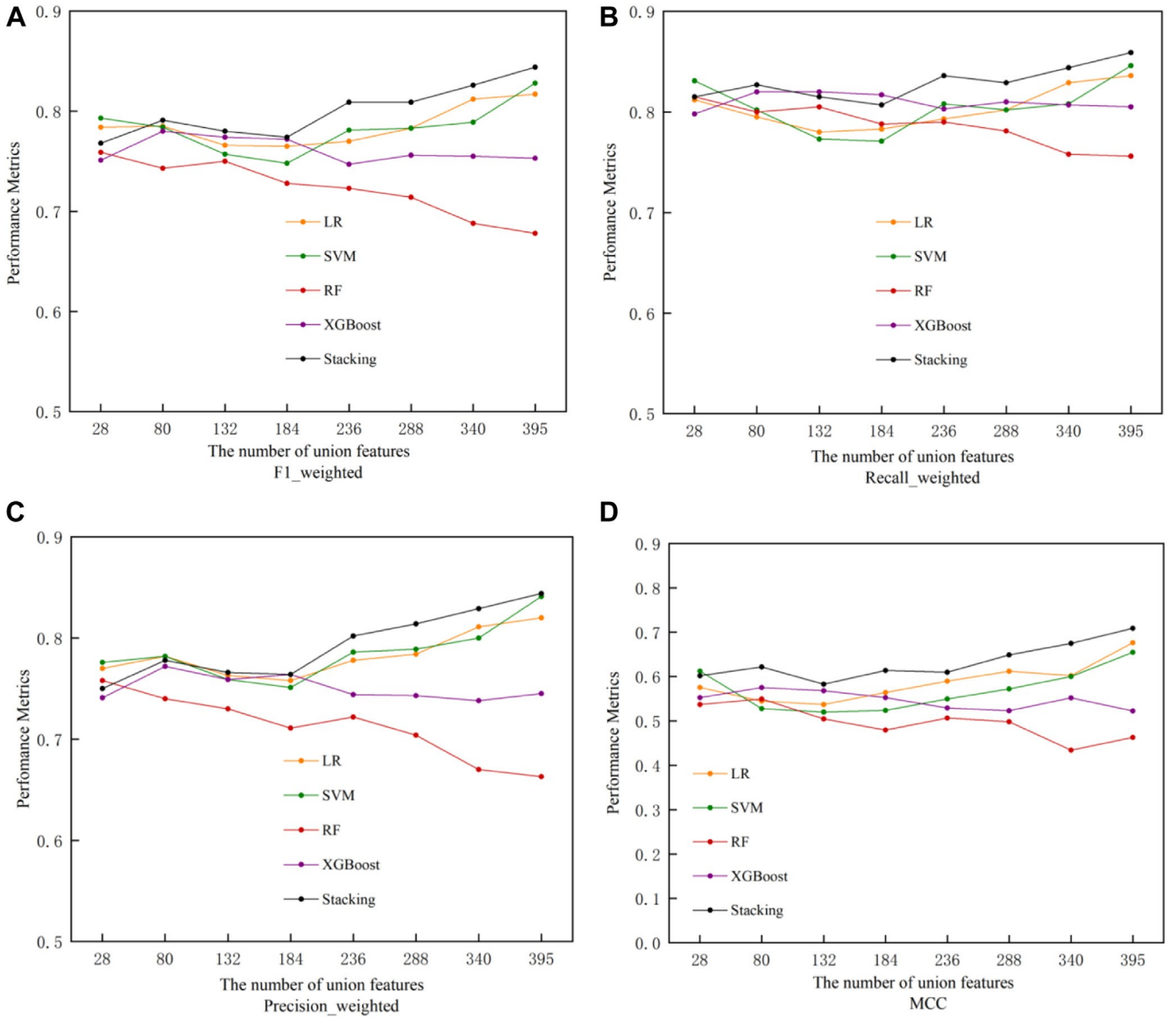

Figure 5.
The classification performance metrics of different classifiers for union features when the target feature number was 70.

Overall, the Stacking classifier outperformed the other classifiers, whereas the RF classifier performed relatively poorly.

