Lei *et al. Journal of Translational Medicine* (2026) 24:376
https://doi.org/10.1186/s12967-026-07847-2

Journal of Translational Medicine

RESEARCH Open Access

A machine learning-based predictive model for radiosensitivity in nasopharyngeal carcinoma utilizing serum proteomics

Mingjun Lei[1], Li Xu[2], Pengzhen Wei[2], Liangfangshen[1] and Zhanzhan Li[1,3*]

Abstract

Background Nasopharyngeal carcinoma (NPC) remains highly sensitive to radiotherapy; however, radioresistance in a subset of patients leads to local recurrence and distant metastasis. Serum proteomics provides a minimally invasive approach to capturing dynamic physiological changes, and machine learning enables efficient construction of predictive models. This study aimed to develop and validate a serum proteomics–based machine-learning model for predicting radiotherapy sensitivity in nasopharyngeal carcinoma (NPC).

Methods Pretreatment serum samples from newly diagnosed NPC patients were analyzed using SELDI-TOF-MS. Differentially expressed proteins between radiosensitive and radioresistant groups were identified using limma. GO and KEGG analyses were performed to explore functional enrichment. Twelve machine-learning algorithms were used to construct predictive models, and the top-performing models were optimized through feature selection. A Random Forest model with seven features was identified as the optimal model. External validation was performed using an independent cohort with ELISA-quantified protein levels. Model performance was assessed using Receiver operating characteristic curve (ROC), calibration analysis, decision curve analysis (DCA), and 10-fold cross-validation. SHapley Additive exPlanations (SHAP) analysis was applied for model interpretability, and the final model was deployed via a ShinyAPP.

Results A total of 96 differentially expressed proteins were identified, which involved multiple function and signaling pathways. The Random Forest model demonstrated the best predictive performance, achieving an area under the curve (AUC) of 0.963 in the training set and 0.975 in the validation set. Cross-validation yielded an average AUC of 0.965. DCA indicated high clinical utility across a broad threshold range, and calibration curves showed good model agreement. Seven proteins (PLXND1, GSR, PGD, PTPRC, OR2T29, ACTG2, CHAD) were selected as final features. SHAP analysis provided global and individual-level interpretability. A web-based tool was developed to facilitate clinical application.

Conclusion This study establishes a robust serum proteomics–based machine-learning model capable of accurately predicting radiotherapy sensitivity in NPC. The model offers clinical interpretability and practical implementation, supporting personalized radiotherapy decision-making.

*Correspondence:
Zhanzhan Li
lizhanzhan@csu.edu.cn

Full list of author information is available at the end of the article

Keywords Nasopharyngeal carcinoma, Radiosensitivity, Machine learning, Proteomics, Prediction model

Background

Nasopharyngeal carcinoma (NPC) is a malignant tumor arising from the mucosal epithelium or glandular structures of the nasopharynx [1]. More than 90% of NPC patients are associated with Epstein–Barr virus (EBV) infection; therefore, plasma EBV DNA detection, in combination with other auxiliary methods, can be used for early screening [2]. Compared with other head and neck malignancies, NPC is more sensitive to ionizing radiation, particularly World Health Organization type II and type III subtypes. As a result, radiotherapy remains the cornerstone of treatment [3]. With advances in computing technology, radiotherapy techniques have rapidly evolved, and the application of novel approaches such as intensity-modulated radiotherapy has markedly improved treatment outcomes. A proportion of patients with early-stage disease can achieve clinical cure with radiotherapy alone [4]. However, due to the nonspecific nature of early symptoms, most patients are diagnosed at locally advanced stages, and treatment responses in some cases remain suboptimal [5]. Radioresistance, leading to local recurrence and distant metastasis, is a major obstacle to improving prognosis. Although increasing radiation doses may enhance local tumor control and reduce the risk of distant metastasis, it also raises the likelihood of radiation-induced damage to normal tissues such as the brainstem and spinal cord, resulting in severe acute and late toxicities [6]. Recent data indicate that the 5-year disease-specific survival rate for stage I–IVB NPC has exceeded 80%, yet approximately 15–20% of patients still experience local recurrence or distant metastasis [7]. Although the risk factors and mechanisms underlying NPC tumorigenesis have been partially elucidated, the specific mechanisms of radiotherapy resistance remain inadequately understood [8]. With the development of genomics, proteomics, and bioinformatics, NPC research is entering a new era. Leveraging big data and artificial intelligence allows for more accurate risk prediction, diagnostic support, assessment of therapeutic response, and the development of personalized treatment strategies. Therefore, advancing research on NPC not only deepens our understanding of the disease but also provides new directions and approaches for its prevention and management.

Proteomics enables the rapid and precise detection of differentially expressed proteins in samples such as serum, cerebrospinal fluid, and diseased tissues, thereby distinguishing healthy individuals from patients [9]. With the continued progress of human genome and post-genomic research, proteomics-based platforms have become convenient tools for disease diagnosis. This technology supports high-throughput analysis and offers high image resolution, allowing systematic screening of the overall protein composition of a sample and providing deeper insights into relevant biological processes. As a powerful and comprehensive protein-screening approach, proteomics facilitates the identification of disease-specific proteins, elucidation of the molecular mechanisms underlying disease onset and progression, and discovery of potential therapeutic targets [10]. On this basis, more specific diagnostic biomarkers and therapeutic agents can be identified, supporting early diagnosis, differential diagnosis, treatment, and prognostic evaluation. With the rapid development of artificial intelligence, machine learning and deep learning algorithms have been increasingly applied in the medical field, demonstrating outstanding performance, particularly in predictive modeling [11]. By integrating multidimensional information—including clinical data, imaging features, radiotherapy dose parameters, and biomarkers—artificial intelligence technologies can substantially improve the accuracy and clinical utility of disease risk prediction [12]. Machine learning has been widely used in developing predictive and prognostic models for various cancers [11, 13]. However, in nasopharyngeal carcinoma, multigene predictive models based on proteomics for radiotherapy sensitivity have not yet been developed or applied.

Serum samples offer the advantages of being easy to obtain while continuously and rapidly reflecting dynamic physiological states. In the present study, we employed surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS) to analyze serum proteomic differences between radiotherapy-sensitive and radiotherapy-resistant patients with nasopharyngeal carcinoma. Proteins associated with radiotherapy sensitivity were identified, and based on these differential proteins, twelve machine-learning algorithms were used to construct serum-proteomics–based predictive models for radiotherapy sensitivity in nasopharyngeal carcinoma. Additional datasets were then used to validate the performance of these models. This study is expected to facilitate the selection of appropriate radiotherapy strategies for NPC patients and ultimately improve clinical outcomes and prognosis.

Methods and materials

Patients and samples

This study is part of a nasopharyngeal carcinoma (NPC) big-data research project initiated by our institution in 2016. We enrolled patients who were newly diagnosed with NPC between September 2018 and September

2019 and were receiving radiotherapy for the first time (n = 135). Prior to radiotherapy, 10 mL of venous blood was collected from each patient and stored at − 80 °C for subsequent serum proteomic analysis. The quantitative label-free proteomics using ultra-definition mass spectrometry of detecting the serum protein were described in previous study [14]. Briefly, serum samples were clarified by centrifugation, the top 12 high-abundance proteins were depleted, and total protein concentration was quantified by BCA kit. Proteins were reduced and alkylated, ultrafiltered, and digested overnight with trypsin; peptides were then quantified, desalted, and TMT-labeled. Labeled peptides were fractionated by high-pH reverse-phase high performance liquid chromatography fractionation, combined into nine fractions, and analyzed by nanoLC–MS/MS on an Orbitrap Fusion Lumos using a data-dependent acquisition method. Raw spectra were searched in MaxQuant against the human database with standard trypsin/P settings, specified modifications, and an false discovery rate < 1%. At the same time, another cohort of patients was collected as an independent validation dataset for model construction (n = 164). The inclusion criteria were as follows: patients with NPC who had not received radiotherapy, surgery, chemotherapy/immunotherapy, or other treatments; and age ≥ 18 years. The exclusion criteria were patients with severe infectious diseases or concurrent malignancies; those with severe hepatic, renal, or other major organ dysfunction; those with immune system disorders; those with serious hematological conditions; or those who had previously received any anticancer therapy.

