

Integrating Inflammation-Nutrition Markers and Machine Learning for Prognostic Stratification in Advanced Gastric Cancer: A Multicenter Cohort Study

Guochao Deng

dgc1213@mail.nankai.edu.cn

Chinese PLA General Hospital

Yanrong Wang

Chinese PLA General Hospital

Fangfang Liu

Chinese PLA General Hospital

Xue Han

Chinese Academy of Medical Sciences & Peking Union Medical College

Yuting Pan

Zhejiang Cancer Hospital

Ru Jia

Chinese PLA General Hospital

Nan Zhang

Chinese PLA General Hospital

Yushan Jia

Chinese PLA General Hospital

Quanli Han

Chinese PLA General Hospital

Mengjiao Fan

Chinese PLA General Hospital

Yingtai Chen

Capital Medical University

Guanghai Dai

Chinese PLA General Hospital

Keywords: Machine learning model, Gastric adenocarcinoma, Prognosis, Laboratory parameters, Clinicopathological factors

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10894751/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Integrating Inflammation-Nutrition Markers and Machine Learning for Prognostic Stratification in Advanced Gastric Cancer: A Multicenter Cohort Study

Guochao Deng^{1,†}, Yanrong Wang^{1,†}, Fangfang Liu^{1,†}, Xue Han^{2,†},
Yuting Pan³, Ru Jia¹, Nan Zhang¹, Yushan Jia¹, Quanli Han¹,
Mengjiao Fan¹, Yingtai Chen^{4*}, Guanghai Dai^{1*}

Guochao Deng, Yanrong Wang, Fangfang Liu, Xue Han contributed equally to this work.

¹Senior Department of Oncology, Chinese PLA General Hospital, Beijing, China

²National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China

³ Department of Lymphoma, The Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), Institute of Basic Medicine and Cancer (IBMC), Chinese Academy of Sciences, Hangzhou, Zhejiang 310022, China.

⁴Department of General Surgery, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China.

Corresponding Authors:

Yingtai Chen, Department of General Surgery, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China. E-mail:yingtaichen@126.com

Guang-Hai Dai, Senior Department of Oncology, Chinese PLA General Hospital, Beijing, China, E-mail: daigh301@vip.sina.com

Abstract

Background

Advanced gastric/gastroesophageal junction adenocarcinoma (GEJC) has poor prognosis despite emerging therapies. This study constructed and validated machine learning (ML) survival model combining inflammatory-nutritional biomarkers and

clinicopathological variables to stratify advanced gastric/GEJC patients into distinct prognostic subgroups.

Methods

796 eligible patients from one medical center were split into training (n=560) and internal validation (n=236) cohorts at a 7:3 ratio, with an external validation cohort of 186 patients from two additional hospitals. Univariate Cox and LASSO-Cox regression filtered predictors for overall survival (OS) and progression-free survival (PFS). Seven ML survival models were compared; the optimal model was interpreted via SHAP analysis, and K-means clustering identified prognostic subtypes. Model performance was evaluated by C-index and time-dependent AUC.

Results

Eight core predictors were selected: WBC, LDH, CA724, prognostic nutritional index (PNI), platelet-to-lymphocyte ratio (PLR), surgical history, immunotherapy, and treatment efficacy. XGBoost achieved stable predictive performance for OS and PFS, with average C-indexes of 0.684 and 0.678 respectively. Time-dependent ROC curves showed favorable AUC values for 1/2/3-year OS, and 3/6/12-month PFS across all datasets. K-means clustering identified two subtypes with distinct

survival ($p<0.05$). Subtype 1 (low inflammation, high nutrition; better prognosis), Subtype 2 (high inflammation, low nutrition; worse prognosis). SHAP revealed nonlinear associations and variable interactions, with treatment efficacy as the leading protective factor for both endpoints.

Conclusion

The eight-variable XGBoost model showed robust predictive accuracy and generalizability for OS and PFS in advanced gastric/GEJC. Data-driven subtyping and SHAP visualization facilitate personalized risk stratification and clinical management, while prospective verification remains warranted.

