scientific reports

Check for updates

OPEN

Predicting pathological response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer with two step feature selection and ensemble learning

Changshun Qian[1,2,5], Shuxin Yang[1,5], Yijing Chen[2,3], Ran Ge[1], Fangmin Shi[2,3], Chengnan Liu[2,3,4], Hui Wang[4]✉ & You Guo[1,2]✉

Patients with locally advanced rectal cancer (LARC) show substantial individual variability and a pronounced imbalance in response distribution to neoadjuvant chemoradiotherapy (nCRT), posing significant challenges to treatment response prediction. This study aims to identify effective predictive biomarkers and develop an ensemble learning-based prediction model to assess the response of LARC patients to nCRT. A two-step feature selection method was developed to identify predictive biomarkers by deriving stable reversal gene pairs through within-sample relative expression orderings (REOs) from LARC patients undergoing nCRT. Preliminary screening utilized four methods—MDFS, Boruta, MCFS, and VSOLassoBag—to form a candidate feature set. Secondary screening ranked these features by permutation importance, applying Incremental Feature Selection (IFS) with an Extreme Gradient Boosting (XGBoost) to determine final predictive gene pairs. The ensemble model BoostForest, combining boosting and bagging, served as the predictive framework, with SHAP employed for interpretability. Through two-step feature selection, the 32-gene pair signature (32-GPS) was established as the final predictive biomarker. In the test set, the model achieved an area under the precision-recall curve (AUPRC) of 0.983 and an accuracy of 0.988. In the validation cohort, the AUPRC was 0.785, with an accuracy of 0.898, indicating strong model performance. The study further demonstrated that BoostForest achieved superior overall performance compared to Random Forest, Support Vector Machine (SVM), and XGBoost. To evaluate the effectiveness of the 32-GPS, its performance was compared with two alternative feature sets: the lasso-gene pair signature (lasso-GPS), derived through lasso regression, and the 15-shared gene pair signature (15-SGPS), consisting of gene pairs identified by all four feature selection methods. The 32-GPS demonstrated superior performance in both comparisons. The two-step feature selection method identified robust predictive biomarkers, and BoostForest outperformed Random Forest, Support Vector Machine, and XGBoost in classification performance and predictive capability.

Keywords Feature selection, Locally advanced rectal cancer, Neoadjuvant chemoradiotherapy, Class imbalance problem, Ensemble learning

Neoadjuvant chemoradiotherapy (nCRT) prior to total mesorectal excision (TME) in locally advanced rectal cancer (LARC) patients reduces tumor volume, minimizes the risk of local recurrence, and improves the likelihood of successful resection, thereby enhancing prognosis[1–3]. Additionally, patients who achieve pathological complete response (pCR) after nCRT may avoid surgery, thereby enhancing their quality of life[4,5]. Despite the numerous benefits of nCRT, only about 15–27% of patients achieve pCR, indicating variability in

[1]School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China. [2]Medical Big Data and Bioinformatics Research Centre, First Affiliated Hospital of Gannan Medical University, Ganzhou 341000, China. [3]School of Public Health and Health Management, Gannan Medical University, Ganzhou 341000, China. [4]State Key Laboratory of Oncogenes and Related Genes, Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China. [5]Changshun Qian and Shuxin Yang contributed equally to this work. ✉email: huiwang@shsmu.edu.cn; gy@gmu.edu.cn

response due to the heterogeneity of tumor biology and the complexity of individual gene expression[6]. Such variability underscores the heterogeneity of tumor biology and the complexity of individual gene expression profiles. Current preclinical evaluation methods lack precision in predicting patient responses to nCRT, resulting in some patients enduring unnecessary side effects without deriving therapeutic benefits. In the context of precision medicine, identifying new biomarkers to predict response to nCRT is imperative. Such biomarkers would enable more precise treatment decisions, reducing both under-treatment and over-treatment and ultimately optimizing therapeutic outcomes for patients with LARC[7,8].

Gene expression analysis has demonstrated significant potential in predicting treatment responses[9–12]. Specific gene expression profiles offer valuable insights into patient treatment responses and prognosis[13,14], while also facilitating the identification of risk factors that can inform the development of novel therapeutic targets[15]. Although some progress has been made in predicting the response of rectal cancer patients to nCRT[16,17], there remain several limitations and challenges. Firstly, batch effects in gene expression data and the issue of high-dimensional feature selection continue to be critical factors influencing the performance of predictive models. Batch effects can lead to data discrepancies between experiments, thereby affecting the reliability of analysis results and the generalizability of the models. Additionally, feature selection in high-dimensional data involves identifying the most predictive features from a large number of variables, which is crucial for improving prediction accuracy and reducing model overfitting. Currently, three main types of feature selection methods are widely used, each based on different principles, leading to variations in the features selected.

Furthermore, the significant imbalance in the proportion of patients responding to nCRT presents a challenge for feature selection, complicating the prediction of pCR in LARC patients[18]. Among patients receiving nCRT, only a small fraction achieves pCR. This imbalanced sample distribution complicates the development of accurate predictive models. Although some studies have attempted to address the issue of imbalanced sample distribution through resampling techniques or synthetic sample generation[19,20], these methods may alter the original characteristics of the dataset or introduce biases. The generated samples may not fully represent the real clinical situations of patients, potentially leading to unstable model performance on new datasets.

Zheng et al. demonstrated that relative rank methods, such as ROGER, provide a robust alternative to traditional normalization for compositional NGS data analysis[21]. Among these, gene pair-based within-sample relative expression orderings (REOs) offer an innovative strategy to mitigate batch effects between samples. REOs are an analysis method based on qualitative transcript features. Since REOs rely solely on the relative order of gene expression, they exhibit strong robustness against batch effects, effectively eliminating systematic biases arising from different experimental conditions[22–25]. Additionally, ensemble learning methods, which combine multiple base learners, provide an effective solution to the issue of imbalanced sample distribution, particularly when the minority class consists of the positive samples that researchers focus on more heavily[26,27]. Traditional machine learning models are often sensitive to the majority class, which can result in poor prediction performance for minority class samples. However, ensemble learning enhances the ability to identify minority class samples through the combined decisions of multiple models, thereby improving overall model performance. Methods such as Bagging and Boosting can reduce the influence of majority class samples by adjusting sample weights or employing resampling strategies, ultimately improving the balance of classification results.

This study employs REOs to identify stable reversal gene pairs and adjusts the feature selection process with the goal of discovering new predictive biomarkers. By applying ensemble learning, the study aims to improve the prediction probability for pCR patients, potentially providing therapeutic guidance for LARC patients prior to surgery. Regarding feature selection, the study utilizes two rounds of feature screening: an initial screening followed by incremental feature selection. This approach is designed to reduce bias introduced by any single feature selection method, thereby increasing the likelihood of identifying relevant features and minimizing the risk of overfitting. Additionally, BoostForest is introduced as the predictive model[28], aimed at mitigating the impact of sample distribution imbalance on prediction accuracy, ultimately improving the model's ability to predict nCRT response in LARC patients.

Methods and materials

Data pre-processing

The workflow of this study is presented in Fig. 1. The cohort and validation cohort used in this study were both sourced from the Gene Expression Omnibus (GEO) database. The dataset GSE87211[29] was detected using the Agilent-026652 Whole Human Genome Microarray 4 × 44 K v2 (GPL13497) platform, which includes 363 samples. For this study, 106 LARC samples from the GSE87211 cohort were selected for analysis. GSE40492[30] served as the validation cohort, consisting of 118 samples used for experimental validation. All samples from both GSE87211 and GSE40492 cohorts were included in the prognostic analysis. All samples were collected from cancer patients. The data in this study have not undergone standardization across samples. It has been verified that all expression values have been log2-transformed. For genes represented by multiple probes, the arithmetic mean of the probe values is used, with values expressed on the log2 scale. The inclusion and exclusion criteria for sample selection were as follows: (1) patients who did not receive neoadjuvant chemoradiotherapy (nCRT) were excluded, (2) patients with a T stage of 4 were excluded[31], and (3) Disease-free survival (DFS) is a critical indicator for evaluating the efficacy of neoadjuvant chemoradiotherapy and for assessing patient prognosis. Missing DFS data can compromise the reliability of prognostic information. In this study, patients with incomplete DFS data were excluded from the survival analysis. The sample distributions for GSE87211 and GSE40492 are shown in Fig. 2a.

Differential gene identification

In the GSE87211 dataset, the study participants were classified into the pCR and non-pCR groups based on the tumor regression grade (TRG) from the American Joint Committee on Cancer (AJCC). Patients with TRG0