Data collection and definition

We collected patients' age, sex, degree of differentiation, pathological stage, gross tumor volume (GTV) of the primary nasopharyngeal lesion, pretreatment lymph node GTV, lymph node remission rate, overall tumor shrinkage rate, routine hematological parameters (counts of red blood cells, white blood cells, neutrophils, and lymphocytes), as well as levels of hemoglobin, alpha-fetoprotein (AFP), carcinoembryonic antigen (CEA), CA125, CA242, CA19-9, EBV DNA, EA-IgA, and VCA-IgA from the hospital medical record system. Tumor biomarkers were quantitatively measured using chemiluminescence assays (Immulite; DPC, USA), whereas EBV DNA was quantified by fluorescent PCR. Radiotherapy sensitivity was subsequently evaluated based on the reduction in the maximum cross-sectional area of the primary lesion, calculated as the product of the maximum transverse and longitudinal diameters. A reduction of > 50% was defined as radiotherapy-sensitive, while ≤ 50% was defined as radiotherapy-resistant [15]. Magnetic resonance imaging (MRI) was performed at baseline (pre-radiotherapy) 6 weeks after completion of radiotherapy. Tumor response was evaluated according to the Response Evaluation Criteria in Solid Tumors (RECIST) based on changes in the primary lesion [16], as follows: complete disappearance of the primary lesion was defined as complete remission (CR); a > 30% decrease in lesion size was defined as partial remission (PR); a < 30% decrease was considered stable disease (SD); and a > 20% increase was considered progressive disease (PD). PR, SD, and PD were collectively categorized as non-CR. To ensure standardization, imaging measurements and response assessments in both the training and validation cohorts were performed by the same senior radiology technologist, following a consistent protocol. Importantly, the assessor was blinded to patients' radiosensitivity status at the time of evaluation. According to the criteria, the training cohort consisted of 49 radiotherpay sensitive patients and 86 radiotherapy resistant patients. The testing cohorts included 65 and 99 patients, respectively.

Analysis of differentially expressed proteins

To identify differentially expressed proteins, we analyzed the proteomic data using the limma package [17]. The voom function was used to transform raw counts for linear modeling, followed by fitting an empirical Bayes model to evaluate expression differences. The treat method set a minimum fold-change threshold ($|\log_2 FC| > 0.01$). P-values were adjusted using the Benjamini–Hochberg method, and genes with a false discovery rate (FDR) < 0.05 that met the fold-change cutoff were considered significantly differentially expressed.

GO enrichment and KEGG pathway analysis

We performed Gene Ontology (GO) functional enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis on the set of significantly differentially expressed genes [18, 19]. All analysis was conducted in the R environment. The clusterProfiler package was used to analyze GO terms (including biological process, molecular function, and cellular component) and to perform KEGG pathway annotation and enrichment. P-values were adjusted with the Benjamini–Hochberg method, and terms or pathways with an FDR < 0.05 were considered significantly enriched. The results were visualized with enrichplot and related packages.

Development of machine learning model of radiosensitivity prediction in NPC

Using the training dataset, we built a predictive model for radiotherapy sensitivity in nasopharyngeal carcinoma, using radiotherapy response as the dependent variable and the selected differentially expressed proteins as independent variables. We applied 12 machine-learning algorithms to develop the models and calculated the Area Under the Curve (AUC) and its 95% Confidence Interval

(95% CI). The methods included: Logistic Regression (LR), Neural Network (NN), Bayesian Method (Bayes), Support Vector Machine with Kernel Function (SVM-Kernel), Gradient Boosting Machine (GBM), Adaptive Boosting (AdaBoost), Linear Discriminant Analysis (LDA), Partial Least Squares Model (PLS), Least Absolute Shrinkage and Selection Operator (LASSO) k-Nearest Neighbors (k-NN), Boosting Method (Boosting), and Random Forest (RF).

Feature selection and identification of final model

Based on the AUC values, we first selected the top five models with include features. For each of these top five models, we then ranked all included features by their importance, rebuild the models and recalculated the AUCs using decreasing numbers of features by one each time in the training dataset. The number of retained features was constrained to a predefined range (minimum 5, maximum 15) to balance parsimony and stability. Specifically, using fewer than five features may lead to underfitting and instability (i.e., overly sensitive selection driven by sampling variation), whereas allowing substantially more features in a modest training cohort ($n = 135$) increases model complexity and overfitting risk, potentially inflating training AUC while reducing generalizability. In addition, a 5–15 feature signature remains practical for downstream experimental verification and potential clinical implementation. Finally, we identified the model that achieved a relatively high AUC while using the fewer features, and this model was chosen as the final optimal model. The evaluation parameters including sensitivity, specificity, positive predict value (PPV), negative predict value (NPV), accuracy, and F1 score were calculated for final model.

Validation and assessment of radiosensitivity prediction model

After determining the final number of features in the selected optimal model, we measured the protein levels of these model features in the validation cohort using ELISA kits. We then applied the final model developed from the training dataset to evaluate its external predictive performance. The 10-fold cross validation was also performed to avoid problems with overfitting. In addition, we conducted decision curve analysis and calibration curve analysis (using a clamping/smoothing approach) in both the training and validation sets to assess the model's predictive accuracy and clinical utility.

Explanation and interactive application of model

SHapley Additive exPlanations (SHAP) analysis was applied to interpret feature contributions and assess variable importance [20]. The SHAP method provides both global and local interpretability for the model. The global explanation yields consistent and accurate attribution values for each feature in the model, demonstrating the association between the input features and radiotherapy sensitivity. The local explanation allows for specific, patient-level predictions by providing the attributed prediction for each individual data input. Finally, to make the model more convenient for clinical application, we deployed the final model into a web application using ShinyAPP (https://www.shinyapps.io/). Clinicians can now estimate a patient's radiotherapy resistance risk probability simply by inputting the specific numerical values for each feature.

Statistical analysis

All analyses were performed using the R 4.5.1 platform. For continuous data, normally distributed variables were expressed as the mean±standard deviation, and comparisons between the two groups were performed using the independent samples t-test. Data that did not conform to a normal distribution were presented as the median and interquartile range (IQR), and comparisons between groups were conducted using non-parametric tests. Categorical data were presented as counts and percentages, and comparisons between the two groups were performed using the chi-square test. Area Under the Receiver Operating Characteristic Curve (AUCs) were used to evaluate the predictive ability of the model. The differences between the AUCs were compared using DeLong's test [21]. A two-sided P-value < 0.05 was considered statistically significant.

Results

Identification of differentially expressed proteins

Figure 1 present study workflow for the development of a serum proteomics–based machine learning model to predict radiotherapy sensitivity in NPC. Based on the predefined screening criteria, we identified a total of 96 differentially expressed proteins between the radioresistant and radiosensitive groups, including 54 upregulated and 42 downregulated proteins (Table S1, Fig. 2A). Several top differentially expressed protein s are presented in Fig. 2B. Gene Ontology (GO) enrichment analysis indicated that these genes encoding these proteins were primarily involved in the acute inflammatory response, negative regulation of hydrolase and peptidase activities, collagen-containing extracellular matrix, blood microparticles, endoplasmic reticulum lumen, glycosaminoglycan, sulfur compound and heparin binding, and endopeptidase inhibitor activity (Fig. 2C). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis further revealed that these genes were mainly enriched in complement and coagulation cascades, ECM–receptor interaction, focal adhesion, PI3K–Akt signaling pathway,