Determination of the Final Union Feature Set

Through the above analyses, we obtained the corresponding performance metrics after applying different classifiers for classification prediction under a series of union feature sets corresponding to different target feature numbers. It is noteworthy that when the target feature numbers are different, the specific features contained in the union feature sets may be different, despite the fact that they contain the same number of features. Most of the classification models corresponding under these different union sets outperform the classification models before feature selection in the 4 performance metrics. However, it is unclear how to determine the final set of union features.

We then first identify all models with accuracy above 0.8 (the accuracy of models without feature selection was only about 0.6 to 0.7). Then, a strategy of F1_weighted first, Recall_weighted second, Precision_weighted third, and MCC last was adopted to rank all models. The parameters of the models with optimal performance metrics corresponding to each target feature number are listed in Table 4 and show that the Stacking classification model performed optimally in most cases. This finding is also consistent with the results of the previous analysis. Comparing the information in Table 4 longitudinally, the performance metrics of the Stacking classification model were optimal when the target feature number was 50, and the corresponding union feature set contains 298 features.

The error of a ML model arises from the combination of biases and variances. In general, increasing the number of features reduces the bias, and decreasing the number of features reduces the variance. However, reducing bias increases variance and vice versa (bias-variance trade-off).[61] Therefore, the number of features in the final feature subset should neither be as high as it could be nor