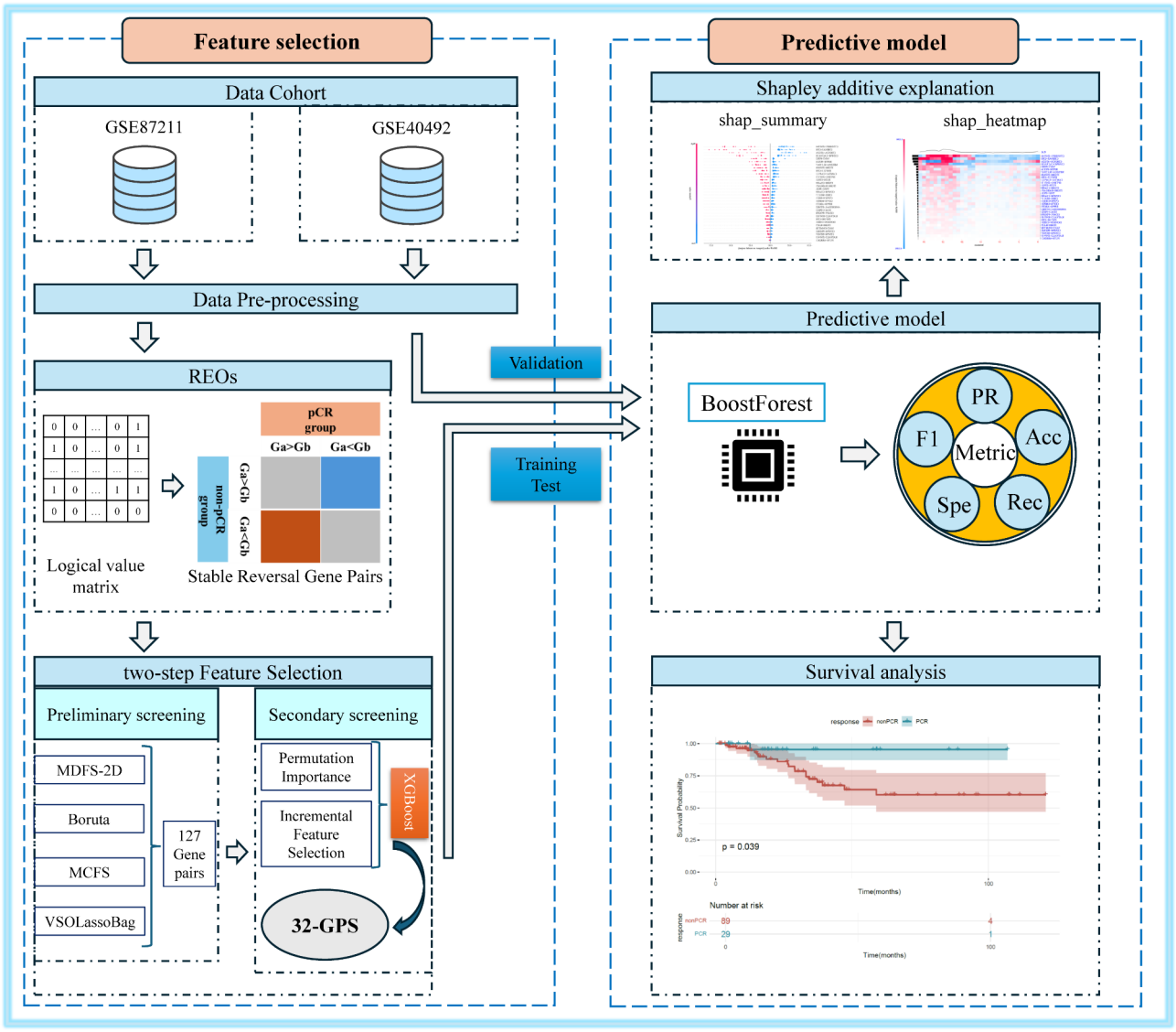


Fig. 1. Overview of the research workflow diagram.

were assigned to the pCR group, while those with TRG1, TRG2, and TRG3 were categorized as non-pCR. The GSE40492 dataset used Dowrak criteria to assess the pathological response status of patients. Although the two evaluation standards slightly differ, the differences between them are minimal[32]. In this study, the *limma* algorithm was applied for differential expression analysis to identify differentially expressed genes (DEGs) between the pCR and non-pCR groups, with a false discovery rate (FDR) of less than 5% and $|\log FC| > 0.5$. The detailed process of the *limma* algorithm is as follows: A design matrix was constructed based on the grouping information from the GSE87211 dataset to represent the different pCR and non-pCR groups. To perform the inter-group differential comparison, the *makeContrasts* function was used to create a contrast matrix, defining pCR - non-pCR as the comparison group. Next, we followed the standard workflow of the *limma* algorithm, initially fitting the data using the *lmFit* function, and then applying the contrast matrix to the fitted results with *contrasts.fit*. Finally, empirical *Bayes* adjustment was applied using *eBayes* to assess the significance of gene expression. It is important to note that the *eBayes* function does not enable trend adjustment by default.

We assessed the consistency of differentially expressed genes (DEGs) between the GSE87211 and GSE40492 datasets using the *limma* algorithm. Consistency was defined as the proportion of overlapping DEGs with the same direction of differential expression. Under completely random conditions, the probability that a gene exhibits consistent differential expression direction in two datasets is 0.5. The number of differentially expressed genes common to both datasets is denoted as n, and the number of genes with consistent differential expression direction is denoted as s. Therefore, consistency is defined as s divided by n. The statistical significance of the consistency of differentially expressed genes is then evaluated using the binom.test function.

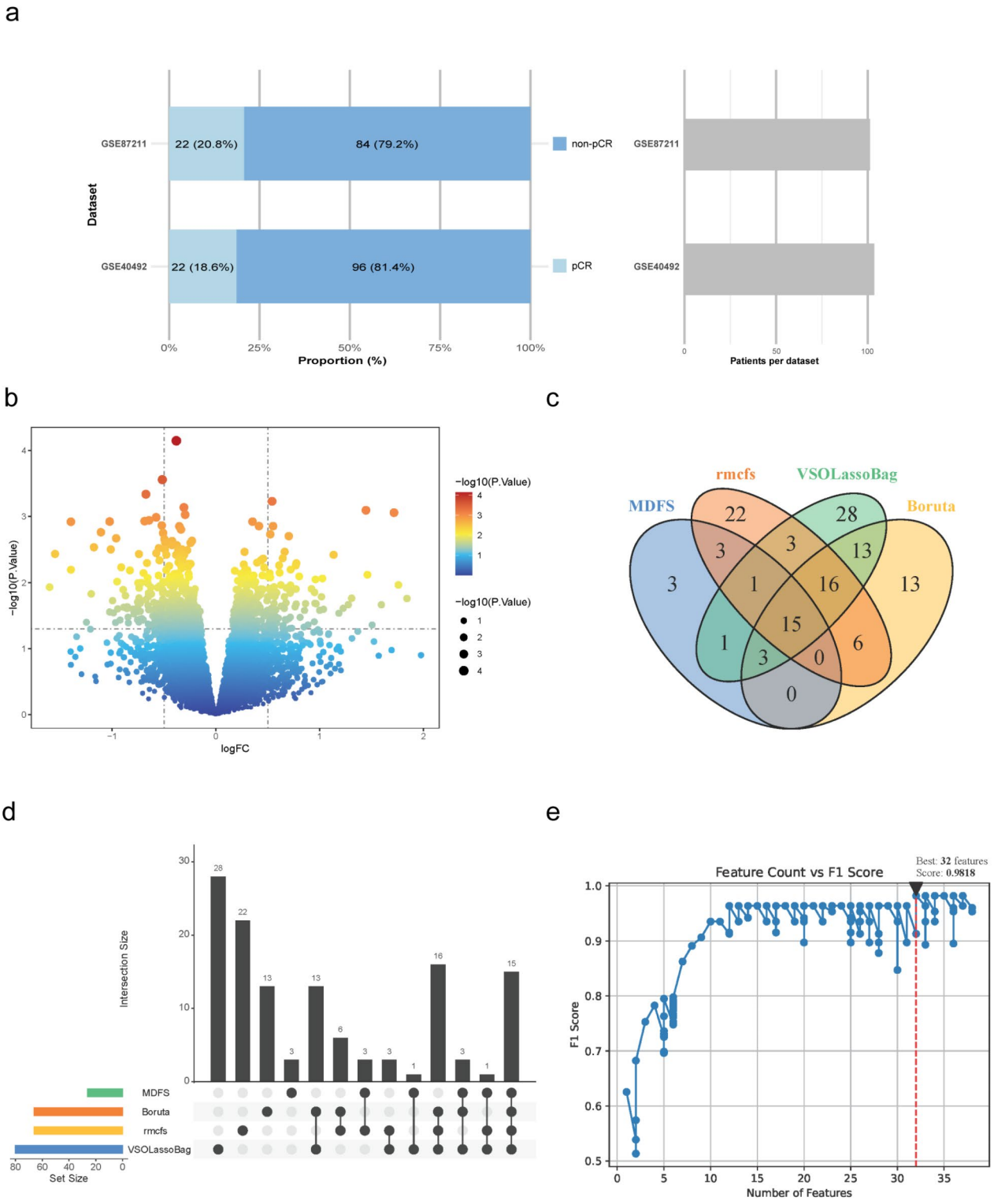


Fig. 2. **(a)** Distribution of pCR and non-pCR sample proportions in GSE87211 and GSE40492. **(b)** Volcano Plot of Differential Expression Analysis. **(c)** Venn diagram showing the overlapping genes identified by the four feature selection methods. **(d)** Upset plot showing the overlapping genes identified by the four feature selection methods. **(e)** Visualization of the feature selection process in the secondary feature selection.

Construction of gene pairs

In this study, we utilized REOs to construct gene pairs across two groups of samples and identify stable reversal gene pairs, with the excellent performance of REOs validated in the study by Xue et al.[7]. First, we constructed gene pairs separately within each sample group. For each gene pair, we test whether the expression of one gene is

significantly higher than the other, with the hypothesis that the proportion of times one gene's expression exceeds the other's is greater than 50%. This assumes that each comparison is an independent trial with two possible outcomes—either gene A's expression is higher than gene B's, or vice versa. After performing the binomial test, we obtained p-values for each gene pair and selected those with significant p-values. Stable reversal gene pairs were identified from this subset. Subsequently, a binomial distribution test was applied to calculate the *p*-value for each gene pair, and gene pairs with *p*-values less than 0.05 were selected, where at least one gene in each pair was identified as differentially expressed. We applied the Benjamini-Hochberg (BH) method for multiple comparisons correction of *p*-values (Supplementary Data 1). For any two genes *a* and *b* within the expression matrix, their expression values are represented by G_a and G_b, respectively, and can exhibit one of two relationships: $G_a > G_b$ or $G_b > G_a$. Stable reversal gene pairs are defined as gene pairs where the expression relationship between the genes exhibits opposite patterns in the pCR and non-pCR groups. In the pCR group, when the number of samples with consistent expression relationships exceeds 60%, and in the non-pCR group, when the number of samples with reverse expression relationships exceeds 60%, the gene pair is classified as a stable reversal gene pair. As shown in the Stable Reversal Gene Pairs section of Fig. 1, there are four possible expression relationships between gene pairs in the pCR and non-pCR groups: ① G*a* > G*b*, G*a* > G*b*; ② G*a* > G*b*, G*a* < G*b*; ③ G*a* < G*b*, G*a* > G*b*; ④ G*a* < G*b*, G*a* < G*b*. Gene pairs meeting the criteria specified in the red and blue regions (i.e., ② and ③) are identified as stable reversal gene pairs. The algorithm is provided in Supplementary Algorithm 1.

Selection of gene pairs

To select gene pairs capable of predicting the response of LARC patients to nCRT from the numerous stable reversal gene pairs, this study employed a two-step feature selection method. Preliminary screening was conducted using four distinct methods, namely MultiDimensional Feature Selection (MDFS)[33], Boruta[34], Monte Carlo Feature Selection (MCFS)[35], and VSOLassoBag[36], each offering unique advantages. To ensure comprehensive feature capture, the union of the selected features from all four methods was used as a new candidate gene set.

MDFS (MultiDimensional Feature Selection) is a type of filter method that accounts for the relationships between multiple features and decision variables. MDFS, by calculating multidimensional information gain, can identify informative variables that are related to the target variable and simultaneously capture interactions between variables. This method is particularly effective in identifying gene pairs with potential interactions[33]. Based on the recommendations of the method developers, this study employs the more comprehensive MDFS-2D approach.