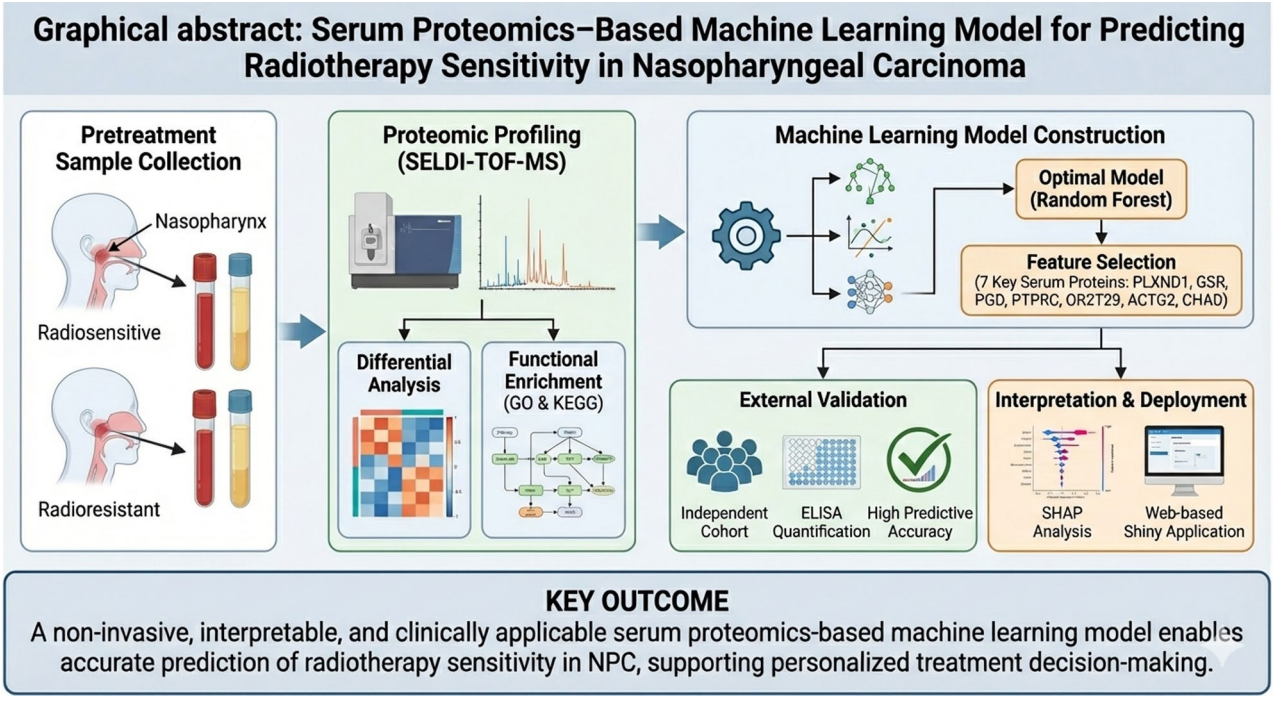

Fig. 1 Study workflow for developing and validating a serum proteomics–based machine-learning model to predict radiotherapy sensitivity in nasopharyngeal carcinoma (NPC)

pentose phosphate pathway, HIF-1 signaling pathway, and PPAR signaling pathway (Fig. 2D).

Development of machine learning model of radiosensitivity prediction in NPC

We focused model development on proteomic features because most clinical characteristics showed no statistically significant differences between the radiosensitive and radioresistant groups (Table S3). Using 96 DGEs and 12 machine learning algorithms, we constructed predictive models for radiotherapy sensitivity of nasopharyngeal carcinoma in training dataset (Fig. 3A). The top five models with the highest predictive performance were Random Forest (AUC: 0.968, 95% CI: 0.939–0.998), SVM (AUC: 0.965, 95% CI: 0.936–0.994), PLS (AUC: 0.963, 95% CI: 0.932–0.994), LASSO (AUC: 0.959, 95% CI: 0.927–0.992), Boosting (AUC: 0.956, 95% CI: 0.917–0.995). These were followed by k-NN (AUC: 0.951, 95% CI: 0.906–0.995), AdaBoost (AUC: 0.949, 95% CI: 0.905–0.993), GBM (AUC: 0.949, 95% CI: 0.907–0.991), Bayes (AUC: 0.942, 95% CI: 0.907–0.991), NN (AUC: 0.938, 95% CI: 0.892–0.984), LDA (AUC: 0.850, 95% CI: 0.779–0.922), and LR (AUC: 0.839, 95% CI: 0.767–0.910). The other parameters were presented in Table S2. These models incorporated 15 features: PLXND1, PGD, GSR, PTPRC, OR2T29, BMP1, ACTG2, ADIPOQ, CHAD, BASP1, WDFY4, INHBC, SAA2, CCL18, and S100A9. For each of these top five models, we ranked 15 features by importance and progressively reduced the number of features by one. The number of retained features was constrained to a predefined range (minimum 5, maximum 15) to balance parsimony and stability. For each reduced set, we reconstructed the top five models and recalculated their AUC values. The Random Forest model demonstrated consistently robust performance across different number features (Fig. 3B). When the number of features was reduced to seven, the Random Forest model maintained an AUC of 0.963, which remained higher than the AUCs of the other four models (0.948 for SVM, 0.952 for PLS, 0.951 for LASSO, and 0.959 for Boosting). Although the Random Forest model achieved a slightly higher AUC of 0.971 when 14 features were used, this improvement was not statistically significant compared with using seven features ($\Delta AUC = 0.008$, $P > 0.005$). Therefore, we selected seven features as the final feature set: PLXND1, GSR, PGD, PTPRC, OR2T29, ACTG2, and CHAD. The trends of various evaluation metrics for the Random Forest model under different feature counts are presented in Fig. 3C.

Validation and assessment of radiosensitivity prediction model

We next validated the Random Forest model containing the seven selected features. First, a 10-fold cross-validation was performed, which yielded an average AUC of 0.965. We then validated the model using an independent dataset collected during a different period, and the validation set and the training set are similar in

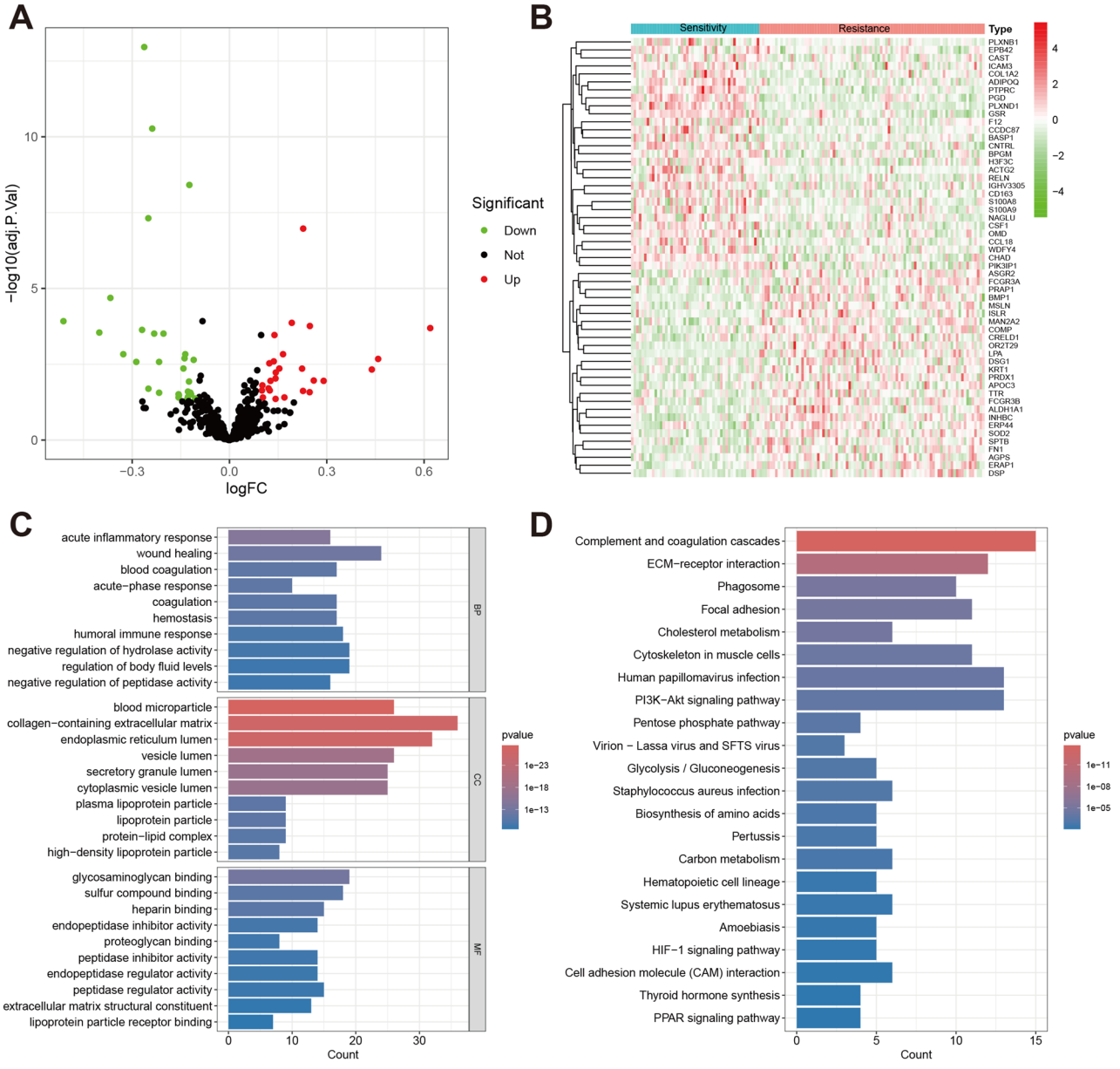

Fig. 2 Differentially expressed proteins between radiosensitivity and radioresistant NPC patients. **A**: Volcano plot showing upregulated and downregulated proteins. **B**: Heatmap of tope differentially expressed proteins associated with radiosensitivity. **C** and **D**: GO enrichment and KEGG pathway analysis based on the differentially expressed proteins genes