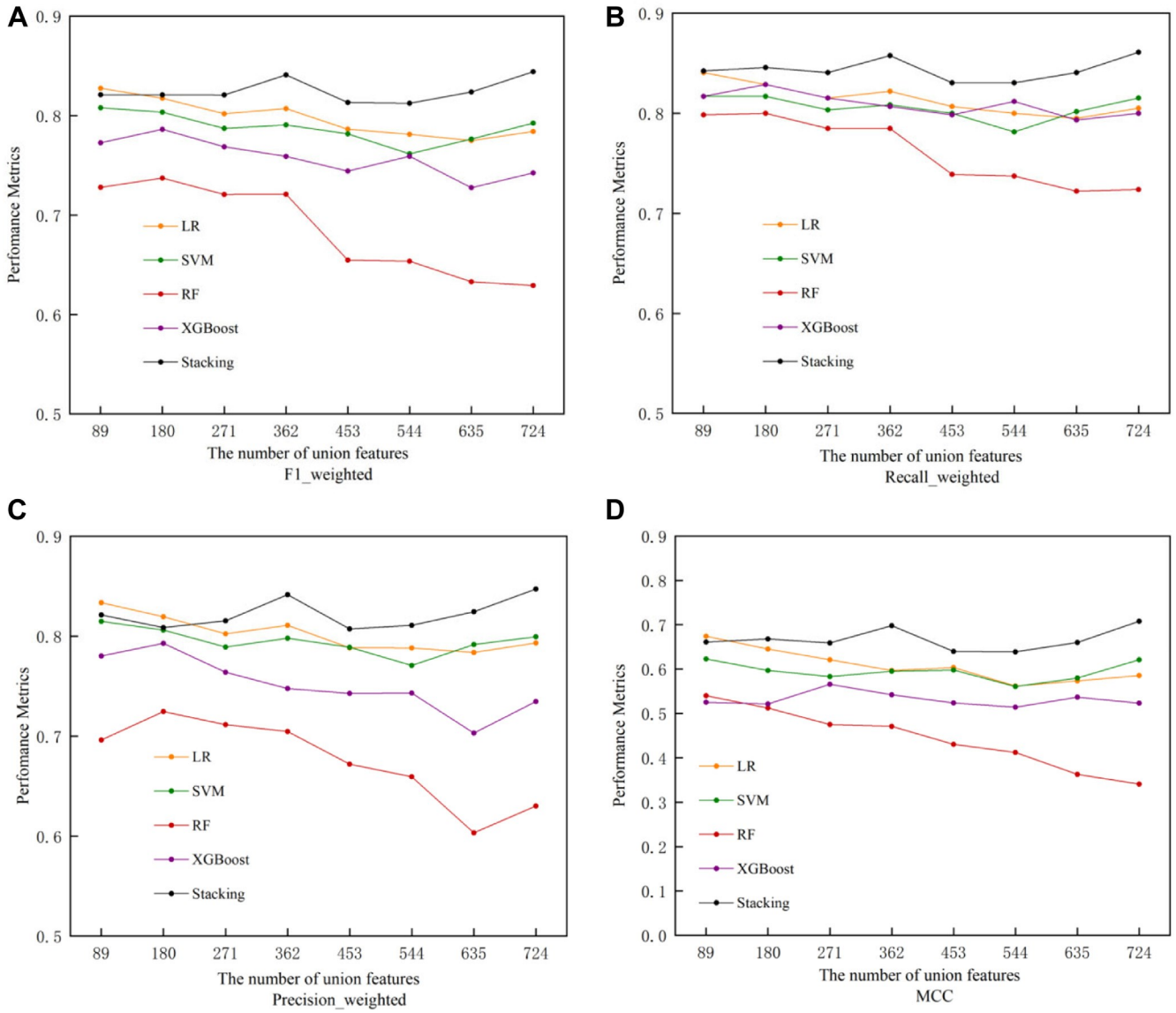

Figure 6.
The classification performance metrics of different classifiers for union features when target feature number was 150.

Table 4
The optimal classification performance with different basic features numbers

TFN	NFUF	NFA	Optimal classifier	F1_weighted	Recall_weighed	Precision_weighed	Accuracy
10	41	35	XGBoost	0.79	0.827	0.79	0.827
20	125	118	Stacking	0.834	0.851	0.837	0.851
30	164	158	Stacking	0.82	0.837	0.814	0.837
40	246	232	Stacking	0.841	0.849	0.851	0.849
50	298	278	Stacking	0.851	0.864	0.854	0.864
60	347	325	Stacking	0.843	0.858	0.852	0.858
70	395	367	Stacking	0.844	0.859	0.844	0.859
80	435	399	Stacking	0.838	0.856	0.839	0.856
90	482	439	Stacking	0.843	0.859	0.84	0.859
100	526	475	Stacking	0.843	0.858	0.854	0.858
150	724	635	Stacking	0.844	0.861	0.847	0.861
200	923	795	Stacking	0.831	0.847	0.837	0.847

NFA, number of features added to the base set; NFUF, number of features contained in the union feature set; TFN, the target feature number.

Table 5
The classification error corresponding to the optimal classification model

TFN	NFUF	Bias	Variance	MSE
10	41	0.849	0.290	1.139
20	125	0.709	0.145	0.853
30	164	0.801	0.140	0.941
40	246	0.764	0.100	0.864
50	298	0.592	0.076	0.668
60	347	0.623	0.099	0.722
70	395	0.629	0.082	0.711
80	435	0.679	0.111	0.790
90	482	0.697	0.103	0.800
100	526	0.687	0.124	0.811
150	724	0.693	0.076	0.769
200	923	0.758	0.092	0.850

MSE, mean squared error; NFUF, number of features contained in the union feature set; TFN, the target feature number.

as low as it could be. Due to the relative proximity of the multiple models listed in Table 4 in terms of performance metrics, we further evaluated the performance of the above classification models in terms of bias and variance (Table 5). When the target feature number had a value of 50 and the corresponding union feature set contained 298 features, the Stacking classification model still performed best (Bias, Variance, and Mean squared error were all minimized).

Discussion

Model Selection

Based on the above analyses, when the target feature number selected by the RFE base estimator was set to 50, a union feature set with 298 features was obtained. Based on this union feature set, the classification prediction using ML models can achieve excellent classification performance metrics and small classification errors, with the Stacking model performing even better. Given the imbalanced nature of CRC data, we were very concerned about the performance of the classification models on the minority categories with small sample sizes.[62,63] Therefore, we selected a union feature set containing 298 and 395 features obtained with target feature numbers of 50 and 70, respectively. The classification performance of the minority categories was observed using the Stacking classifier and compared with the performance of the classification model before feature selection. The results are shown in Table 6. Due to the high dimensionality of the original CRC data (17,719 features), the classification model would have been time consuming if the Stacking algorithm had been used directly for classification. Therefore, we compared the performance of the LR, SVM, RF, and XGBoost algorithms and present here the classification results of the best-performing RF classifier.

As shown in Table 6, these models hardly ever correctly predicted a sample in the minority categories before feature selection was performed. The 2 models that underwent U-RFE feature selection performed well in the majority category and both performed significantly in the minority categories, although their performance in the DWNP category was still unsatisfactory. Although the union feature set with 298 features had the best overall classification performance and the lowest corresponding classification error, it performed slightly worse on some minority categories (DWP) than the union feature set with 395 features. It is believed that similar results would be observed if more classification models with different union sets were compared. Therefore, we believe that in practical applications, the selection of feature sets and models should not only depend on certain performance metrics but also be more focused on the objectives of the application.[64,65] For example, if we tend to accurately predict samples belonging to the DWP category, then a model that performs well in this minority category should be selected, even if its corresponding overall performance is not the best. It, therefore, makes sense to perform careful feature selection and model comparison.