Boruta is a feature selection algorithm based on random forests. Its fundamental principle involves evaluating the importance of each feature by comparing it with random shadow features[34]. In the context of nCRT response prediction, the high dimensionality of features necessitates the identification of key gene pairs from a large pool of redundant candidates. By iteratively calculating feature importance scores and performing global comparisons with shadow variables, Boruta effectively prioritizes gene pairs with superior predictive value.

MCFS evaluates both the importance of individual features and their interactions, thereby facilitating the identification of complex dependencies. MCFS has the capability to simultaneously evaluate the association between features and the target variable as well as the interactions among features[35]. In the prediction of nCRT response in LARC patients, potential interactions between features may not only have a significant impact on the performance of the predictive model but also substantially influence the outcomes of feature selection, thereby affecting the identification of key features.

VSOLassoBag combines the LASSO algorithm with a bagging strategy, demonstrating strong performance in the analysis of high-dimensional, low-sample-size biological data. Liang et al. have shown that VSOLassoBag is applicable in studies of colorectal cancer and can effectively mitigate issues related to multicollinearity, thereby avoiding the selection of highly correlated features[36]. LASSO inherently possesses sparsity, favoring the selection of a smaller number of features and typically selecting one feature from a group of highly correlated features during the feature selection process. By integrating multiple LASSO models, VSOLassoBag further optimizes the feature selection process.

For secondary screening, the study applied an Incremental Feature Selection (IFS) algorithm to identify the optimal predictive feature combination[37]. Permutation importance was used to rank the importance of gene pairs in the new candidate gene set[38], and Extreme Gradient Boosting (XGBoost) was selected as the classifier[39]. In the feature selection process, the IFS algorithm starts with the most important features ranked by importance and progressively adds them to the current feature subset. The performance of this feature subset is then evaluated using the XGBoost classifier, continuing until all features have been evaluated, at which point the algorithm terminates. Through cross-validation (GridSearchCV) or post-training performance evaluation, it is determined whether the addition of a new feature to the current subset contributes to improving the model's performance. Permutation Importance, a model-agnostic method, evaluates the importance of features by randomly shuffling the values of a given feature while holding others constant to assess the resulting decline in model performance. A significant performance drop upon shuffling indicates high importance for the feature, whereas minimal impact suggests low importance. In secondary screening, two evaluation metrics were employed: the performance of the XGBoost model after each training iteration and an evaluation metric for selecting optimal hyperparameters through grid search. Considering data imbalance, this study used a combination of log-loss and AUC as evaluation metrics for the XGBoost model and selected the F1 score as the grid search metric. To enhance model performance on minority classes, the ratio of pCR to non-pCR samples was used to adjust the weight in the loss function. In this study, the class imbalance ratio was calculated as the number of non-pCR samples (class 0) divided by the number of pCR samples (class 1). This ratio was directly assigned to the *scale_pos_weight* parameter in XGBoost to balance the contributions of both classes during training. Specifically,

adjusting *scale_pos_weight* helps prevent the model from being biased toward the majority class (non-pCR), thereby improving its ability to identify pCR cases.

Establishment of the predictive model

In this study, BoostForest was employed as the final predictive model. BoostForest is an ensemble learning method that combines the principles of boosting and bagging. First, multiple distinct subsets are generated from the original training set using the bootstrap method, and a BoostTree model is trained on each subset. The final prediction probability was obtained by calculating the arithmetic mean of the predicted probabilities from all BoostTree models. BoostTree utilizes gradient boosting to decompose the original problem into multiple sub-regression tasks and applies logistic regression to solve each sub-regression task at each node. A parameter pool sampling strategy is employed to facilitate the hyperparameter tuning process, where all possible values for each parameter are stored in a parameter pool[28]. BoostTree then randomly selects a parameter combination from this pool. Given an input, BoostTree assigns it to a leaf node and calculates the final prediction by summing the outputs of the models at all nodes along the path from the root to the leaf node. The accumulated output is then transformed into a probability using the Sigmoid function. BoostTree effectively combines the structured representation capability of decision trees with the strong learning ability of gradient boosting. When integrated with the BoostForest strategy, Zhao et al.'s study validated the predictive performance of BoostForest across multiple datasets by comparing it with models such as RandomForest, XGBoost, and LightGBM[28]. The results indicated that BoostForest outperformed the other models across multiple datasets and demonstrated superior generalization ability compared to the models in the comparison. In our study, we employed grid search for hyperparameter optimization. The hyperparameters adjusted for the BoostForest model include: *max_leafs*, with possible values of None, 1, 3, 5, or 7; *min_sample_leaf_list*, with possible values of None, 1, 3, 5, 7, or 9; *reg_alpha_list*, with values of 0.01, 0.05, 0.1, 0.2, or 0.3; and *n_estimators*, with possible values of 50, 100, 200, or 300. The algorithm for this section is provided in Supplementary Algorithm 2.

Pathway enrichment analysis

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were performed using KOBAS 3.0 (KEGG Orthology Based Annotation System)[40]. The results of the GO and KEGG enrichment analyses were visualized using R 4.4.0 software. A threshold of P.adj < 0.05 was applied to reduce the false discovery rate.

Interpretability analysis

To enhance the interpretability of the model, shapley additive explanation (SHAP)[41], a Python-based tool, was used for analysis on both the test and validation cohort. The primary function of SHAP analysis is to help understand the contribution of each feature to the model's prediction, thereby providing interpretability and aiding in the understanding of how the model makes decisions. The SHAP method quantifies the impact of each feature on the model's prediction. SHAP values demonstrate the direction and magnitude of each feature's effect on the prediction, assisting in identifying which features play a critical role in the model's decision-making process. This analysis helps reveal the decision-making logic of the model across different datasets, thereby improving its transparency and interpretability.

Statistical analysis

The statistical analysis in this study was performed using R 4.4.0 and Python 3.9.6. In R, the *limma* package was used to identify differentially expressed genes, while the *MDFS*, *Boruta*, *rmcfs*, and *VSOLassoBag* packages were utilized for Preliminary screening. The *survival* package was employed for survival analysis. In this study, gene pairs obtained using the REOs algorithm were evaluated using the binomial test ($p < 0.05$), and the p-values were subsequently corrected for multiple comparisons using the Benjamini-Hochberg (BH) method. GO and KEGG enrichment analyses were performed using the online tool KOBAS 3.0, with terms having a p-value < 0.05 considered statistically significant. The predictive ability of features was assessed using the precision-recall (PR) curve and the area under the precision-recall curve (AUPRC). Survival analysis of the dataset was conducted using the Kaplan-Meier method, with statistical comparisons made using the log-rank test. In Python, modeling and evaluation were conducted using the *scikit-learn* and *xgboost* packages, and the *shap* package was used for model interpretability analysis. Statistical significance was defined as a p-value less than 0.05. We used TOPSIS analysis to evaluate multiple metrics and determine the relative performance of each model[42].

Results

Differential analysis of pCR and non-pCR

In this study, differential expression analysis was conducted for 21 pCR samples and 85 non-pCR samples selected from the training cohort GSE87211. A total of 201 differentially expressed genes (DEGs) were identified. The identified DEGs were visualized using a volcano plot (Fig. 2b), with 86 genes upregulated and 115 genes downregulated.

The consistency analysis of DEGs between the GSE87211 and GSE40492 datasets revealed a consistency rate of 0.966 (95% confidence interval: 0.932–0.986). A two-sided exact binomial test yielded a p-value of 1.52×10^{-50}, indicating statistically significant consistency.

Feature selection

A total of 5,487 stable reversal gene pairs were initially selected using the REOs algorithm. Subsequently, a two-step feature selection method was applied to identify the final predictive gene pairs. As shown in Fig. 2c-d, during the preliminary screening of the two-step feature selection, MDFS-2D selected 26 gene pairs, Boruta selected 66

gene pairs, rmcfs selected 66 gene pairs, and VSOLassoBag selected 81 gene pairs. A total of 15 gene pairs were common across all four methods, and the union of gene pairs selected by the four methods contained 127 gene pairs. During the Secondary screening, the parameters of the XGBoost model were set as *learning_rate* = 0.1, *max_depth* = 4, and *n_estimators* = 150, which yielded the best performance of the classification model. The ratio of pCR to non-pCR sample numbers was calculated and assigned to the *scale_pos_weight* parameter. As shown in Fig. 2e, when the number of gene pairs reached 32, the F1 score peaked at 0.9818. The details of these 32-gene pair signature are provided in Table 1, hereafter referred to as the 32-GPS.

Predictive biomarkers for pathological complete response to nCRT

The GSE87211 dataset, consisting of 106 samples, was selected as the training cohort for this study. Internal validation was performed using the holdout method, with the training and test sets randomly split in a 7:3 ratio. The training set contained 74 samples, while the test set included 32 samples, with both sets maintaining the same proportion of positive and negative samples as the original dataset. The optimal parameter settings for the BoostForest classifier were *max_leafs* = None, *min_sample_leaf_list* = 5, *reg_alpha_list* = 0.1, and *n_estimators* = 200, with bootstrap sampling applied with replacement, through which 200 subsets were used to train 200 BoostTrees. In the case of imbalanced data, where the majority of samples are negative, the model may still exhibit a high true positive rate (TPR) and low false positive rate (FPR), resulting in a high AUC value. However, this may not accurately reflect the model's performance in identifying positive cases. The PR curve, which focuses more on the performance for the positive class, was used in this study instead of the ROC curve to evaluate the model's performance, providing a more accurate assessment of its ability to recognize the positive class in imbalanced datasets.

As shown in Fig. 3a, in the test set, the AUPRC of BoostForest was 0.983 (0.951, 1.000), with an accuracy of 0.988 (0.963, 1.000), recall of 0.95 (0.852, 1.000), specificity of 0.983 (0.951, 1.000), and an F1 score of 0.867 (0.667, 1.000). The F1 score remains lower than accuracy, which may be attributed to data imbalance. Even if the model performs well on the majority class, accuracy can still be high, whereas the F1 score is influenced by the trade-off between precision and recall. As shown in Fig. 3b, in the GSE40492 dataset, BoostForest achieved an AUPRC of 0.785, accuracy of 0.898, recall of 0.727, specificity of 0.938, and an F1 score of 0.727.