general clinical characteristics and are comparable (Table S4). The results showed that the testing set achieved an AUC of 0.975, indicating strong predictive performance (Fig. 4A and B). Subsequently, we conducted decision curve analysis (DCA) to assess the clinical utility of the model in both the training and testing sets. The results demonstrated that, across a threshold probability range of 0.1–1.0, patients with NPC would benefit from the intervention guided by this model (Fig. 4C and D). Calibration analysis further showed that the Random Forest model exhibited good agreement between predicted and observed outcomes in both the training and testing sets (Fig. 4E and F).

Explanation and interactive application of model

To enhance understanding and facilitate the use of the model, we applied SHAP analysis to quantify the contribution of each feature protein and to interpret the model outputs. We performed both global interpretation of feature importance and local interpretation at the individual level. Global interpretation describes the overall behavior of the model. As shown in Fig. 5A, we calculated the mean SHAP value for each feature protein to assess its

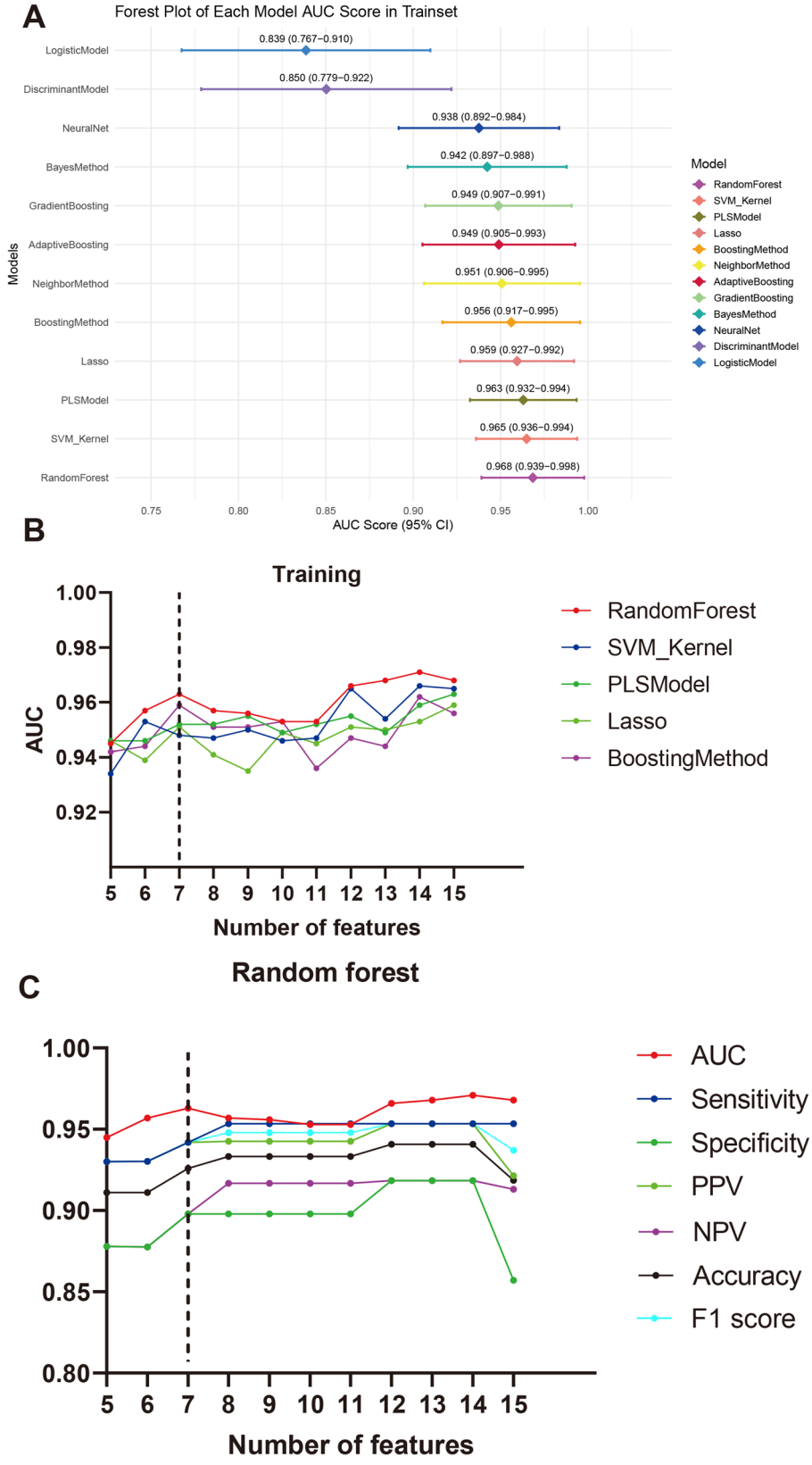

Fig. 3 Identification of the optimal prediction model for radiosensitivity in NPC. **A**: Forest plot summarizing the AUCs of the evaluated machine learning algorithms. **B**: AUCs of top five models across different numbers of retained features. **C**: Changes in key random forest model parameters across different features numbers