Keywords

Machine learning model, Gastric adenocarcinoma, Prognosis, Laboratory parameters, Clinicopathological factors

Introduction

In recent years, the incidence and mortality of gastric cancer have gradually increased. Gastric cancer is associated with an extremely high mortality rate and poor prognosis, ranking as the fifth most commonly diagnosed cancer and the fourth leading cause of cancer-related death worldwide[1, 2]. Current clinical studies indicate that first-line treatment for patients with advanced gastric cancer is primarily based on

combination chemotherapy with fluorouracil and platinum agents. Depending on HER-2 (Human Epidermal Growth Factor Receptor-2) positivity and PD-L1(Programmed Cell Death Ligand 1) expression status, adding anti-HER2 targeted therapy and/or immunotherapy to chemotherapy could significantly improve patient survival outcomes[3]. Although the survival of patients with advanced gastric cancer has improved to some extent with the incorporation of targeted therapy and immunotherapy, the current 5-year survival rate for gastric cancer is merely 33.53%, highlighting the urgent need for improved prognostic tools and therapeutic strategies[4]. The efficacy of anti-tumor drugs varies among patients with advanced gastric cancer; therefore, assessing the prognosis of patients under different conditions is crucial for both patients and oncologists. Previous studies showed that systemic inflammation and nutritional abnormalities were associated with tumor development and progression[5]. Moreover, as inflammation-nutrition biomarkers reflect systemic inflammatory status and nutritional condition, they have become important prognostic indicators for various malignancies[6, 7].

Machine learning (ML) is increasingly being applied in clinical oncology for tumor diagnosis, prognosis, and treatment

guidance[8, 9]. Compared with traditional prediction models, machine learning demonstrates distinct advantages including automated feature extraction, powerful generalization capability, superior performance in processing complex high-dimensional data, strong adaptability, and the ability to mine potential correlations. These merits render it highly applicable in clinical settings[10]. Shapley Additive Explanations (SHAP) can quantify the contribution of variables to risk and enable visualization, thereby effectively improving the interpretability of prediction models.

Our study hypothesized that a clinical prediction model incorporating multiple parameters—including inflammatory markers, nutritional indicators, and clinicopathological characteristics—will be effectively applicable to survival prediction in advanced gastric cancer[11, 12].

Materials and Methods

Patient population

We retrospectively collected clinicopathological characteristics, hematological biomarkers, and follow-up data from patients with advanced gastroesophageal adenocarcinoma. Eligible patients were collected from three medical centers of the Chinese PLA General Hospital (the First

Medical Center and the Fifth Medical Center) as well as National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College between January 1, 2016 and June 30, 2024. Follow-up was performed every 6 months via telephone or medical record evaluation.

Inclusion criteria: 1) Aged ≥ 18 years; 2) Eastern Cooperative Oncology Group Performance Status (ECOG PS) score of 0–2; 3) Histopathologically confirmed gastric adenocarcinoma, gastroesophageal junction adenocarcinoma, or esophageal adenocarcinoma, with unresectable locally advanced lesions or distant metastases; or disease progression/recurrence occurring at least 6 months after radical surgery; 4) Receipt of at least 2 cycles of first-line therapy with PD-1 inhibitor monotherapy, PD-1 inhibitor combined with chemotherapy, or targeted therapy; 6) Underwent at least one tumor response assessment during first-line treatment.

Clinicopathological Variables Collection

The collected data were categorized into three categories: general information, baseline clinical characteristics, and hematological biomarkers. General clinical information mainly included gender and age. Baseline clinical characteristics

comprised ECOG PS score, tumor location, histological differentiation grade, Lauren classification, diagnosis time, HER2 expression status, microsatellite instability (MSI) status, programmed death-ligand 1 combined positive score (PD-L1 CPS), liver metastasis, lung metastasis, peritoneal metastasis, lymph node metastasis, number of metastatic sites, chemotherapy regimen, immunotherapy administration in first-line treatment, surgical history, best overall response to first-line treatment, progression-free survival (PFS), time of death, and the date of the last outpatient visit or discharge in our hospital.

Hematological biomarkers included baseline hemoglobin (HB), white blood cell count (WBC), neutrophil count (N), lymphocyte count (LN), platelet count (PLT), albumin (ALB), lactate dehydrogenase (LDH), carcinoembryonic antigen (CEA), alpha-fetoprotein (AFP), carbohydrate antigen 199 (CA199), carbohydrate antigen 724 (CA724), platelet-to-albumin ratio (PAR), neutrophil-to-platelet ratio (NPR), prognostic nutritional index (PNI), hemoglobin-albumin-lymphocyte-platelet (HALP) score, neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio (PLR), and lung immune prognostic index (LIPI). These markers were calculated as follows: PAR= platelet