To validate the importance of BoostForest in this study, random forest (RF) was added as the baseline, and Support Vector Machine (SVM) and XGBoost were included as comparison models. All four prediction models were evaluated under the same experimental conditions. As shown in Fig. 3a, the AUPRC of all four models was 0.983 in the test set, with no significant difference, which may be attributed to the relatively small sample size of the test set, making it difficult to distinguish model performance. In the GSE40492 dataset, BoostForest achieved the highest AUPRC of 0.785. We observed that the AUPRC of BoostForest in the GSE40492 dataset is lower than its performance in the test set. This discrepancy may be attributed to the relatively small sample size of the test set. Additionally, although the test and training sets were partitioned using the holdout method, both sets originate from the GSE87211 dataset, which may have led to overfitting in the test set, thus affecting the model's performance in the GSE40492 dataset. To ensure the reliability of the results, all algorithms were executed five times on each dataset, analyzed using TOPSIS, and statistically compared using the Mann-Whitney U test between BoostForest and other models. The results, presented in Table 2, are reported as mean values. The best-performing results are highlighted in bold, and * indicates the statistically significant superiority of BoostForest. In the GSE87211 dataset, BoostForest achieved an average TOPSIS ranking of 2.2, second only to SVM. In the GSE40492 dataset, BoostForest ranked first in the TOPSIS analysis, with its F1 Score and AUPRC significantly higher than those of other models (p-value < 0.05). We found that although BoostForest had a lower recall than XGBoost in the GSE40492 dataset, it still achieved a recall of 0.7, which is above the average level. Except for recall, BoostForest exhibited significantly higher performance in all other metrics compared to XGBoost, and its TOPSIS ranking was also higher than that of XGBoost. We found that the specificity of BoostForest was significantly lower than that of Baseline (p-value < 0.05). However, BoostForest exhibited significantly higher recall, F1 score, and AUPRC compared to Baseline (p-value < 0.05). Moreover, the improvements in recall and

ID	Gene pair ($G_a > G_b$)	ID	Gene pair ($G_a > G_b$)	ID	Gene pair ($G_a > G_b$)
1	CTNNBIP1 > ZNF544	2	MFSD2A > SLC39A7	3	VAV2 > PER2
4	ANKRD20A1 > ETNK2	5	ITGB8 > C2orf49	6	RNF213 > CFH
7	CENPM > CACNA1H	8	STAC3 > PBX3	9	CDKN2A > ATG2A
10	RPP38 > ABCF3	11	CHRNA5 > CFH	12	RPP38 > PEX16
13	CENPM > PRDM2	14	SPTBN1 > ZNF175	15	ITGB8 > ZNF658
16	FUT8 > ITFG1	17	CAMTA1 > UAP1L1	18	CENPM > BEND7
19	GDAP1 > PDGFB	20	CENPM > C2orf49	21	CEP76 > SMPD1
22	KATNAL2 > RNF185	23	KATNAL2 > ZNF415	24	ITGB8 > KAT5
25	FUT8 > ARRDC1	26	CYP2C9 > RSU1	27	TNS1 > PIGL
28	CBR3 > MUC17	29	ZNF180 > CFH	30	DONSON > CHID1
31	VAV2 > DNMT3B	32	ITGB8 > RABGAP1		

Table 1. The specific description of the 32-gene pair signature. *Ga* > *Gb* represents pathological complete response to neoadjuvant chemoradiation in locally advanced rectal cancer.

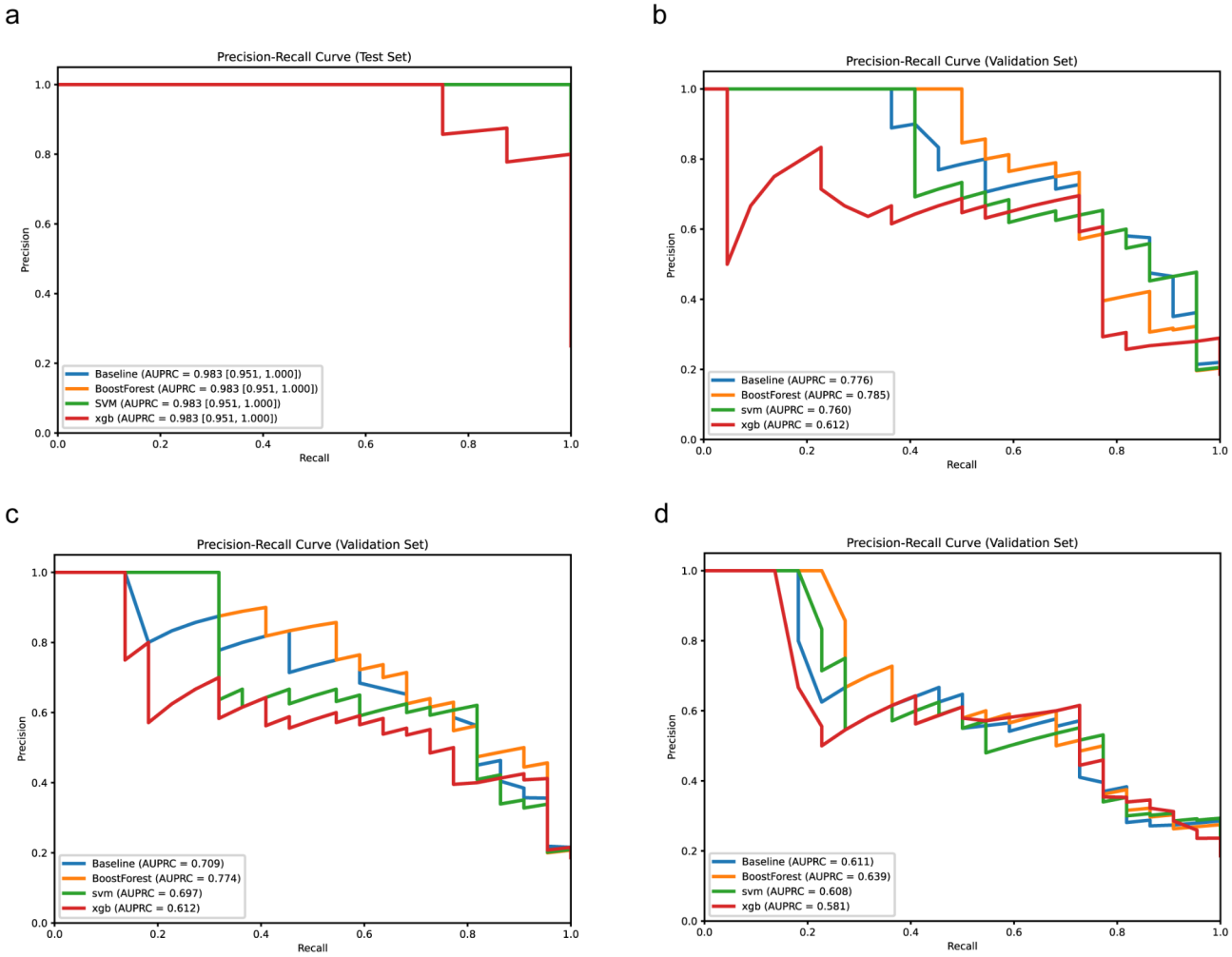


Fig. 3. Precision-recall curves of four predictive models in the test and validation cohort. **(a)** PR curves of four predictive models in the test set, with baseline representing the random forest model. **(b)** PR curves of four predictive models in the validation cohort. **(c)** AUPRC Performance of Four Prediction Models Using lasso gene pair signature (lasso-GPS) as Predictive Features in the GSE40492 Dataset. **(d)** AUPRC Performance of Four Prediction Models Using 15-shared gene pair signature (15-SGPS) as Predictive Features in the GSE40492 Dataset.

Dataset	Test				GSE40492			
Model	Baseline	SVM	XGBoost	BoostForest	Baseline	SVM	XGBoost	BoostForest
Accuracy	0.9688	**0.9880**	0.9320*	0.9654	0.8746	0.8658	0.8140*	**0.8882**
Recall	0.8400	**0.9500**	0.7000*	0.9400	0.4456*	0.5182*	**0.7730**	0.7000
Specificity	**1.0000**	**1.0000**	0.9830*	0.9966	**0.9728**	0.9460	0.8230*	0.9312
F1_Score	0.8730	**0.9670**	0.7000*	0.8798	0.5696*	0.5900*	0.6070*	**0.7020**
AUPRC	**0.9898**	0.9830	0.9830	0.9830	0.7510*	0.7600*	0.6120*	**0.7710**
Topsis_Rank	2.80	1.00	4.00	2.20	3.80	3.20	2.00	1.00

Table 2. Performance of four prediction models on the GSE87211 and GSE40492 datasets. The best-performing results are highlighted in bold, and * indicates the statistically significant superiority of BoostForest. The term "Topsis_Rank" refers to the ranking obtained using TOPSIS analysis.

F1 score were more pronounced than the reduction in specificity. These results were obtained through statistical significance testing, ensuring that the observed differences are not due to random variation.

In this study, lasso regression was also employed as an alternative feature selection method, and the features selected by lasso were compared with those selected by the two-step feature selection method, while all other conditions were kept constant. To enhance the robustness of the selected features, the 1-SE rule was applied

		Accuracy	AUPRC	Recall	Specificity	F1_Score
Test	lasso-GPS	0.975	1	0.8	1	0.867
	32-GPS	0.988	0.983	0.95	0.983	0.867
	Δ	0.013	− 0.017	0.15	− 0.017	0
GSE40492	lasso-GPS	0.864	0.774	0.318	0.99	0.467
	32-GPS	0.898	0.785	0.727	0.938	0.727
	Δ	0.034	0.011	0.409	-0.052	0.26

Table 3. Comparison of prediction results between lasso-GPS and 32-GPS. 32-GPS refers to the 32-gene pair signature, lasso-GPS refers to the lasso gene pair signature, and Δ represents the improvement of 32-GPS over lasso-GPS.