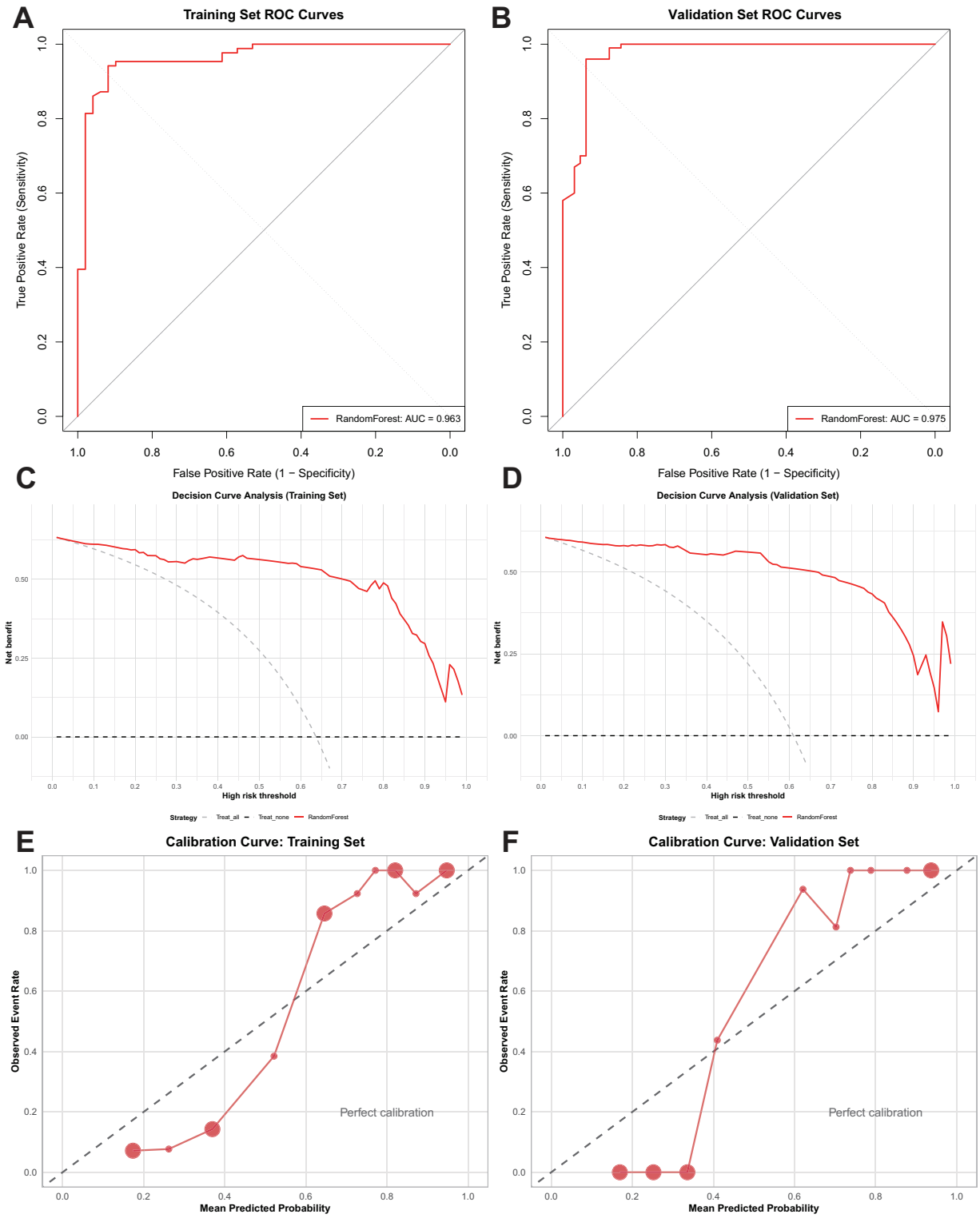

Fig. 4 Performance of random forest model predicting radiosensitivity. **A** and **B**: ROC curves and AUCs in training and validation cohorts. **C** and **D**: Decision curve analysis (DCA) in training and validation cohort. **E** and **F**: Calibration plots o in training and validation cohorts

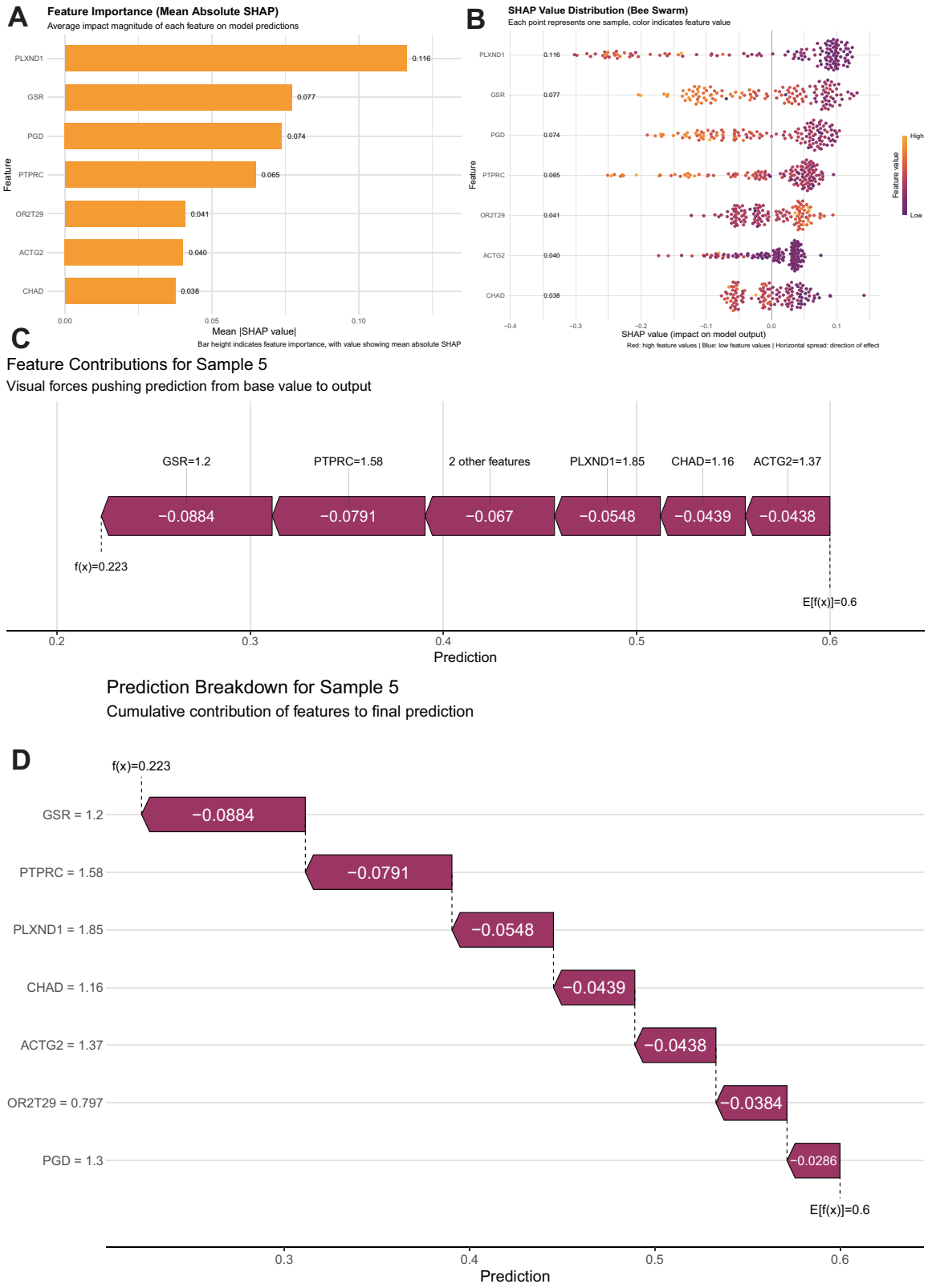

Fig. 5 Global model explanation using Shapley additive explanations (SHAP). **A**: SHAP bar plot that showed the contributions of each feature to the model. **B**: SHAP dot plot showed the probability of radioresistance increased with the SHAP values of the features. **C**: Visual forces pushing prediction from base value to output. **D**: SHAP waterfall plot shows the cumulative contribution of features to final prediction model

contribution. The seven feature proteins ranked from highest to lowest contribution were: PLXND1, GSR, PGD, PTPRC, OR2T29, ACTG2, and CHAD. In addition, the SHAP summary beeswarm plot illustrates both the direction and magnitude of each feature's influence on the model prediction. For example, higher expression levels of PLXND1, GSR, PGD, PTPRC, and ACTG2 were clearly associated with an increased risk of radiotherapy resistance in NPC patients (Fig. 5B). The visual forces and waterfall plots (Fig. 5C and D) showed the contribution of each feature to the model prediction of radioresistance for the fifth patient with NPC. The specific value of each feature and its corresponding SHAP value in the plot indicate the positive and negative impact of the feature on the prediction model. A GSR level of 1.2 and PTPRC of 1.58 made significant negative contributions of -0.0884 and −0.0791 to the prediction result, respectively. Similarly, a PLXND1 of 1.85, CHAD of 1.16, ACTG2 of 1.37, OR2T29 of 0.797, and PGD of 1.3 also made negative contributions to prediction result. Through the accumulation of SHAP values, the waterfall chart provides a visual depiction of the step-by-step construction of an individual patient's prediction, facilitating a more comprehensive interpretation of the model's decision logic. In addition, the SHAP dependence plot provides insight into how individual features influence the model's predictions. Figure 6A and G compares the actual values and SHAP values of the seven selected features. Features with SHAP values below zero correspond to negative contributions to the prediction, that is, they indicate a lower risk of radioresistance. For PLXND1, PGD, GSR, CHAD, PTPRC, and ACTG2, higher expression levels were associated with lower SHAP values, thereby pushing the model's decision toward radiosensitivity. In contrast, lower expression levels of these proteins resulted in higher SHAP values, which shifted the prediction toward radioresistance.