count ($\times 10^9/L$) / serum albumin (g/dL). NPR= neutrophil count($\times 10^9/L$) / platelet count ($\times 10^9/L$). NLR= neutrophil count($\times 10^9/L$)/ lymphocyte count($\times 10^9/L$). PLR= platelet count ($\times 10^9/L$) / lymphocyte count($\times 10^9/L$). PNI = $10 \times$ serum albumin (g/dL) + $0.005 \times$ total lymphocyte count (/mm³). The HALP score was calculated using the formula: HALP = hemoglobin (g/L) $\times$ albumin (g/dL) $\times$ lymphocyte count ($\times 10^9/L$) / platelet count ($\times 10^9/L$). LIPI was constructed based on baseline NLR and LDH, with the cutoff value of NLR set at 3 and the cutoff value of LDH set at 250 U/L (the upper limit of the normal reference range). Patients were stratified into three LIPI subgroups: the poor-risk group (NLR > 3 and LDH > 250 U/L), the intermediate-risk group (NLR > 3 or LDH > 250 U/L), and the good-risk group (NLR $\leq$ 3 and LDH $\leq$ 250 U/L). Efficacy was evaluated in accordance with the Response Evaluation Criteria in Solid Tumors (RECIST) 1.1, and patients were categorized into progressive disease (PD) or non-progressive disease (non-PD) subgroups according to objective tumor responses. PFS was defined as the interval from the start of first-line chemotherapy to the first radiologically confirmed disease progression, all-cause mortality, or the final follow-up assessment. OS was defined as the interval from the initiation of

first-line chemotherapy to death from any cause or the final follow-up assessment date.

Feature screening and cluster analysis

This study applied a dual-endpoint-driven variable screening strategy to reduce bias caused by using a single endpoint, with the intersection of results from PFS and OS screenings serving as the final feature set. We performed univariate survival analysis for both PFS and OS simultaneously on all variables, and also conducted dual-outcome feature selection using Least Absolute Shrinkage and Selection Operator-Cox (Lasso Cox) regression for PFS and OS, respectively. Candidate variables were identified by taking the intersection of features consistently selected across both outcomes, with the intersection visualized using a Venn diagram.

For patient stratification, unsupervised K-means clustering was conducted on the entire cohort (n=796) based on the intersection prognostic features screened by LASSO Cox regression and univariate Cox analysis. The optimal number of clusters was determined comprehensively using the elbow method combined with the silhouette coefficient to balance within-cluster variance and inter-cluster separation. An independent external cohort was applied to further confirm the

stability and reproducibility of the subtypes. To ensure clinical relevance, we validated the clustering results via Kaplan-Meier survival analysis, which revealed significant differences in PFS and OS among the identified patient subgroups. Radar plots were constructed to visualize the top 5 variables—selected via ANOVA F-statistics for maximum inter-cluster discrimination—and highlight the distinct feature profiles of each subgroup, with principal component analysis (PCA) further validating clear segregation among the clusters. Combined dual-outcome survival curves were also created to demonstrate the clinical utility of the proposed stratification approach.

Model development and interpretation

Based on the aforementioned feature preprocessing techniques involving low-variance feature elimination and highly correlated feature removal, this study developed seven survival machine learning models: Extreme Gradient Boosting (XGBoost), Lasso Cox Regression (Lasso), CoxBoost, Gradient Boosting Machine (GBM), Supervised Principal Components (SuperPC), Cox Partial Least Squares Regression (plsRCox), and Survival Support Vector Machine (survival-SVM). Lasso Cox regression was chosen as a baseline model due to its simplicity, interpretability, and inherent feature selection

capability for prognostic modeling. XGBoost was selected for its superior performance in processing survival data, robust ability to capture complex non-linear correlations between features and overall survival, and enhanced generalization with hyperparameter optimization to mitigate overfitting. CoxBoost and Gradient Boosting Machine were included for their advanced ensemble learning frameworks tailored for survival analysis, which boost prediction accuracy and reduce model variance. Supervised Principal Components and Cox Partial Least Squares Regression were incorporated to effectively extract key prognostic information from high-dimensional data and address potential multicollinearity among features. Survival Support Vector Machine was added to explore the potential advantages of kernel-based methods in capturing nonlinear patterns in prognostic feature datasets.

These models were systematically evaluated in the independent training and test sets using a consistent set of evaluation metrics. Model performance was primarily assessed based on discrimination (C-index and ROC curves). Ultimately, the best performing model was selected based on the integrated evaluation of predictive performance (average C-index). A

detailed performance analysis and risk score prediction were conducted for the selected optimal model.