		Accuracy	AUPRC	Recall	Specificity	F1_Score
Test	15-SGPS	0.959	0.933	0.95	1	0.633
	32-GPS	0.988	0.983	0.95	0.983	0.867
	Δ	0.029	0.05	0	− 0.017	0.234
GSE40492	15-SGPS	0.839	0.639	0.364	0.948	0.457
	32-GPS	0.898	0.785	0.727	0.938	0.727
	Δ	0.059	0.146	0.363	-0.01	0.27

Table 4. Comparison of prediction results between the 15-Shared Gene Pair Signature from four feature selection methods and the 32-Gene Pair Signature in the preliminary feature selection phase. 32-GPS refers to the 32-gene pair signature, 15-SGPS refers to the 15-shared gene pair signature, and Δ represents the improvement of 32-GPS over 15-SGPS.

to determine the regularization parameter (λ), resulting in the selection of 30 features, hereafter referred to as lasso gene pair signature (lasso-GPS). Table 3 presents the performance of BoostForest on the two feature sets: in the test set, lasso-GPS exhibited lower accuracy and recall compared to 32-GPS, but demonstrated certain advantages in terms of AUPRC and specificity. However, in the validation cohort, except for specificity, 32-GPS outperformed lasso-GPS across all other evaluation metrics, particularly in the prediction of small sample sizes (pCR), where the accuracy of 32-GPS was significantly improved. These results suggest that, compared to lasso-GPS, 32-GPS demonstrates superior robustness in feature selection. In GSE40492, lasso-GPS was utilized as a predictive biomarker, and the AUPRC values of various models are presented in Fig. 3c. The performance of BoostForest surpassed that of the other four models.

To further validate the importance of secondary screening, the study used the 15 shared gene pair signature (15-SGPS), which was selected from the four feature selection methods during the preliminary screening phase, as the predictive features. The performance of BoostForest is shown in Table 4. The results indicate that, in the test set and GSE40492 dataset, except for an improvement in specificity, all other metrics showed a decline compared to the 32-GPS features, suggesting that using fewer features in the 15-SGPS has certain limitations in predictive performance. Moreover, the AUPRC performance of each prediction model using the 15-SGPS features in the GSE40492 dataset is shown in Fig. 3d. All models were run five times on the GSE40492 dataset, and the Mann-Whitney U test was used to statistically compare the performance of BoostForest with other models. The results, presented in Supplementary Table S1 (Supplementary File 1), show the mean AUPRC. In the GSE40492 dataset, BoostForest achieved a mean AUPRC of 0.6404, which was significantly higher than that of Baseline, SVM, and XGBoost (p-value < 0.05). BoostForest still demonstrated the best performance, exhibiting strong robustness and superior predictive ability. This further confirms the key role of the 32-GPS in enhancing the overall performance of the model. Compared to the 15-SGPS, the 32-GPS features provide more comprehensive information, significantly improving the model's predictive performance.

We conducted a comparative analysis of two class imbalance handling methods: SMOTE (Synthetic Minority Over-sampling Technique) and loss function weight adjustment. The results, presented in Table 5, demonstrate the performance of BoostForest under these two strategies. Specifically, in the GSE87211 dataset, the AUPRC of the loss function weight adjustment method (32_GPS) reached 0.9830, exceeding that of SMOTE (9_GPS) at 0.9666. Similarly, in the GSE40492 dataset, the AUPRC for the loss function weight adjustment method was 0.7710, significantly higher than that of SMOTE (AUPRC = 0.5632). Moreover, across multiple evaluation metrics, including Accuracy, Recall, Specificity, and F1 Score, the loss function weight adjustment method consistently outperformed SMOTE. Notably, statistical analysis confirmed that these differences were statistically significant ($p < 0.05$, denoted by "*" in Table 5). These findings indicate that, in this study, the loss function weight adjustment method is more effective than SMOTE for addressing class imbalance. We appreciate your valuable suggestions, as this analysis has strengthened our evaluation of class imbalance handling methods.

Additionally, during internal validation, we evaluated the model's performance using both the bootstrap method and 10×10-fold stratified cross-validation. All algorithms were executed five times on each dataset, and statistical comparisons between BoostForest and other models were conducted using the Mann-Whitney U test.

		Accuracy	Recall	Specificity	F1_Score	AUPRC
Test	9_GPS	0.9158*	0.6900*	0.9766*	0.7668	0.9666
	32_GPS	**0.9654**	**0.9400**	**0.9966**	**0.8798**	**0.9830**
GSE40492	9_GPS	0.7458*	0.7362	0.7480*	0.5192*	0.5632*
	32_GPS	**0.8882**	0.7000	**0.9312**	**0.7020**	**0.7710**

Table 5. Mean AUPRC of BoostForest under SMOTE and loss function weight adjustment for imbalance handling. 9_GPS represents the features selected using the SMOTE technique, while 32_GPS represents the features selected through the loss function weight adjustment method. The best-performing results are highlighted in bold. * Indicates the statistically significant superiority of 32_GPS.

The results, presented in Supplementary Table S2 (Supplementary File 1), report the mean AUPRC. Under the bootstrap method, in the GSE87211 dataset, BoostForest achieved an AUPRC of 0.972, with minimal differences compared to other models and no statistically significant differences. In the GSE40492 dataset, BoostForest obtained a mean AUPRC of 0.7536, which was significantly higher than those of Baseline, SVM, and XGBoost (p-value < 0.05). Under 10 × 10-fold stratified cross-validation, in the GSE87211 dataset, BoostForest achieved an AUPRC of 0.9868, which was significantly lower than those of Baseline and SVM. In the GSE40492 dataset, BoostForest achieved an AUPRC of 0.7692, which was significantly higher than that of XGBoost and also higher than SVM and Baseline, although the differences with SVM and Baseline were not statistically significant.

Permutation importance is also sensitive to noisy data. If the dataset contains noise or outliers, permuting the values of these features may lead to abnormal fluctuations in model performance. To further evaluate the impact of different feature importance ranking methods, we conducted additional experiments comparing ANOVA, Random Forest (Mean Decrease Gini, MDG), and permutation-based importance. As shown in Supplementary Table S3 (Supplementary File 1), the permutation-based method demonstrated superior performance in terms of accuracy, F1 score, and AUPRC.

To further assess the impact of XGBoost in secondary screening, we conducted additional experiments comparing it with Random Forest and SVM. As presented in Supplementary Table S4 (Supplementary File 1), feature selection based on XGBoost consistently achieved the best classification performance, attaining the highest accuracy, recall, and AUPRC in both the test set and external validation dataset (GSE40492). In contrast, Random Forest and SVM exhibited lower recall, particularly in the external validation dataset.

Guo et al. employed forward feature selection and a voting method[9], ultimately identifying 27_GPS. Xue et al. applied forward feature selection and a voting method combined with Random Forest's Mean Decrease Gini (MDG) for Variable Importance (VIM) ranking, resulting in 41_GPS[7]. We utilized 27_GPS and 41_GPS as predictive features and compared them with the 32_GPS identified in this study. As presented in Supplementary Table S5 (Supplementary File 1), 32_GPS outperformed both 27_GPS and 41_GPS across all evaluation metrics.

Furthermore, we applied forward feature selection combined with a voting method for feature selection and prediction, identifying a total of 10 gene pairs, and evaluated their performance. The results, detailed in Supplementary Table S6 (Supplementary File 1), demonstrate that the proposed method in this study outperforms the forward feature selection and voting strategy.

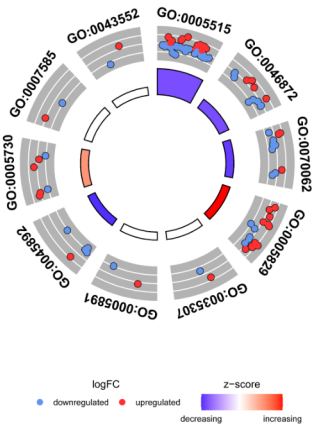
Pathway enrichment analysis of predictive biomarkers

Among the 32-GPS, a total of 51 genes were identified. The GO enrichment analysis revealed that 72.54% of these genes were enriched in protein binding, 19.61% in extracellular exosomes, and 3.92% in pathways such as the positive regulation of protein dephosphorylation, as shown in Fig. 4a-b (P.adj < 0.05). Among the genes enriched in these pathways, the number of downregulated genes was greater than that of upregulated genes. In the KEGG enrichment analysis, as shown in Fig. 4c (P.adj < 0.05), the genes in the 32-GPS were primarily enriched in pathways such as Transcriptional Misregulation in Cancer, Metabolic Pathways, and Regulation of Actin Cytoskeleton. In addition, we provide detailed information on GO enrichment pathways and KEGG enrichment pathways in Supplementary Data 2 and Supplementary Data 3.

Validation of predictive gene pairs using survival analysis

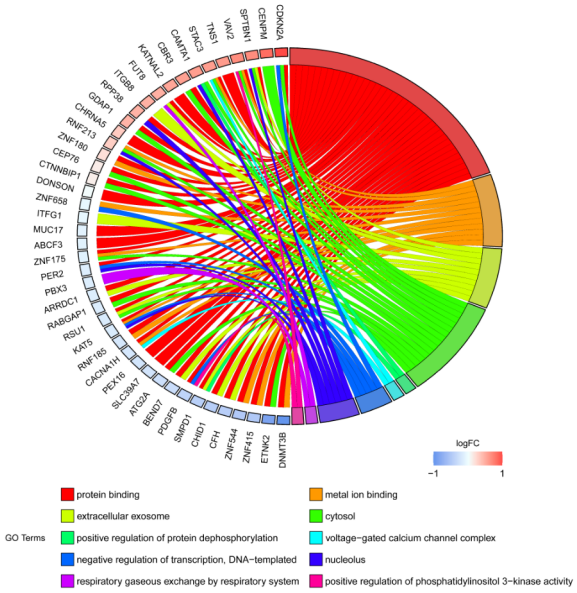
In this study, survival analysis of DFS was conducted on the patients from the GSE87211 and GSE40492 datasets. As shown in Fig. 5a, in GSE87211, the pCR group demonstrated significantly better survival probability compared to the non-pCR group, with a statistically significant difference (p-value = 0.015). Over the 100-month follow-up period, the survival probability of the pCR group remained at a high level, while the survival rate of the non-pCR group significantly declined over time. In GSE40492 (Fig. 5b), the survival rate of the pCR group was also significantly higher than that of the non-pCR group (p-value = 0.11). During the tuning process, we observed a trend where the p-value of the survival curve tended to decrease as the *min_sample_leaf_list* gradually decreased. The survival analysis results corresponding to *min_sample_leaf_list* values of 4, 3, and 2 are presented in Fig. 5c and d, and 5e, respectively. When *min_sample_leaf_list* was set to 1 (Fig. 5f), the p-value was 0.021. After performing multiple testing correction using Benjamini-Hochberg (BH) adjustment on the p-values obtained for different *min_sample_leaf_list* values, the adjusted p-value was 0.105. Although the initial analysis suggested that the DFS in the pCR group was superior to that in the non-pCR group, the adjusted p-value did not reach statistical significance after the correction for multiple comparisons.