As shown in Fig. 7, the final Random Forest model was integrated into a web-based application. By simply entering the relative expression levels of the seven feature

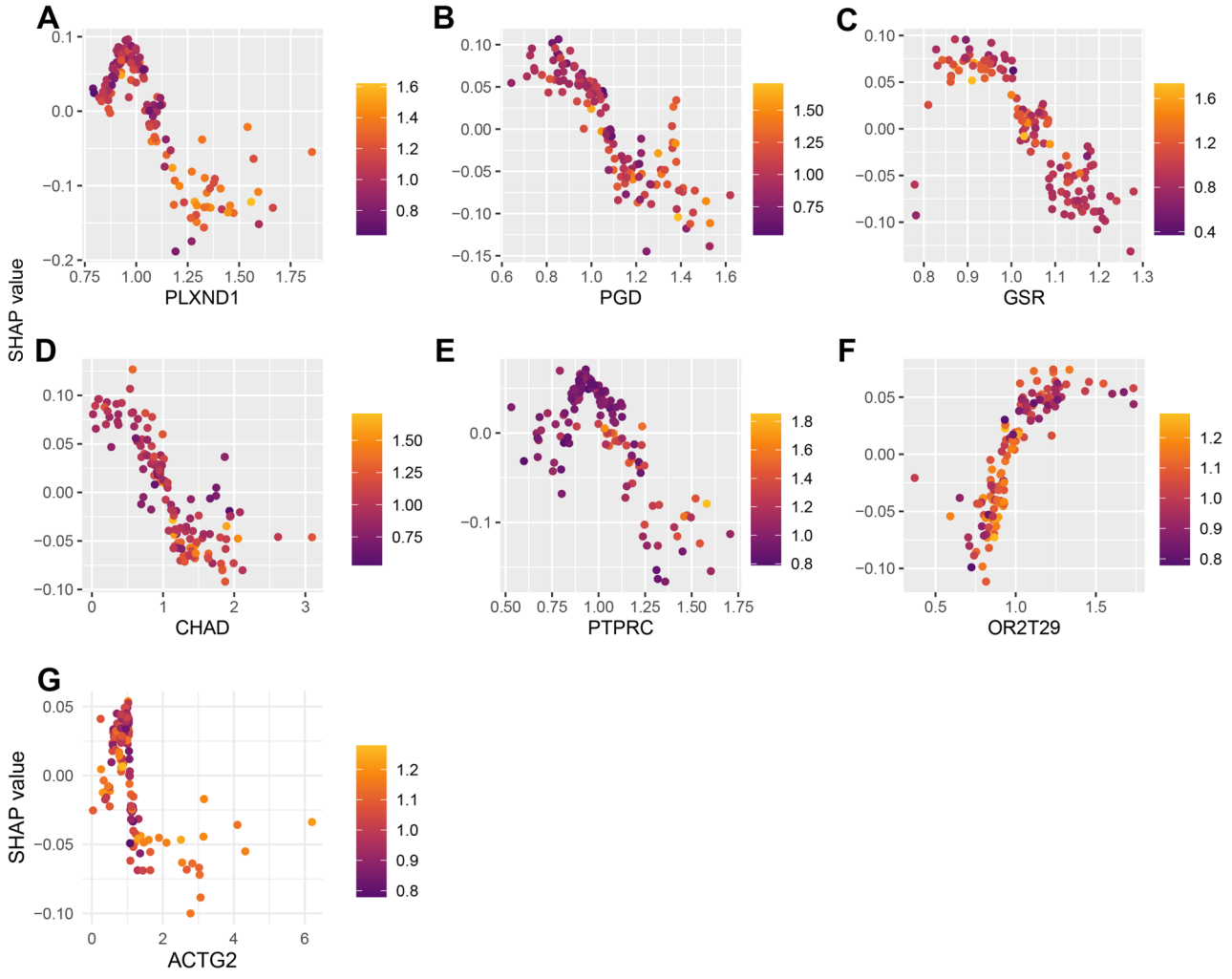

Fig. 6 Local explanation using SHAP dependence plots for individual features. **A**: PLXND1. **B**: GSR. **C**: PGD. **D**: PTPRC. **E**: OR2T29. **F**: ACTG2. **G**: CHAD

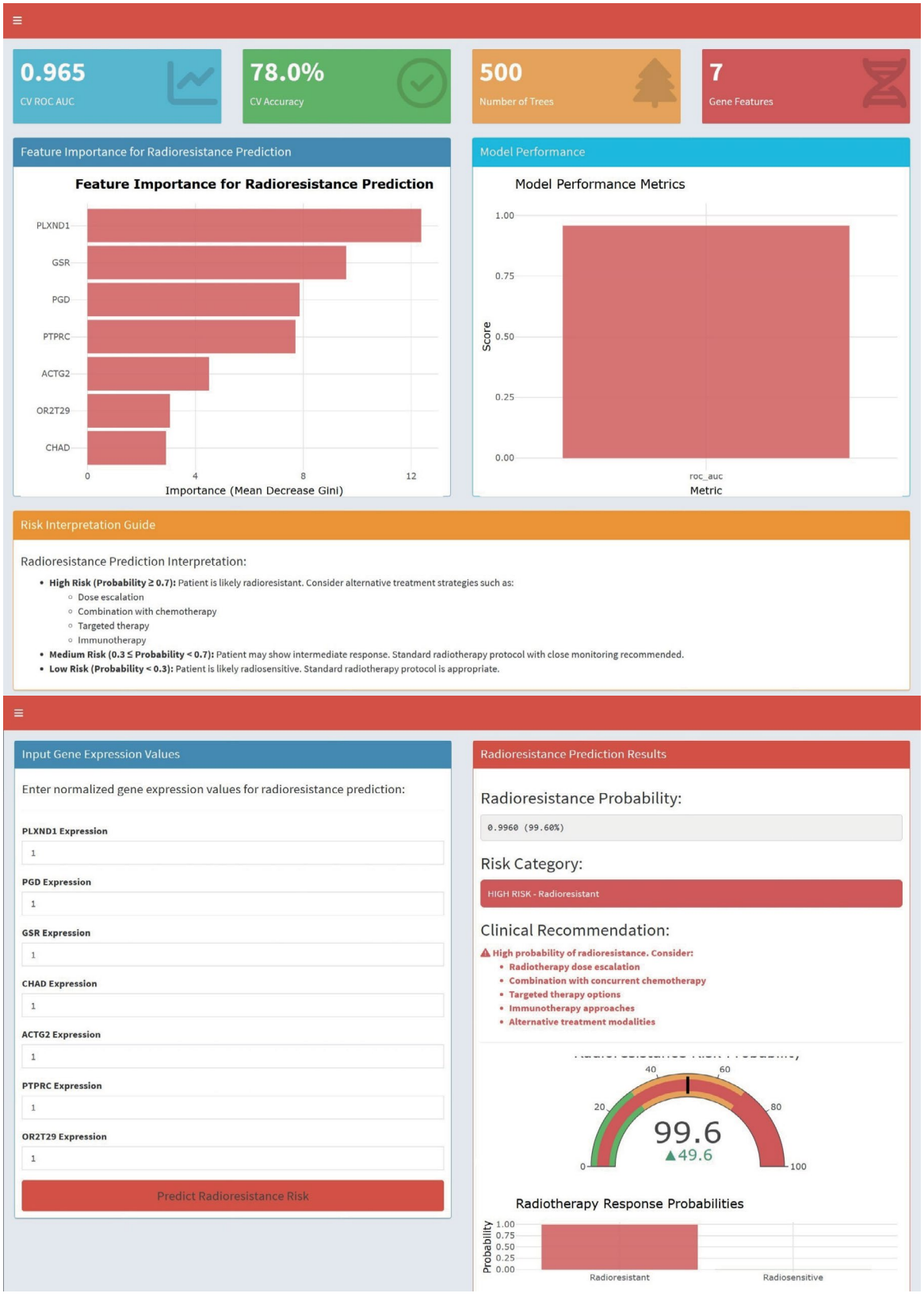

Fig. 7 Web-based interactive calculator for predicting radioresistance in NPC using the random forest model. By entering the relative expression levels of the seven proteins, the tool outputs the predicted probability of radioresistance for an individual patient. The range of seven proteins expression: PLXND1(0.79–1.854), PGD(0.642–1.62), GSR (0.779–1.279), CHAD (0.016–3.093), ACTG2 (0.034–6.202), PTPRC(0.531–1.707), and OR2T29 (0.37–1.735)

proteins, the application automatically calculates the probability of radiotherapy resistance in NPC patients, provides risk stratification, and offers corresponding treatment recommendations. The web application is available at the following link: https://lizzedu.shinyapps.io/NPCShinyAPP/.