Ranking of Features

Further comparison of the specific features included in the different union feature sets. We found that as the target feature numbers of the RFE base estimator were set to increasingly larger values, the number of features included in the final resulting union feature set increased accordingly. As shown in Figure 7, features in the previous level were fully included in the next union feature set, which is consistent with the mechanism of recursive feature selection in RFE.

However, the position of a particular feature in the different union feature sets may vary. This is because changes in the value of the target feature number determine changes in the subset of features selected. Some features that were originally in a union of 3 subsets may in another case belong to only 2 of these subsets or may even change to belong to only one subset. As a result, their ranking positions in the final union set eventually change. The change in feature importance ranking also further affects the classification performance of the final model. Appendix 1 lists the ranking of the 298 selected features when the target feature number value was 50.

Conclusion

To address the poor performance of ML on the TCGA CRC clinical and omic data that have a large number of features, a small number of samples, and a strong sample imbalance, we propose an interesting U-RFE feature selection method to generate a union feature set. Based on the union feature set, the final decisive union feature set and classification model were determined by comparing the classification performance and classification error of several classifiers including LR, SVM, RF, XGBoost, and Stacking. These results show that the U-RFE feature selection method greatly reduces feature redundancy and effectively improves the overall performance of the classification model for multicategory CRC outcomes.

The comprehensive performance of the model is not the only criterion for model selection.[66] This article presents a method for selecting the appropriate set of features and classification models when the data are imbalanced, which allows us to take into account the classification of the minority categories while focusing on comprehensive performance.[67,68] Our recursive feature eliminationebased approach of feature selection improves the performances of classifying CRC deaths using clinical and omic data or those using other data with high feature redundancy and imbalance.

Some limitations of this study are noteworthy. Although the U-RFE feature selection method proposed here is time consuming in the feature selection process, it can provide useful proof in principle for refined model selection, and the generated models will be more efficient in future sample classification prediction. Moreover, a major limitation of this study was the lack of validation of the algorithm on another dataset. This limitation is largely attributable to the lack of a high-quality CRC dataset with a 4-

Table 6
Optimal classification performance of different classifiers on each category

TFN	NFUF	Classifiers	Categories	Accuracy	F1	Recall	Precision	Runtime(s)[a]
			AWNP	0.689	0.816	0.992	0.693	
			AWP		0	0	0	
e	17,719[b]	RF	DWNP		0	0	0	0.014
			DWP		0	0	0	
			Weighted		0.567	0.689	0.482	
			AWNP	0.739	0.852	1.0	0.742	
			AWP		0.365	0.267	0.682	
		RF	DWNP		0.0	0.0	0.0	0.013
			DWP		0.235	0.175	0.541	
			Weighted		0.656	0.739	0.662	
50	298		AWNP	0.864	0.936	0.961	0.912	
			AWP		0.779	0.833	0.754	
		Stacking	DWNP		0.392	0.300	0.600	0.053
			DWP		0.703	0.675	0.754	
			Weighted		0.851	0.864	0.854	
			AWNP	0.756	0.864	1.0	0.761	
			AWP		0.498	0.417	0.787	
		RF	DWNP		0.0	0.0	0.0	0.013
			DWP		0.204	0.138	0.4	
			Weighted		0.678	0.756	0.663	
70	395		AWNP	0.859	0.931	0.956	0.908	
			AWP		0.778	0.833	0.744	
		Stacking	DWNP		0.318	0.25	0.5	0.055
			DWP		0.708	0.688	0.761	
			Weighted		0.844	0.859	0.844	

[a] Runtime(s) of a single test sample.
[b] 17,719 indicates the classification result without feature selection.

category COD. Furthermore, there are some features with missing data in the TCGA CRC dataset. The median imputing method and the creation of the missing data category used here may not be the best method to input all missing features. Future works seem warranted to address this limitation. In addition, the two-stage feature selection process may be time consuming and overly complex. Automation of our proposed framework may help end users, whereas the underlying scientific merits of combining various algorithms may still overwhelming to some users. Finally, the sample size is relatively small, whereas it was reasonably large for omics datasets with detailed survival outcomes, and it is difficult to find a similar dataset. Future research may be focused on addressing these limitations.