The SHAP is an attribution method based on game theory that interprets predictions derived from complex machine learning models, including "black box" models. In this study, this method was used to quantify the contribution of individual clinical features to survival outcomes (OS and PFS) in patients with advanced gastric/GEJC. In the SHAP framework, positive values indicate clinical features that increase the risk of adverse survival outcomes in advanced gastric/GEJC patients, whereas negative values represent protective factors that reduce this survival risk. We adopted the XGBoost model—renowned for its robust performance in predicting cancer patients' survival outcomes—as the base model. By implementing the SHAP method via Kernelshap and Shapviz packages, we systematically ranked feature importance in the survival prediction model, with the mean absolute SHAP values used to quantify the overall contribution of each clinical feature. Beyond the global ranking of feature importance, we conducted individualized interpretations by generating SHAP waterfall plots and force plots for representative patient samples, which

elucidated the specific contribution of each clinical feature to the individual survival risk score of these patients.

Statistical analysis was performed based on R software version 4.3.3 (<http://www.r-project.org>).

Results

Patients Characteristics

A total of 796 patients with advanced gastric/GEJC were enrolled and randomly divided into training and internal validation datasets at a 7:3 ratio. In addition, 186 patients from other healthcare centers served as external validation (Figure1). Baseline characteristics were compared between the sets, as shown in Table1. In the training dataset, the median OS and median PFS were 10.8 months and 6.15 months, respectively. In the training set, OS status was significantly correlated with Lauren classification, mismatch repair (MMR) status, liver metastasis, peritoneal metastasis, immunotherapy, treatment efficacy, and levels of CA199 and HALP ($p < 0.05$), whereas PFS status was associated with Her-2 status, Lauren classification, peritoneal metastasis, immunotherapy, treatment efficacy, CA199, and LN levels (Table S1). In the internal validation set, the median OS and the median PFS were 11.1 months and 6.30 months, respectively. Variables significantly

associated with OS status included peritoneal metastasis, immunotherapy, chemotherapy regimen and treatment efficacy. Furthermore, gender, age, MMR, PD-L1 expression, immunotherapy, chemotherapy regimen, and treatment efficacy were identified as significant factors related to PFS status (Table S2). Both datasets collectively indicate that the line of immunotherapy, treatment efficacy, and specific metastatic status (such as peritoneal metastasis) are significantly associated with both OS and PFS in both the training and test sets.

Variable selection

Before initiating variable screening, we performed Spearman correlation analysis on all variables in the dataset. The results are intuitively presented as a heatmap in the Figure S1A-C, where the color intensity reflects the strength of the correlation among variables. The results showed that variables with a correlation coefficient greater than 0.7 included PLT and PAR ($r=0.957$), ALB and PNI ($r=0.824$), N and NLR ($r=0.770$), NPR and HALP ($r=0.733$). In external dataset, A strong positive correlation was identified for PLT with PAR ($r = 0.932$) and for ALB with PNI ($r = 0.914$). Furthermore, moderate to strong positive correlations were found for WBC with N ($r = 0.864$), N with NLR ($r = 0.792$), WBC with NLR ($r = 0.767$), and N with

NPR ($r = 0.720$). Finally, NLR, NPR, PLT, ALB, and N were excluded from subsequent analyses.

There were 28 variables included in univariate survival analysis and Lasso Cox analysis. In the univariate Cox survival analysis for PFS, 13 variables showed statistical significance, while in the univariate Cox survival analysis for OS, 15 variables were statistically significant. Ten variables overlapped between the PFS and OS survival analyses. In the Lasso regression analysis, both the PFS-Lasso and OS-Lasso models selected 25 variables each. After taking the intersection of the results from these two approaches, a final set of eight variables (WBC, LDH, PLR, PNI, CA724, surgery, first-line immunotherapy, treatment efficacy) was screened for subsequent analysis (Figure 2).

Model Construction and evaluation

We incorporated the selected eight variables into seven machine learning models (CoxBoost, GBM, Lasso, plsRCox, SuperPC, survival-SVM, XGBoost). The performance of each model was shown in Table 2. Among the eight models, XGBoost demonstrated the best performance in survival prediction for both PFS and OS. XGBoost achieved C-index values of 0.721 on the training set, 0.688 on the internal