Discussion

In the present study, we first analyzed serum proteomic differences between radiosensitive and radioresistant NPC patients and identified 96 differentially expressed proteins which involved multiple biological functions and signaling pathways. Next, we developed twelve machine learning-based predictive models for radiosensitivity in NPC patients, and the random forest model with seven proteins (PLXND1, GSR, PGD, PTPRC, OR2T29, ACTG2, and CHAD) was ultimately selected to construct the final model. This model demonstrated excellent performance in both the training cohort (AUC = 0.963) and independent validation cohort (AUC = 0.975), indicating strong predictive ability and good generalizability. Subsequently, by using SHAP analysis, the model provides a transparent explanation of how individual features influence prediction outcomes. Finally, we integrated final model into a web-based application, enhancing its clinical utility and value. Previous studies have also attempted to predict radiation sensitivity in NPC using machine learning. For example, Li et al. developed a multigene predictive model based on transcriptomic data, which included 18 genes and achieved an AUC of 0.996 in the training set and 0.823 in the validation set, with a combined AUC of 0.932 [21, 22]. Several important differences exist between their study and ours. First, the sample sources were fundamentally different: the previous model relied on transcriptomic data from NPC tissues, whereas our model was constructed using serum proteomics, which reflects more accessible and dynamic biomarkers. Second, the earlier study was limited by a small sample size—34 samples in the training set and 20 in the validation set—which increases the risk of overfitting and may explain the substantial discrepancy between training and validation performance [23]. Third, the two studies differ in their scientific focus. Prior work mainly emphasized identifying differential genes and exploring potential molecular mechanisms but did not assess model interpretability or clinical utility. In contrast, our study not only rigorously evaluated and validated the predictive performance of the model but also conducted an in-depth analysis of its interpretability and underlying predictive mechanisms. Finally, compared with nasopharyngeal carcinoma tissue samples, blood samples are easier to obtain, more suitable for clinical applications, and more amenable to real-time monitoring. Taken together, these strengths underscore that our model offers a more clinically accessible, interpretable, and generalizable approach for predicting radiosensitivity in NPC.

A key advantage of this study lies in its integration of SHAP-based interpretability, which considerably improves model transparency and facilitates potential clinical adoption [24]. By quantifying how each feature contributes to individual predictions, SHAP values enable clinicians to identify not only the most influential proteins but also the rationale behind classifying a patient as radiosensitive or radioresistant. This interpretability is essential for fostering trust in AI-driven clinical tools. Through SHAP analysis, we identified PLXND1, GSR, and PGD as the most influential contributors. The biological relevance of these molecules further supports the credibility of the model.

The function of PLXND1 exhibits marked tumor-type specificity, yet its role consistently converges on the core adaptive mechanisms that enable tumors to cope with microenvironmental stress and therapeutic pressure. In prostate cancer, PLXND1 serves as a central regulatory hub driving acquired resistance. It facilitates transdifferentiation of adenocarcinoma cells toward a neuroendocrine lineage, allowing them to disengage completely from androgen receptor dependence and activate pro-survival signaling through the ErbB3/MET pathway. This transition ultimately results in the emergence of a highly aggressive, treatment-refractory neuroendocrine phenotype [25]. In non–small cell lung cancer, a circular RNA derived from PLXND1 (circ_PLXND1) promotes tumor progression by functioning as a competing endogenous RNA that sequesters miR-1287-5p and consequently upregulates targets such as ERBB3, revealing a novel post-transcriptional oncogenic mechanism [26]. Moreover, PLXND1 is broadly expressed in the vascular endothelium of multiple solid tumors, suggesting that its function extends beyond cancer cells to include remodeling of the tumor microenvironment, particularly tumor angiogenesis [27–29]. Collectively, these findings position PLXND1 as a multifunctional node that integrates intrinsic tumor cell plasticity with microenvironmental interactions. Together with its association with radioresistance observed in our study, PLXND1 provides new insight into the extensive adaptive evolutionary strategies employed by tumors. GSR and PGD are central regulators of redox balance and metabolic reprogramming, both key processes known to shape radiation response [30]. The metabolic reprogramming of tumor cells and the maintenance of redox balance are crucial bases for their resistance to therapeutic methods such as radiotherapy [31]. This process may be closely related to the strong antioxidant defense system of tumor cells. For instance, the key enzyme G6PD (also known as PGD) of the pentose phosphate pathway is responsible for generating a large amount of NADPH, which is a necessary cofactor for the regeneration of reduced glutathione (GSH) by glutathione reductase (GSR) [32]. GSH is the core antioxidant molecule for cells to combat reactive oxygen species (ROS). Studies have shown that simultaneously inhibiting glycolysis and glutathione, thioredoxin

system (including GSR), can weaken the defense of cancer cells against ROS, thus becoming one of the strategies for enhancing radiotherapy sensitivity [33]. Notably, their contribution patterns—where lower expression levels correlate with increased radiosensitivity—align well with previously reported mechanisms related to radiation-induced oxidative stress. At the individual patient level, SHAP waterfall and force plots offer intuitive visual explanations of prediction outputs, enabling clinicians to understand the specific factors driving a radioresistance prediction. This individual interpretability is valuable for shared decision-making and personalized treatment planning. Moreover, integrating the final model into a ShinyAPP substantially enhances its clinical practicality. Clinicians can input quantitative levels of the seven proteins and immediately receive not only the predicted probability of radioresistance but also an explanation of how each biomarker shapes the prediction. Such a tool can be seamlessly incorporated into routine workflows and may inform several aspects of clinical care, including identifying high-risk patients who may benefit from treatment intensification, guiding early therapeutic adjustments, and supporting closer surveillance during radiotherapy.

Despite these advantages, several limitations should be acknowledged. First, although the sample size was sufficient for initial model development, it remains relatively small, and this study was conducted at a single center. These factors may limit the model's applicability to broader populations with diverse ethnic backgrounds, treatment regimens, or EBV characteristics. Multicenter prospective studies are needed to further validate the model's robustness and generalizability. Second, although SELDI-TOF-MS provides high-throughput proteomic profiling, the technique primarily captures low-molecular-weight proteins and may miss high-molecular-weight or low-abundance proteins that could have important biological significance. Third, the biological functions of some identified proteins, such as OR2T29 and CHAD, have not been extensively studied in NPC, and their association with radioresistance requires further experimental validation. Finally, although the model demonstrates strong predictive performance, the absolute risk estimates should be interpreted with caution and future work could incorporate probability calibration techniques, such as Platt scaling or isotonic regression, prior to clinical implementation. It reflects correlation rather than causation, and functional studies are necessary to determine whether these proteins directly mediate radiation response or act as surrogate biomarkers.

Conclusions

In conclusion, this study integrates serum proteomics with machine learning to establish a robust, interpretable, and clinically practical predictive model for radiotherapy sensitivity in NPC. The model demonstrates excellent discriminatory performance, supports clinical decision-making effectively, and is readily deployable via a user-friendly web application. By enabling early identification of patients at high risk of radioresistance, this tool may assist clinicians in tailoring treatment strategies—such as intensifying systemic therapy or adjusting radiotherapy regimens. Future multicenter validations, mechanistic explorations, and refinements in proteomic profiling technology will further enhance the utility and translational potential of this predictive framework. Once validated in larger cohorts, this approach could represent a meaningful step toward personalized radiotherapy in NPC.

Abbreviations

NPC	Nasopharyngeal carcinoma
SELDI-TOF-MS	Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry
FC	Fold change
CI	Confidence interval
GO	Gene ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
ROC	Receiver operating characteristic curve
SHAP	SHapley Additive exPlanations
AUC	Area under the curve
GTV	Gross tumor volume
AFP	Alpha-fetoprotein
EBV	Epstein–Barr virus
FDR	False discovery rate
LR	Logistic Regression
NN	Neural Network
SVM-kernel	Support Vector Machine with Kernel Function
GBM	Gradient Boosting Machine
AdaBoost	Adaptive Boosting
LDA	Linear Discriminant Analysis
PLS	Partial Least Squares Model
LASSO	Least Absolute Shrinkage and Selection Operator
k-NN	k-Nearest Neighbors
RF	Random Forest

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-026-07847-2.

Supplementary Material 1: Table S1 Differentially expressed proteins between radioresistanceand radiosensitivity groups in NPC

Supplementary Material 2: Table S2 Performance of the machine learning models for radiosensitivity

Supplementary Material 3

Supplementary Material 4: Table S3 General characteristics between training set and validation set

Supplementary Material 5

Acknowledgements

None.