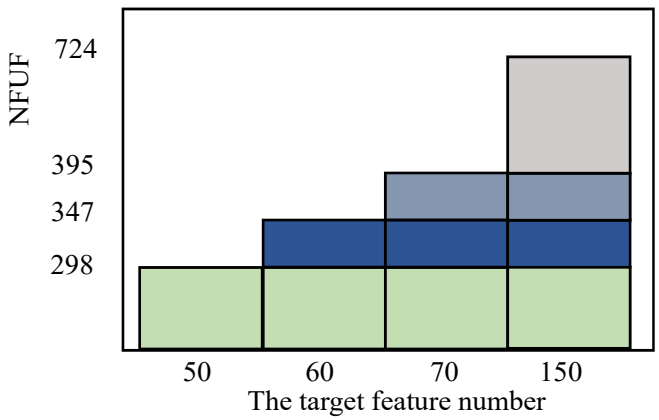

Figure 7.
The number of features included in the union features sets (NFUF) by the targeted feature number.

The model developed in this study can help us understand the molecular mechanisms of CRC progression and, to some extent, assist physicians to more effectively manage CRC patients. At the same time, the analysis method proposed in this article is also applicable to other types of datasets.

Acknowledgment

This work was in part supported by U.S. National Science Foundation (IIS-2128307 to L.Z.).

Author Contributions

F.D. and L.Z. designed this study, and all authors participated in the discussion and implementation of research. L.Z. drafted the manuscript, and all authors discussed and edited the manuscript. F.D. and L.Z. supervised the work and revised the manuscript.

Data Availability

The datasets used and/or analyzed in this study are available on the cBioPortal website (https://www.cbioportal.org/). The program coding is deposited to the GitHub (https://github.com/FeiDeng-RUTGERS/URFE.git) and available from the corresponding authors on reasonable request.

Funding

The funder has no roles in writing this work or the decision to submit it for publication.

Declaration of Competing Interest

None to disclose.

Ethics Approval and Consent to Participate

This exempt study using publicly available deidentified data did not require an IRB review.