a



ID	Description
GO:0005515	protein binding
GO:0046872	metal ion binding
GO:0070062	extracellular exosome
GO:0005829	cytosol
GO:0035307	positive regulation of protein dephosphorylation
GO:0005891	voltage−gated calcium channel complex
GO:0045892	negative regulation of transcription, DNA−templated
GO:0005730	nucleolus
GO:0007585	respiratory gaseous exchange by respiratory system
GO:0043552	positive regulation of phosphatidylinositol 3−kinase activity

b



c

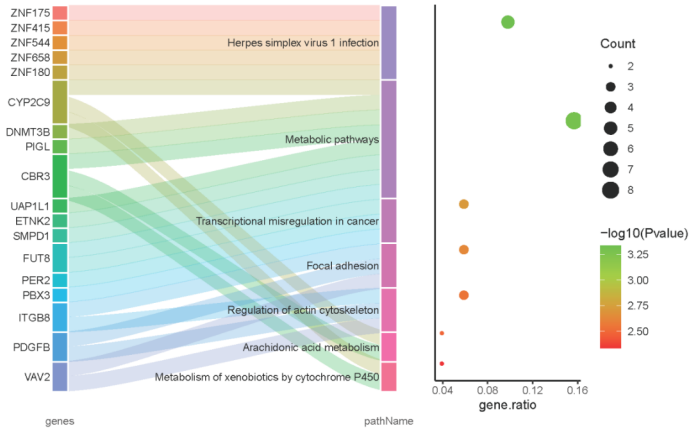


Fig. 4. Pathway enrichment analysis of genes included in the 32-GPS. **(a)** Gene enrichment distribution from the GO enrichment analysis. **(b)** Specific distribution of gene enrichment in the GO analysis. **(c)** Gene enrichment distribution from the KEGG enrichment analysis.

Interpretation of BoostForest model output using SHAP

To gain a deeper understanding of the contribution of each feature to the model's predictions, this study employed the SHAP method for interpretability analysis. The results are shown in the figure below.

As shown in Fig. 6a, the average absolute SHAP values for each feature in the test set are displayed on the left, and the SHAP value distribution plot is presented on the right. Features are ranked from top to bottom based on their importance, with higher absolute SHAP values indicating greater contribution to the model's predictions. Among them, CTNNBIP1 > ZNF544, CHRNA5 > CFH, CDKN2A > ATG2A, and ENPM > CACNA1H have the most significant impact on the model's predictions. As shown in Fig. 6b, positive SHAP values indicate that the feature contributes to increasing the probability of a positive response, while negative SHAP values suggest that the feature reduces the probability of a positive response. Figure 6c presents a heatmap of SHAP values, where red indicates a contribution to positive predictions, and blue signifies a contribution to negative predictions. Some features exhibited clear discriminative ability between pCR and non-pCR patients, with similar results observed in the GSE40492 dataset (Supplementary Figure S1).

In datasets GSE87211 and GSE40492, we performed a directional consistency validation for the 32-GPS, utilizing *limma* analysis to compute logFC. The results revealed that 25 out of 32 gene pairs exhibited consistent directionalities, yielding a consistency rate of 0.78, which was statistically significant (p-value = 0.002), as shown in Supplementary Data 4.

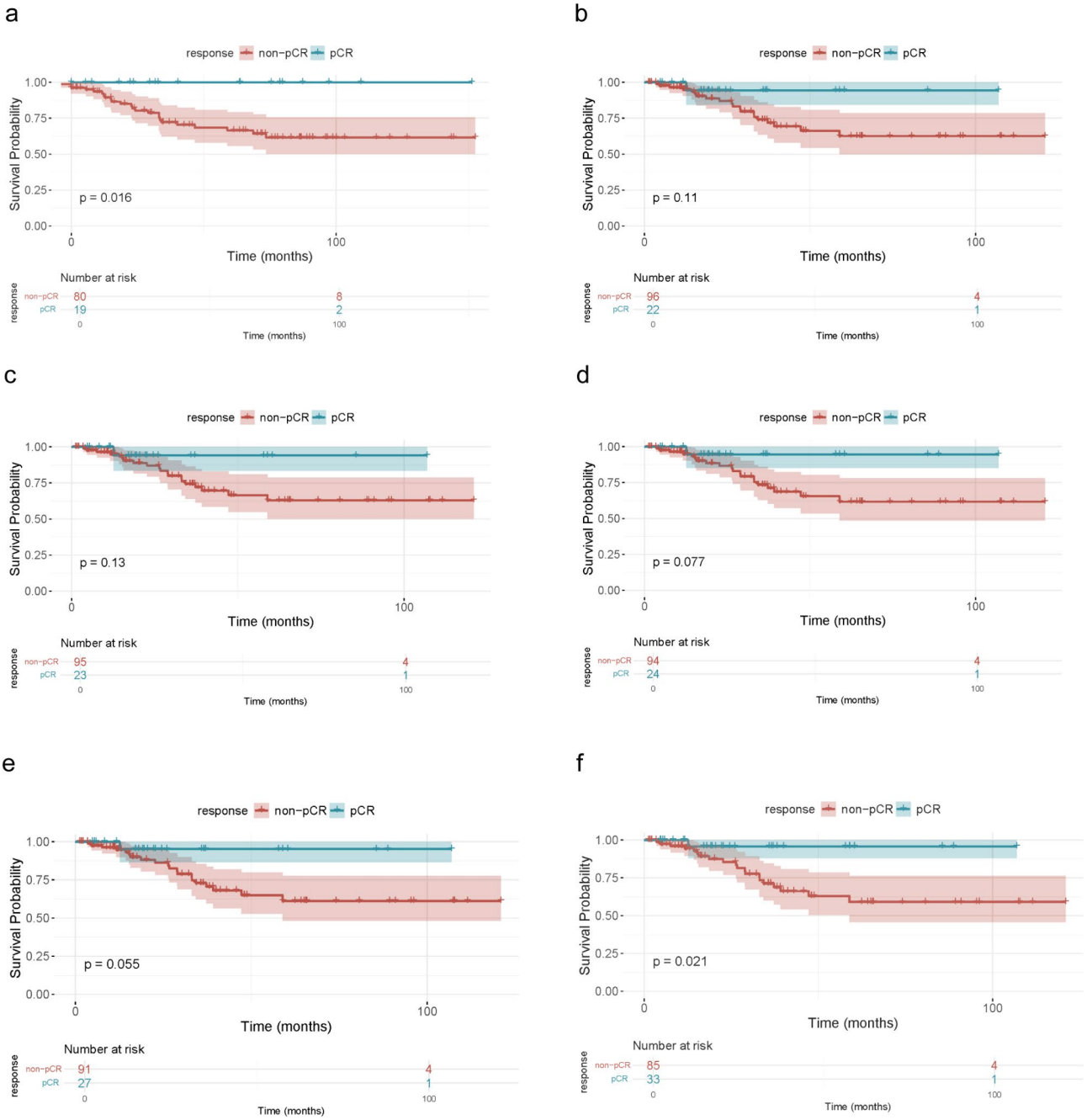


Fig. 5. Survival analysis validation of the 32-GPS. **(a)** Survival analysis of GSE87211 dataset. **(b)** Survival analysis of GSE40492 dataset (with *min_sample_leaf_list* set to 5). **(c–f)** Survival analysis of GSE40492 dataset (with *min_sample_leaf_list* set to 4, 3, 2, and 1, respectively).

Analysis of samples with mismatched prediction results and true labels

Subsequently, the distribution of samples with mismatched prediction results and true labels was analyzed. First, the 32-GPS score for each sample in the GSE40492 dataset was calculated. As shown in Fig. 6d, samples with a score below 16 were predominantly predicted as non-pCR patients (using a prediction probability threshold of 0.5), while samples with a score above 21 were all predicted as pCR patients. The samples with mismatched predictions were mainly distributed between 16 and 21 points. For these challenging-to-predict samples, the discrepancy may arise from the higher weight assigned to the minority class (pCR) in the loss function during feature selection and model training. This causes the model to be more sensitive to the minority class samples, leading the selected features to potentially reflect characteristics more associated with the minority class, which in turn increases the likelihood of misclassifying a small number of majority class (non-pCR) samples as minority class samples.