set, 0.642 on the external set and an average of 0.684 for OS prediction. For PFS prediction, the corresponding C-index values were 0.701 on the training set, 0.667 on the internal set, 0.667 on the external set, with an average of 0.678. The differences between training and validation sets were both less than 0.05 for both OS and PFS groups, indicating a moderate decline in the model's generalization ability on unseen test data. To evaluate model performance, time-dependent ROC curves were plotted using XGBoost. In the training set, the AUC values for OS at 1, 2, and 3 years were 0.733, 0.740, and 0.780, respectively. For PFS, the corresponding AUC values at 3, 6, and 12 months were 0.873, 0.726, and 0.695. In the internal set, the AUC values for OS at 1, 2, and 3 years were 0.716, 0.714, and 0.701, respectively. Similarly, for PFS at 3, 6, and 12 months, they were 0.941, 0.665, and 0.657. In the external set, the AUC values for OS at 1, 2, and 3 years were 0.733, 0.678, and 0.659, respectively, while for PFS at 3, 6, and 12 months, they were 0.724, 0.735, and 0.786. (Figure 3A-F). Survival outcomes differed significantly between high- and low-risk groups in both the training and validation cohorts (Figure 4A-F).

Heterogeneous characteristics in tumor microenvironment clusters

We used the remaining eight variables for the K-means clustering algorithm.

According to the elbow method (based on within-cluster sum of squares), the optimal number of clusters is $k = 2$ (Figure 5A). The mean values of each standardized feature across the clusters were calculated, and features with high discriminatory power (most significant intergroup differences) were selected by calculating the F-statistic via one-way ANOVA, choosing the top 5 features by F-value. Based on the principal component analysis results, the patients could be classified into two subtypes, and different outcome groups exhibit distinct clustering patterns in the dimensionality-reduced space. The top five features were WBC, PNI, CA724, efficacy and immunotherapy. Radar charts were used to display the characteristic expression patterns of different subtypes, clarifying the differences in their biological features (Figure 5B). Based on the principal component analysis (PCA), patients were classified into two subtypes, which showed distinct clustering patterns in the principal component space (Figure 5C-D). Subtype 1 exhibited lower levels of WBC and CA724, comparable nutritional status (PNI), and slightly better treatment efficacy compared to Subtype 2. In contrast, Subtype 2 was

characterized by significantly elevated WBC and CA724, alongside a high immunotherapy-related signature, but had significantly worse overall survival. Significant differences among the two subtypes were observed in both progression-free survival (PFS) and overall survival (OS) ($p < 0.0001$) (Figure 5E-H). Trends for PFS and OS were entirely consistent: Subtype 1 had the better survival prognosis than Subtype 2.

Feature Contribution Analysis via SHAP Values

In this study, an optimized XGBoost machine learning model was employed, and the SHAP method was applied to rank the importance of features. The results demonstrated the contribution of each feature variable to predicting survival outcomes in advanced gastric/GEJC patients. The SHAP summary plot visually illustrates the direction and magnitude of feature contributions, where positive SHAP values indicate an increased risk of adverse survival outcomes, while negative values represent a protective effect. The SHAP summary plot indicates that treatment efficacy, WBC, CA724, and PNI are the most influential features in predicting overall survival (OS) in patients with advanced gastric/GEJC (Figure 6A-B). We further applied the XGBoost algorithm to perform a specific SHAP value decomposition for the overall survival (OS) risk of a

representative case. In one patient, the combined effects of normal LDH (164 U/L), normal WBC ($7.1 \times 10^9/L$), and Efficacy=1(effective treatment) reduced the prediction score from the baseline value ($E[f(x)] = 0.623$) to a final value of 0.524. In contrast, risk factors including a high PNI (43) and no surgery (Surgery=0) slightly increased the score. Overall, the net effect of these features resulted in a lower predicted mortality risk for this patient compared to the average level (Figure 6C-D). Results from the SHAP dependence plot revealed that all 8 incorporated core features exhibited significant nonlinear correlations and threshold effects in their predictive contributions to OS. Among these features, elevated levels of tumor markers CA724 and LDH and inflammatory indicators (WBC, PLR) were associated with positive SHAP values, which significantly increased patients' mortality risk and thus served as core drivers of poor prognosis. In contrast, the prognostic nutritional index (PNI) showed a shift from positive to negative SHAP values with increasing numerical values, presenting a marked protective effect; treatment efficacy, surgical intervention, and immunotherapy exerted negative SHAP values within appropriate value ranges, which could reduce patients' survival risk. Meanwhile, the color gradient uncovered

significant interactive effects between inflammatory indicators, the immune marker PD-L1, and core prognostic features: a hyper-inflammatory state synergistically amplified the adverse prognostic effects of tumor markers, whereas treatment efficacy could partially offset the risk increment induced by high-risk features. These findings suggest that systemic inflammation, tumor burden, nutritional and immune status, and therapeutic factors jointly form a regulatory network for OS prognosis, with complex combined regulatory effects among these factors (Figure 6E).