References

1. Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020. CA Cancer J Clin. 2020;70(3):145e164.
2. Liu Z, Liu L, Weng S, et al. Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. Nat Commun. 2022;13:1e14.
3. Koncina E, Haan S, Rauh S, Letellier E. Prognostic and predictive molecular biomarkers for colorectal cancer: updates and challenges. Cancers. 2020;12(2):319.
4. American Cancer Society. Colorectal Cancer Facts & Figures 2020-2022. American Cancer Society; 2020.
5. Zhang M, Hu W, Hu K, et al. Association of KRAS mutation with tumor deposit status and overall survival of colorectal cancer. Cancer Causes Control. 2020;31:683e689.
6. Amin MB, Greene FL, Edge SB, et al. The eighth edition AJCC cancer staging manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin. 2017;67:93e99.
7. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol. 2010;17:1471e1474.
8. Mayo E, Llanos AM, Yi X, Duan SZ, Zhang L. Prognostic value of tumour deposit and perineural invasion status in colorectal cancer patients: a SEER-based population study. Histopathology. 2016;69:230e238.
9. Chavali LB, Llanos AAM, Yun JP, Hill SM, Tan XL, Zhang L. Radiotherapy for patients with resected tumor deposit-positive colorectal cancer: a surveillance, epidemiology, and end results-based population study. Arch Pathol Lab Med. 2018;142:721e729.
10. Li MX, Liu XM, Zhang XF, et al. Prognostic role of neutrophil-to-lymphocyte ratio in colorectal cancer: a systematic review and meta-analysis. J Natl Cancer Inst. 2014;134:2403e2413.
11. Xu Y, Ju L, Tong J, Zhou C, Yang J. Machine learning algorithms for predicting the recurrence of stage IV colorectal cancer after tumor resection. Sci Rep. 2020;10:1e9.
12. Reddy GT, Reddy MPK, Lakshmanna K, et al. Analysis of dimensionality reduction techniques on big data. IEEE Access. 2020;8:54776e54788.
13. Maros ME, Capper D, Jones DTW, et al. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. Nat Protoc. 2020;15:479e512.
14. Mohammed M, Mwambi H, Mboya IB, Elbashir MK, Omolo B. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Sci Rep. 2021;11:1e22.
15. Spyropoulos CD. AI planning and scheduling in the medical hospital environment. Artif Intell Med. 2000;20:101e111.
16. Deng F, Shen L, Wang H, Zhang L. Classify multicategory outcome in patients with lung adenocarcinoma using clinical, transcriptomic and clinico-transcriptomic data: machine learning versus multinomial models. Am J Cancer Res. 2020;10:4624e4639.
17. Deng F, Huang J, Yuan X, Cheng C, Zhang L. Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data. Lab Invest. 2021;101:430e441.
18. Deng F, Zhou H, Lin Y, et al. Predict multicategory causes of death in lung cancer patients using clinicopathologic factors. Comput Biol Med. 2021;129:104161.
19. Wang J, Deng F, Zeng F, Shanahan AJ, Li WV, Zhang L. Predicting long-term multicategory cause of death in patients with prostate cancer: random forest versus multinomial model. Am J Cancer Res. 2020;10:1344e1355.
20. Ogunleye A, Wang QG. XGBoost Model for chronic kidney disease diagnosis. IEEE/ACM Trans Comput Biol Bioinform. 2019;17(6):2131e2140.
21. Liu Q, Zhao Q, McMinn A, Yang EJ, Jiang Y. Planktonic microbial eukaryotes in polar surface waters: recent advances in high-throughput sequencing. Mar Life Sci Technol. 2020;3(1):94e102.
22. Yang M, Yang H, Ji L, et al. A multi-omics machine learning framework in predicting the survival of colorectal cancer patients. Comput Biol Med. 2022;146:105516.
23. O'Brien R, Ishwaran H. A random forests quantile classifier for class imbalanced dat. Pattern Recognit. 2019;90:232e249.
24. Yang Y, Jiang J. Adaptive bi-weighting toward automatic initialization and model selection for HMM-based hybrid meta-clustering ensembles. IEEE Trans Cybern. 2018;49:1657e1668.
25. Pouyanfar S, Chen SC. Automatic video event detection for imbalance data using enhanced ensemble deep learning. J. Semantic Comput. 2017;11:85e109.
26. Yang Y, Nan F, Yang P, et al. GAN-based semi-supervised learning approach for clinical decision support in health-IoT platform. IEEE Access. 2019;7:8048e8057.
27. Saarela M, Ryynänen OP, Äyrämö S. Predicting hospital associated disability from imbalanced data using supervised learning. Artif Intell Med. 2019;95:88e95.
28. Yin Q, Chen W, Zhang C, Wei Z. A convolutional neural network model for survival prediction based on prognosis-related cascaded Wx feature selection. Lab Invest. 2022;102:1064e1074.
29. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. 2019;112:103375.
30. Shahrjooihaghighi A, Frigui H, Zhang X, Wei X, Shi B, Trabelsi A. An ensemble feature selection method for biomarker. Disc Proc IEEE Int Symp Signal Proc Inf Tech. 2017:416e421.
31. Plyushchenko I, Shakhmatov D, Bolotnik T, Baygildiev T, Nesterenko PN, Rodin I. An approach for feature selection with data modelling in LC-MS metabolomics. Anal Methods. 2020;12:3582e3591.
32. Xanthopoulos P, Pardalos PM, Trafalis TB. Linear Discriminant Analysis. Robust Data Mining. Springer; 2013:27e33.
33. Zhao H, Wang Z, Nie F. A new formulation of linear discriminant analysis for robust dimensionality reduction. IEEE Trans Knowl Data Eng. 2018;31(4):629e640.
34. Sugiyama M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. J Mach Learn Res. 2007;8(37):1027e1061.
35. Wang S, Lu J, Gu X, Du H, Yang J. Semi-supervised linear discriminant analysis for dimension reduction and classification. Pattern Recognit. 2016;57:179e189.
36. Pandey H, Lee S, Tseng C, et al. A dimension-reduction based multilayer perception method for supporting the medical decision making. Pattern Recognit Lett. 2020;131:15e22.
37. Ng WWY, Hu J, Yeung DS, Yin S, Roli F. Diversified sensitivity-based undersampling for imbalance classification problems. IEEE Trans Cybern. 2014;45(11):2402e2412.
38. Babar V, Ade R. A novel approach for handling imbalanced data in medical diagnosis using undersampling technique. Commun Appl Electron. 2016;5:36e42.
39. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority oversampling technique. J Artif Intell Res. 2002;16:321e357.
40. Jiang K, Lu J, Xia K. A novel algorithm for imbalance data classification based on genetic algorithm improved SMOTE. Arab J Sci Eng. 2016;41(8):3255e3266.
41. Yun J, Ha J, Lee JS. Automatic determination of neighborhood size in SMOTE. Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication; 2016:1e8.
42. Feng CH, Disis ML, Cheng C, Zhang L. Multimetric feature selection for analyzing multicategory outcomes of colorectal cancer: random forest and multinomial logistic regression models. Lab Invest. 2022;102:236e244.
43. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6.pl1-pl1.
44. The Cancer Genome Atlas Research Network, Weinstein J, Collisson E, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113e1120.
45. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389e422.
46. Ananthakrishnan A N, Hoffmann R G, Saeian K. Higher physician density is associated with lower incidence of late-stage colorectal cancer. J Gen Intern Med. 25(11):1164-1171.
47. Sikdar KC, Dickinson J, Winget M. Factors associated with mode of colorectal cancer detection and time to diagnosis: a population level study. BMC Health Serv Res. 2017;17:1e11.
48. Liu Q, Gu Q, Wu Z. Feature selection method based on support vector machine and shape analysis for high-throughput medical data. Comput Biol Med. 2017;1(91):103e111.
49. Majid A, Ali S, Iqbal M, Kausar N. Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. Comput Methods Programs Biomed. 2014;113:792e808.
50. Zhang IY, Hart GR, Qin B, Deng J. Long-term survival and second malignant tumor prediction in pediatric, adolescent, and young adult cancer survivors using Random Survival Forests: a SEER analysis. Sci Rep. 2023;13:1911.
51. Chawla NV, Data mining for imbalanced datasets: An overview. Data Mining and Knowledge Discovery Handbook. Springer; 2010:875e886.
52. Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl. 2004;6:20e29.

53. Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res. 2018;61:863e905.
54. Zhao L, Deng F, Zhang X, Ning Y. RFE based feature selection improves performance of classifying multiple-causes deaths in colorectal cancer7. IEEE 2022 7th International Conference on Intelligent Informatics and Biomedical Science (ICIIBMS); 2022:188e194.
55. Su Q, Liu Q, Lau RI, et al. Faecal microbiome-based machine learning for multi-class disease diagnosis. Nat Commun. 2022;13:6818.
56. Zhou D, Tian F, Tian X, et al. Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer. Nat Commun. 2020;11:2961.
57. Evaluation D Powers. from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011;2(1):37e63.
58. Zhou ZH. Machine Learning. Springer Nature; 2021.
59. Xu C, Zhang R, Duan M, Zhou Y, Bao J, LU H. A polygenic stacking classifier revealed the complicated platelet transcriptomic landscape of adult immune thrombocytopenia. Mol Ther Nucleic Acids. 2022;28:477e487.
60. Liu J, Dong X, Zhao H, Tian Y. Predictive classifier for cardiovascular disease based on stacking model fusion. Processes. 2022;10:749.
61. Ng A. Machine Learning Yearning. 2018. Accessed December 4, 2023. https://www.dbooks.org/machine-learning-yearning-1501/#google_vignette
62. Matsuo H, Nishio M, Kanda T, Kojita Y, Kono AK, Hori M. Diagnostic accuracy of deep-learning with anomaly detection for a small amount of imbalanced data: discriminating malignant parotid tumors in MRI. Sci Rep. 2020;10, 19388.
63. Wang S, Dai Y, Shen J, Xuan J. Research on expansion and classification of imbalanced data based on SMOTE algorithm. Sci Rep. 2021;11:1e11.
64. Caldiera VRBG, Rombach HD, The goal question metric approach. Encyclopedia of Software Engineering. JohnWiley & Sons; 1994:528e532.
65. He H, Ma Y. Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley-IEEE Press; 2013.
66. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21:1263e1284.
67. Ou G, Murphey YL. Multi-class pattern classification using neural networks. Pattern Recognit. 2007;40:4e18.
68. Zhang H, Huang L, Wu CQ, Li Z. An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset. Comput Netw. 2020;177, 107315.