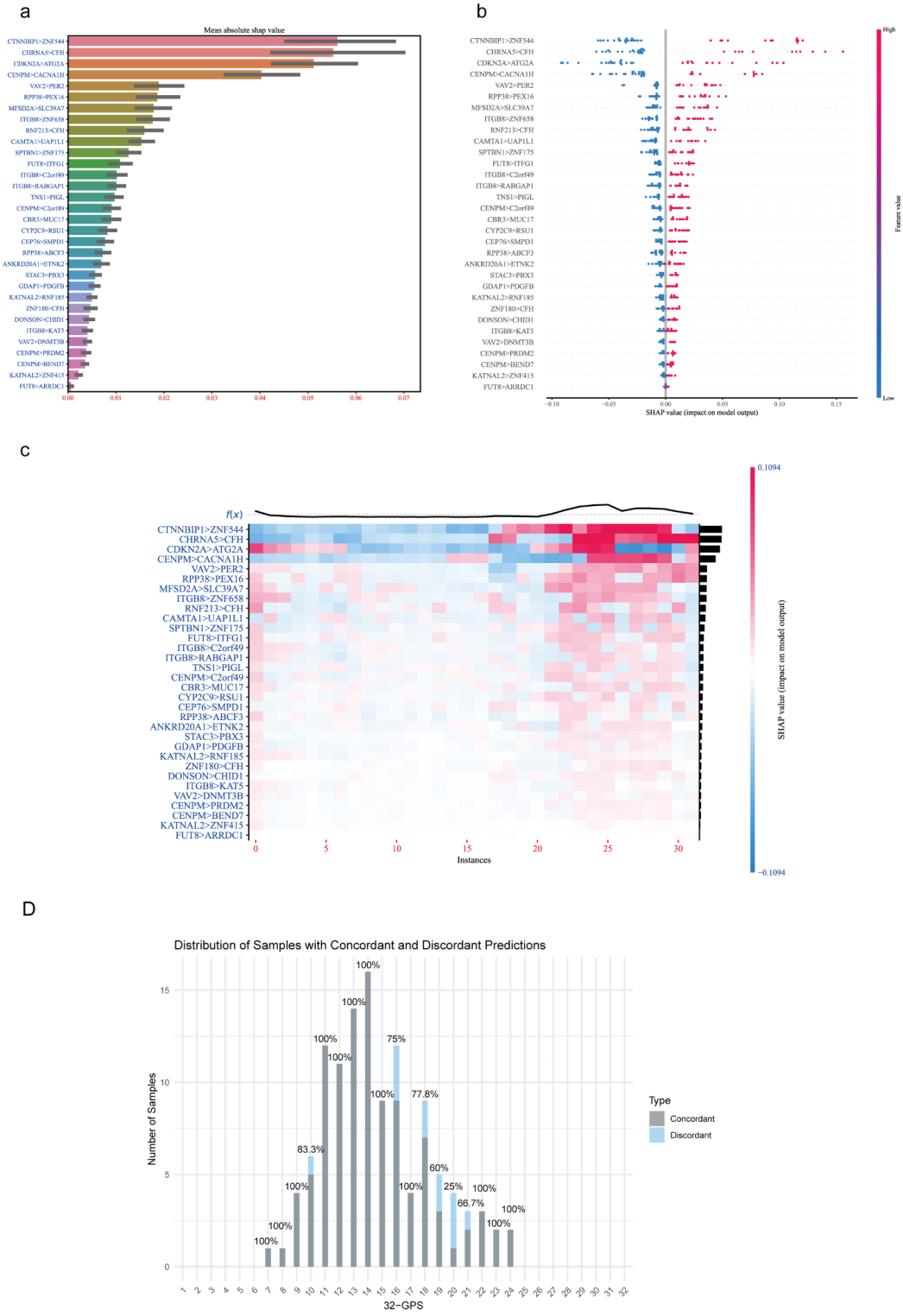


Fig. 6. Interpretability analysis of predictive model. **(a)** A bar chart presents the importance of each feature, with the left section showing the mean absolute SHAP values and the right section displaying the SHAP value distribution. **(b)** A scatter plot shows the distribution of SHAP values. **(c)** A heatmap visualizes the contribution of each feature to the predictions. **(d)** Distribution of Samples with Consistent and Inconsistent Prediction Results Compared to True Labels.

Discussion

In studies on predicting the response of LARC patients to nCRT, existing feature selection methods still face challenges in overcoming reproducibility issues, which in turn results in the inability to validate the predictive models across different datasets. This issue has attracted extensive research attention. For example, Guan et al. employed Weighted Gene Co-Expression Network Analysis (WGCNA) to identify CCT5 and ELF1, which successfully predicted treatment response in LARC patients[43]. Wang et al. used LASSO regression to identify six key indicators and construct the Systemic Inflammation-Nutritional Index (SINI)[44]. However, in these studies, single feature selection methods often struggle to capture the generalizability of features. Additionally, sample imbalance remains a limitation, as it affects predictive accuracy for positive samples. Although some studies have employed synthetic data to balance the distribution of positive and negative samples, this approach is relatively prone to overfitting[19,20]. In this study, we addressed these limitations by using REOs algorithm to construct gene expression order pairs, reducing batch effect influences. A two-step selection process was applied to identify the 32-GPS feature set, effectively predicting the response of LARC patients to nCRT. We then used BoostForest as the predictive model to enhance accuracy, particularly for pCR patients. Experimental results demonstrated that the model achieved an accuracy of 0.898 and an AUPRC of 0.785 on the validation cohort, effectively distinguishing pCR from non-pCR patients. These findings provide a reliable basis for clinical decision-making, helping to mitigate under- or overtreatment in patient care.

Filtering features that are significantly associated with treatment response enables the identification of potential biomarkers, which serve as a foundation for personalized treatment strategies. While methods such as Lasso regression, Random Forest, and Recursive Feature Elimination (RFE) are commonly used to predict treatment responses in cancer[45,46], reliance on a single method presents challenges. These methods can be sensitive to noise and outliers, potentially leading to inconsistent feature selection results across different data distributions. As a result, the generalizability and accuracy of selected feature sets may be compromised. In this study, a two-step feature selection process was implemented to address these challenges. The Preliminary screening involved applying four distinct feature selection methods, with the results integrated to form a candidate feature set. This approach reduces dependence on a single method, resulting in a more robust set of features and improving the identification of treatment-response-related features from multiple perspectives. In the Secondary screening, IFS combined with an XGBoost model was used for secondary selection, ultimately identifying 32 GPS features as the critical predictive feature set. To further enhance the reliability of the feature selection process and avoid convergence on a local optimum, permutation importance was applied to rank the candidate features. This adjustment improves both the efficiency and reliability of the selection process. Finally, to address the issue of imbalanced datasets, especially concerning the minority class (pCR), the loss function weights of the XGBoost model were adjusted to account for class imbalance, and the F1 score was used as the evaluation metric. This approach ensures more accurate predictions for the minority class.

This study employed a combination of four feature selection methods to identify features, resulting in the identification of 127 gene pairs. Specifically, MDFS-2D, Boruta, MCFS, and VSOLassoBag identified 26, 66, 66, and 81 features, respectively. Certain differences were observed in the features selected by different methods. Single feature selection methods, which are typically based on specific statistical assumptions or algorithmic principles, may lead to the omission of important features. In contrast, combining multiple feature selection methods allows for multidimensional analysis, facilitating the identification of features that may be overlooked by individual methods. This approach enables the reassessment of the importance of these features during secondary screening, thereby reducing the likelihood of missing critical features. This has been well demonstrated in the study by Wei et al.[20]. Moreover, the question of how to scientifically, accurately, and efficiently combine feature selection methods remains an intriguing area of investigation, which we plan to explore further in future studies.

While machine learning is widely applied in clinical diagnostics, achieving notable improvements in accuracy and efficiency[47,48], and models such as random forest (RF) and support vector machine (SVM) demonstrate high predictive accuracy in assessing responses to neoadjuvant chemoradiotherapy (nCRT) in patients with locally advanced rectal cancer (LARC)[49,50], results from the present study indicate that RF and SVM tend to favor the majority class (non-pCR), limiting their effectiveness for minority class predictions (pCR). To address this limitation, a more robust model, BoostForest, was introduced. With RF set as the baseline model and SVM and XGBoost included for comparison, all models demonstrated similar predictive results in the test set, while BoostForest achieved the highest AUPRC in the validation cohort, showing an overall performance advantage. Despite the strong predictive power of tree-based models, their complex decision paths and multi-level branching structures present a significant interpretability challenge in clinical practice. SHAP analysis was therefore applied to quantify each feature's contribution to the prediction results, thereby improving model transparency. This approach enhances interpretability, allowing clinicians to understand the factors underlying the model's predictions more clearly and fostering greater trust and applicability in practical diagnostics. In summary, the BoostForest model, used in conjunction with the 32-GPS system, demonstrates effectiveness in supporting clinicians' treatment decision-making by aiding in the selection of personalized treatment strategies tailored to specific patient categories, which may reduce the adverse effects associated with overtreatment.

In selecting BoostForest as the predictive model for response to neoadjuvant chemoradiotherapy, we considered several key aspects. BoostForest utilizes BoostTree as its base learner and introduces randomness into the model through both Bootstrap sampling and random selection of hyperparameters. Additionally, it enhances model diversity by randomly choosing split points during node splitting, which contributes to reducing overfitting. Previous studies, such as those by Zhao et al., have shown that BoostForest generally outperforms or at least performs comparably to established ensemble models like LightGBM and XGBoost, particularly in challenging datasets[28]. In our own research, we have demonstrated that BoostForest provides superior overall performance compared to the XGBoost model.

The limitations of this study must be acknowledged. First, although the proposed method and model have been optimized and achieved certain effectiveness, the existing sample data are insufficient to comprehensively illustrate the advantages of our approach. This limited sample size restricts the performance and potential of the method and model. Due to the relatively small size of the validation cohort (GSE40492), the dataset may only represent a subset of patients, lacking sufficient sample diversity. In future studies, we aim to validate the model using larger and more diverse cohorts. Second, although the proposed signature and model have shown potential in preliminary validation, their performance is currently insufficient to instill full confidence among clinicians, thus limiting their applicability and verification in clinical settings. Future research should include multi-center studies and clinical validation to ensure the model's robustness and applicability, thereby enhancing its impact and reliability in clinical practice.

Conclusions

This study constructed gene pairs using REOs and identified the 32-GPS from numerous gene pairs by applying a two-step selection method. BoostForest was introduced as the predictive model to forecast the response of LARC patients to nCRT. In comparison to traditional classification models such as random forests, SVM, and XGBoost, BoostForest demonstrates superior performance in terms of accuracy and predictive capability. This workflow is expected to improve the accuracy of predicting pCR patients, particularly in scenarios with imbalanced class distributions.

Data availability

The datasets used in this study are available in the online repository, the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/).

Received: 26 November 2024; Accepted: 13 March 2025
Published online: 22 March 2025