Author contributions

ZZL designed this study and directed the research group in all aspects, including planning, execution, and analysis of the study. MJL drafted the manuscript. LX and WZW collected the data. ZZL provided statistical software, performed the data analysis, MJL arranged the Figures and Tables. ZZL and SLF revised the manuscript. All authors have read and approved the final version of the manuscript.

Funding
This study was supported by Hunan Provincial Natural Science Foundation Enterprise Joint Fund of China (No. 2025JJ90289) and Nasopharyngeal Carcinoma Clinical Big Data System Project of Xiangya Hospital, Central South University (No. xyyydsj2).

Data availability
The data underlying this article will be shared on reasonable request to the corresponding author.

Declarations

Ethics approval and consent to participate
This study was approved by the Ethics Committee of Xiangya Hospital, Central South University (No.201404355) and conducted in compliance with the 1964 Helsinki Declaration and its subsequent amendments. All participants were informed of the study procedures and provided written informed consent.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no conflict of interest.

Author details
[1]Department of Oncology, Xiangya Hospital, Central South University, No.87, Xiangya Road, Kaifu District, Changsha, Hunan Province 410008, China
[2]Department of Radiation Oncology, Xiangya Boai Rehabilitation Hospital, Changsha, Hunan 410199, China
[3]National Clinical Research Center for Geriatric Diseasesï Xiangya Hospitalï, Central South University, Changsha, Hunan Province 410008, China

Received: 18 December 2025 / Accepted: 4 February 2026
Published online: 11 February 2026

References

1. Su ZY, Siak PY, Lwin YY, Cheah S. Epidemiology of nasopharyngeal carcinoma: current insights and future outlook. Cancer Metastasis Rev. 2024;43(3):919–39.
2. Lam WKJ, Kang G, Chan CML, Lee VCT, Ma ML, Zhou Q, et al. Fragmentomics profiling and quantification of plasma Epstein-Barr virus DNA enhance prediction of future nasopharyngeal carcinoma. Cancer Cell. 2025;43(4):728–39.
3. Zhang Z, Chen X, Yuan T. Precision radiotherapy for nasopharyngeal carcinoma. Precision Radiation Oncol. 2024;8(1):37–41.
4. Zhu DQ, Su C, Li JJ, Li AW, Luv Y, Fan Q. Update on radiotherapy changes of nasopharyngeal carcinoma tumor microenvironment. World J Oncol. 2023;14(5):350–7.
5. Zhou X, Shao T, Jia H, Hou L, Tang X, Yu C, et al. Current state, challenges, and future perspective of adaptive radiotherapy: A narrative review of nasopharyngeal carcinoma. Oral Oncol. 2024;158:107008.
6. Xu H, Li W, Wang D. The promising role of MiRNAs in radioresistance and chemoresistance of nasopharyngeal carcinoma. Front Oncol. 2024;14:1299249.
7. Siak PY, Heng WS, Teoh SSH, Lwin YY, Cheah S. Precision medicine in nasopharyngeal carcinoma: comprehensive review of past, present, and future prospect. J Transl Med. 2023;21(1):786.
8. Vasudevan HN, Yom SS. Nasopharyngeal carcinoma and its association with Epstein-Barr virus. Hematol Oncol Clin N Am. 2021;35(5):963–71.
9. Alharbi RA. Proteomics approach and techniques in identification of reliable biomarkers for diseases. Saudi J Biol Sci. 2020;27(3):968–74.
10. Barbosa EB, Vidotto A, Polachini GM, Henrique T, Marqui ABTD, Tajara EH. Proteomics: methodologies and applications to the study of human diseases. Rev Assoc Med Bras (1992). 2012;58(3):366–75.
11. Prelaj A, Miskovic V, Zanitti M, Trovo F, Genova C, Viscardi G, et al. Artificial intelligence for predictive biomarker discovery in immuno-oncology: a systematic review. Annals Oncol: Official J Eur Soc Med Oncol. 2024;35(1):29–65.
12. Al-Nafjan A, Aljuhani A, Alshebel A, Alharbi A, Alshehri A. Artificial intelligence in predictive healthcare: A systematic review. J Clin Med. 2025;14(19):6752.
13. Marra A, Morganti S, Pareja F, Campanella G, Bibeau F, Fuchs T, et al. Artificial intelligence entering the pathology arena in oncology: current applications and future perspectives. Ann Oncol. 2025;36(7):712–25.
14. Zhang G, Zhang K, Li C, Li Y, Li Z, Li N, et al. Serum proteomics identify potential biomarkers for nasopharyngeal carcinoma sensitivity to radiotherapy. Bioscience Rep. 2019;39:BSR201900275.
15. Zhao Y, Shen L, Huang X, Jing D, Huang D, Fu J, et al. High expression of Ki-67 acts a poor prognosis indicator in locally advanced nasopharyngeal carcinoma. Biochem Bioph Res Co. 2017;494(1–2):390–6.
16. Nishino M, Jagannathan JP, Van den Ramaiya NH. Revised RECIST guideline version 1.1: what oncologists want to know and what radiologists need to know. AJR Am J Roentgenol. 2010;195(2):281–9.
17. Mohammed FA, Fall EHM, Tune KK, Hammamieh R, Jett M, Muhie S. MultiModalGraphics: an R package for graphical integration of multi-omics datasets. BMC Bioinformatics. 2025;26(1):259.
18. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):R14.
19. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
20. Lundberg S, Lee S. A unified approach to interpreting model predictions. 2017.
21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45.
22. Li K, Liang J, Li N, Fang J, Zhou X, Zhang J, et al. A multi-gene predictive model for the radiation sensitivity of nasopharyngeal carcinoma based on machine learning. Elife. 2025;13:P99849.
23. Aliferis C, Simon G, Overfitting. Underfitting and general model overconfidence and Under-Performance pitfalls and best practices in machine learning and AI. In: Simon GJ, Aliferis C, editors. Artificial intelligence and machine learning in health care and medical sciences: best practices and pitfalls. Cham (CH): Springer; 2024. pp. 477–524.
24. Demir S, Sahin EK. An innovative machine learning approach for slope stability prediction by combining Shap interpretability and stacking ensemble learning. Environ Sci Pollut Res Int. 2025;32(21):12827–43.
25. Wei J, Wang J, Guan W, Li J, Pu T, Corey E, et al. PlexinD1 is a driver and a therapeutic target in advanced prostate cancer. Embo Mol Med. 2025;17(2):336–64.
26. Wu J, Liu C, Yu G. Downregulation of circ_PLXND1 inhibits tumorigenesis of non-small cell lung carcinoma via miR-1287-5p/ERBB3 axis. Thorac Cancer. 2023;14(17):1543–55.
27. Chen W, Guo L, Wang Z, Wang Z, Dai X. PLXND1 from the plexins family is a prognostic factor promoting colon cancer invasion and metastasis. Am J Transl Res. 2025;17(10):8243–64.
28. Hagihara K, Haraguchi N, Nishimura J, Yasueda A, Fujino S, Ogino T, et al. PLXND1/SEMA3E promotes Epithelial-Mesenchymal transition partly via the PI3K/AKT-Signaling pathway and induces heterogenity in colorectal cancer. Ann Surg Oncol. 2022;29(12):7435–45.
29. Li J, Hu K, He D, Zhou L, Wang Z, Tao Y. Prognostic value of PLXND1 and TGF-β1 coexpression and its correlation with immune infiltrates in hepatocellular carcinoma. Front Oncol. 2021;10:604131.
30. Zhang H, Ma J, Hou C, Luo X, Zhu S, Peng Y, et al. A ROS-mediated oxidation-O-GlcNAcylation cascade governs ferroptosis. Nat Cell Biol. 2025;27(8):1288–300.
31. Shi Z, Hu C, Zheng X, Sun C, Li Q. Feedback loop between hypoxia and energy metabolic reprogramming aggravates the radioresistance of cancer cells. Experimental Hematol Oncol. 2024;13(1):55.
32. Park Y, Jeong EM. Glutathione dynamics in the tumor microenvironment: A potential target of cancer stem cells and T cells. Int J Stem Cells. 2024;17(3):270–83.
33. Floberg JM, Schwarz JK. Manipulation of glucose and hydroperoxide metabolism to improve radiation response. Semin Radiat Oncol. 2019;29(1):33–41.

Publisher's note