References

1. Zhao, F. et al. Neoadjuvant radiotherapy improves overall survival for T3/4 N + M0 rectal cancer patients: a population-based study of 20300 patients. *Radiat. Oncol.* **15**, 49. https://doi.org/10.1186/s13014-020-01497-4 (2020).
2. Kong, J. C. et al. Total neoadjuvant therapy in locally advanced rectal cancer: a systematic review and metaanalysis of oncological and operative outcomes. *Ann. Surg. Oncol.* **28**, 7476–7486. https://doi.org/10.1245/s10434-021-09837-8 (2021).
3. Maas, M. et al. Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. *Lancet Oncol.* **11**, 835–844. https://doi.org/10.1016/S1470-2045(10)70172-8 (2010).
4. Chatila, W. K. et al. Genomic and transcriptomic determinants of response to neoadjuvant therapy in rectal cancer. *Nat. Med.* **28**, 1646–1655. https://doi.org/10.1038/s41591-022-01930-z (2022).
5. Lopez-Campos, F. et al. Watch and wait approach in rectal cancer: current controversies and future directions. *World J. Gastroenterol.* **26**, 4218–4239. https://doi.org/10.3748/wjg.v26.i29.4218 (2020).
6. Jin, C. et al. Predicting treatment response from longitudinal images using multi-task deep learning. *Nat. Commun.* **12**, 1851. https://doi.org/10.1038/s41467-021-22188-y (2021).
7. Xue, Z. et al. A 41-Gene pair signature for predicting the pathological response of locally advanced rectal cancer to neoadjuvant chemoradiation. *Front. Med. (Lausanne)*. **8**, 744295. https://doi.org/10.3389/fmed.2021.744295 (2021).
8. Cho, E. et al. A multigene model for predicting tumor responsiveness after preoperative chemoradiotherapy for rectal cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **105**, 834–842. https://doi.org/10.1016/j.ijrobp.2019.07.058 (2019).
9. Guo, Y. et al. A qualitative signature for predicting pathological response to neoadjuvant chemoradiation in locally advanced rectal cancers. *Radiother Oncol.* **129**, 149–153. https://doi.org/10.1016/j.radonc.2018.01.010 (2018).
10. Guan, Q. et al. A qualitative transcriptional signature for the risk assessment of precancerous colorectal lesions. *Front. Genet.* **11**, 573787. https://doi.org/10.3389/fgene.2020.573787 (2020).
11. Zhou, H. et al. Evaluation of the ability of fatty acid metabolism signature to predict response to neoadjuvant chemoradiotherapy and prognosis of patients with locally advanced rectal cancer. *Front. Immunol.* **13**, 1050721. https://doi.org/10.3389/fimmu.2022.1050721 (2022).
12. Smith, F. M., Reynolds, J. V., Miller, N., Stephens, R. B. & Kennedy, M. J. Pathological and molecular predictors of the response of rectal cancer to neoadjuvant radiochemotherapy. *Eur. J. Surg. Oncol.* **32**, 55–64. https://doi.org/10.1016/j.ejso.2005.09.010 (2006).
13. Rimkus, C. et al. Microarray-based prediction of tumor response to neoadjuvant radiochemotherapy of patients with locally advanced rectal cancer. *Clin. Gastroenterol. Hepatol.* **6**, 53–61. https://doi.org/10.1016/j.cgh.2007.10.022 (2008).
14. Park, I. J. et al. Neoadjuvant treatment response as an early response indicator for patients with rectal cancer. *J. Clin. Oncol.* **30**, 1770–1776. https://doi.org/10.1200/JCO.2011.39.7901 (2012).
15. Lu, M. & Zhan, X. The crucial role of multiomic approach in cancer research and clinically relevant outcomes. *EPMA J.* **9**, 77–102. https://doi.org/10.1007/s13167-018-0128-8 (2018).
16. Casado, E. et al. A combined strategy of SAGE and quantitative PCR provides a 13-gene signature that predicts preoperative chemoradiotherapy response and outcome in rectal cancer. *Clin. Cancer Res.* **17**, 4145–4154. https://doi.org/10.1158/1078-0432.CCR-10-2257 (2011).
17. Chen, Y., Yang, B., Chen, M., Li, Z. & Liao, Z. Biomarkers for predicting the response to Radiation-Based neoadjuvant therapy in rectal cancer. *Front. Biosci. (Landmark Ed).* **27**, 201. https://doi.org/10.31083/j.fbl2707201 (2022).
18. Cui, J., Dou, X., Sun, Y. & Yue, J. Consolidation chemotherapy May improve pathological complete response for locally advanced rectal cancer after neoadjuvant chemoradiotherapy: a retrospective study. *PeerJ* **8**, e9513. https://doi.org/10.7717/peerj.9513 (2020).
19. Hu, J. et al. Construction and validation of a progression prediction model for locally advanced rectal cancer patients received neoadjuvant chemoradiotherapy followed by total mesorectal excision based on machine learning. *Front. Oncol.* **13**, 1231508. https://doi.org/10.3389/fonc.2023.1231508 (2023).
20. Wei, W., Li, Y. & Huang, T. Using machine learning methods to study colorectal cancer tumor Micro-Environment and its biomarkers. *Int. J. Mol. Sci.* **24**, 453. https://doi.org/10.3390/ijms241311133 (2023).
21. Zheng, X. et al. Less is more: relative rank is more informative than absolute abundance for compositional NGS data. *Brief. Funct. Genomics*. https://doi.org/10.1093/bfgp/elae045 (2024).
22. Zheng, H. et al. A qualitative transcriptional signature for determining the grade of colorectal adenocarcinoma. *Cancer Gene Ther.* **27**, 680–690. https://doi.org/10.1038/s41417-019-0139-1 (2020).

23. Zhang, Z. M. et al. Early diagnosis of pancreatic ductal adenocarcinoma by combining relative expression orderings with Machine-Learning method. *Front. Cell. Dev. Biol.* **8**, 582864. https://doi.org/10.3389/fcell.2020.582864 (2020).
24. Xue, Z. et al. An immuno-score signature of tumor immune microenvironment predicts clinical outcomes in locally advanced rectal cancer. *Front. Oncol.* **12**, 993726. https://doi.org/10.3389/fonc.2022.993726 (2022).
25. Guan, Q. et al. A qualitative transcriptional signature for the early diagnosis of colorectal cancer. *Cancer Sci.* **110**, 3225–3234. https://doi.org/10.1111/cas.14137 (2019).
26. Zhu, Z., Wang, Z., Li, D., Zhu, Y. & Du, W. Geometric structural ensemble learning for imbalanced problems. *IEEE Trans. Cybern.* **50**, 1617–1629. https://doi.org/10.1109/TCYB.2018.2877663 (2020).
27. Zhao, D. et al. Whale optimized mixed kernel function of support vector machine for colorectal cancer diagnosis. *J. Biomed. Inf.* **92**, 103124. https://doi.org/10.1016/j.jbi.2019.103124 (2019).
28. Zhao, C. et al. BoostTree and BoostForest for ensemble learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 8110–8126. https://doi.org/10.1109/TPAMI.2022.3227370 (2023).
29. Hu, Y. et al. Colorectal cancer susceptibility loci as predictive markers of rectal cancer prognosis after surgery. *Genes Chromosomes Cancer* **57**, 140–149. https://doi.org/10.1002/gcc.22512 (2018).
30. Gaedcke, J. et al. Molecular markers for predicting treatment outcome in patients with rectal cancer: a comprehensive analysis from the german rectal cancer trials. *GEO* (2023). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40492.
31. PDQ Adult Treatment Editorial Board. Rectal cancer treatment (PDQ(R)): health professional version in PDQ Cancer Information Summaries (National Cancer Institute (2024).
32. Kim, S. H. et al. What is the ideal tumor regression grading system in rectal cancer patients after preoperative chemoradiotherapy?? *Cancer Res. Treat.* **48**, 998–1009. https://doi.org/10.4143/crt.2015.254 (2016).
33. Radosław et al. MDFS: multidimensional feature selection in R. *R J.* **11**, 198 (2019).
34. Miron, B., Kursa, Jankowski, A. & Rudnicki, W. R. Boruta—a system for feature selection. *Fundam Inf.* **101**, 271–285 (2010).
35. Michał, D. & Koronacki Jacek. Rmcfs: an R package for Monte Carlo feature selection and interdependency discovery. *J. Stat. Softw.* **85**, 1–28. https://doi.org/10.18637/jss.v085.i12 (2018).
36. Liang, J. et al. VSOLassoBag: a variable-selection oriented LASSO bagging algorithm for biomarker discovery in omic-based translational research. *J. Genet. Genomics.* **50**, 151–162. https://doi.org/10.1016/j.jgg.2022.12.005 (2023).
37. Huan, L. & Setiono, R. Incremental feature selection. *Appl. Intell.* **9**, 217–230. https://doi.org/10.1023/A:1008363719778 (1998).
38. Fisher, A., Rudin, C. & Dominici, F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **20**, 896 (2019).
39. Zuo, D. et al. Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Med. Inf. Decis. Mak.* **23**, 276. https://doi.org/10.1186/s12911-023-02377-z (2023).
40. Bu, D. et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* **49**, W317–W325. https://doi.org/10.1093/nar/gkab447 (2021).
41. Scott, L. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874 (2017).
42. Sir, E. & Batur Sir, G. D. Evaluating treatment modalities in chronic pain treatment by the multi-criteria decision making procedure. *BMC Med. Inf. Decis. Mak.* **19**, 191. https://doi.org/10.1186/s12911-019-0925-6 (2019).
43. Guan, B., Xu, M., Zheng, R., Guan, G. & Xu, B. Novel biomarkers to predict treatment response and prognosis in locally advanced rectal cancer undergoing neoadjuvant chemoradiotherapy. *BMC Cancer.* **23**, 1099. https://doi.org/10.1186/s12885-023-11354-8 (2023).
44. Wang, Y. et al. Prediction and validation of pathologic complete response for locally advanced rectal cancer under neoadjuvant chemoradiotherapy based on a novel predictor using interpretable machine learning. *Eur. J. Surg. Oncol.* **50**, 108738. https://doi.org/10.1016/j.ejso.2024.108738 (2024).
45. Zhang, Z. Y. et al. Metabolic reprogramming-associated genes predict overall survival for rectal cancer. *J. Cell. Mol. Med.* **24**, 5842–5849. https://doi.org/10.1111/jcmm.15254 (2020).
46. Yi et al. Research on radiotherapy related genes and prognostic target identification of rectal cancer based on multi-omics. *J. Translational Med.* **21**, 856. https://doi.org/10.1186/s12967-023-04753-9 (2023).
47. Wang, Z. et al. Machine learning model for prediction of low anterior resection syndrome following laparoscopic anterior resection of rectal cancer: a multicenter study. *World J. Gastroenterol.* **29**, 2979–2991. https://doi.org/10.3748/wjg.v29.i19.2979 (2023).
48. Shu, P. et al. An immune-related gene prognostic prediction risk model for neoadjuvant chemoradiotherapy in rectal cancer using artificial intelligence. *Front. Oncol.* **14**, 1294440. https://doi.org/10.3389/fonc.2024.1294440 (2024).
49. Chen, K. A. et al. Prediction of pathologic complete response for rectal cancer based on pretreatment factors using machine learning. *Dis. Colon Rectum* **67**, 387–397. https://doi.org/10.1097/DCR.0000000000003038 (2024).
50. Qian, L., Lai, X., Gu, B. & Sun, X. An Immune-Related gene signature for predicting neoadjuvant chemoradiotherapy efficacy in rectal carcinoma. *Front. Immunol.* **13**, 784479. https://doi.org/10.3389/fimmu.2022.784479 (2022).

Author contributions

CQ and YG: Conceptualization; CQ, SY, and YG: Methodology; CQ: Data analysis, original draft writing; CQ, RG, SY, HW and YG: Review and editing; YG: Funding acquisition; YC, FS, and CL: revised the article; All authors have read and agreed to the final version of the manuscript.

Funding

This study was supported by the National Natural Science Foundation of China (NSFC) under Grant 82060618, Grant 72261018, Jiangxi Province Key Laboratory of Multidimensional Intelligent Perception and Control (2024SSY03161), Doctoral Fund of First Affiliated Hospital of Gannan Medical University.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-94337-y.

Correspondence and requests for materials should be addressed to H.W. or Y.G.