In predicting PFS in advanced gastric/GEJC patients, the most impactful features were treatment efficacy, PLR, CA724, and WBC (Figure 6F-G). In one patient, positive contributions to progression risk came from elevated CA724 (75.8 ng/mL), WBC ($3.9 \times 10^9/L$) and the absence of surgical intervention (Surgery = 0). These risk-promoting effects were substantially offset by prominent negative contributions from effective treatment (Efficacy = 1), normal LDH (145 U/L), and first-line immunotherapy (Immunotherapy = 1), PNI (35), and PLR (198). Collectively, these features drove the model's prediction score from the baseline expected value ($E[f(x)] = 0.721$) down to a final value of $f(x) = 0.586$, indicating a reduced risk of

progression for this patient. (Figure 6H-I). The SHAP dependence plot for PFS demonstrated that all 8 core features exhibited significant nonlinear trends and threshold effects in predicting disease progression. Treatment efficacy was identified as the most critical protective factor, showing remarkably negative SHAP values at high levels and substantially reducing the risk of disease progression. In contrast, elevated levels of tumor markers (CA724, LDH) and systemic inflammatory indices (PLR, WBC) were accompanied by increased positive SHAP values, serving as key risk factors driving disease progression and poor prognosis. The prognostic nutritional index (PNI) shifted from positive to negative SHAP values with increasing values, presenting a distinct protective effect, while surgery and immunotherapy showed differential prognostic contributions with obvious effect inflection points across various ranges. The color gradient further uncovered significant interactions between treatment efficacy and inflammatory or tumor markers, where a high efficacy status markedly attenuated the adverse prognostic effects of high-risk features. These findings indicate that tumor burden, systemic inflammation, nutritional-immune status, and therapeutic response jointly form a core regulatory network for PFS, with

multidimensional features synergistically modulating the risk of disease progression through nonlinear and interactive effects (Figure 6J).

Discussion

Gastric cancer remains a major global health burden with high mortality and poor long-term prognosis, despite advances in targeted therapy and immunotherapy in recent years[13, 14]. This study systematically explored the prognostic value of inflammation, nutritional indicators, and clinicopathological characteristics in advanced advanced gastric/GEJC based on multicenter real-world data, and constructed a survival prediction model using machine learning algorithms. This study sought to fully mine the predictive potential of routine, low-cost clinical and hematological indicators, and translate these findings into a practical tool to support risk stratification and personalized treatment guidance.

Our main findings are summarized as follows: First, using a dual-endpoint-driven feature selection strategy, we identified 8 core features significantly associated with both OS and PFS. Second, among seven machine learning models, the XGBoost algorithm exhibited the best predictive performance, with favorable discriminative ability and generalization capacity.

Third, K-means clustering analysis based on the screened 10 non-therapeutic variables successfully stratified patients into two subtypes with significant survival differences, revealing distinct tumor immune microenvironment characteristics. Finally, the XGBoost model was interpreted using the SHAP method to quantify the contribution of each feature and uncover their complex nonlinear relationships and interactive effects.

Consistent with previous studies[15, 16], our results confirmed that systemic inflammation and nutritional abnormalities are closely associated with disease progression and prognosis in gastric cancer. Additionally, various inflammatory markers have significant prognostic value in gastric cancer[17-20].Through dual-endpoint-driven variable screening (integrating PFS and OS), we identified 8 core prognostic features, including WBC, LDH, PLR, PNI, CA724, surgical history, immunotherapy, and treatment efficacy. PNI not only can reflect the chronic inflammation, immune status, and nutritional condition of cancer patients but is also associated with the prognosis of various diseases[21-24]. In our research, PNI also showed a significant protective effect in both OS and PFS prediction: with the increase in PNI values, its SHAP value shifted from positive to negative, indicating that better

nutritional status can reduce the risk of disease progression and death. This finding aligns with other studies, which conclude that nutritional deficiency and systemic inflammation may synergistically promote tumor progression by suppressing antitumor immunity and enhancing tumor cell invasiveness[5, 25, 26]. Furthermore, elevated levels of inflammatory indicators (WBC, PLR) were identified as high-risk factors for poor prognosis, which may be related to the fact that persistent systemic inflammation can induce tumor angiogenesis, epithelial-mesenchymal transition (EMT), and immune escape[27]. Previous studies have shown that CA724 has certain diagnostic value for gastric cancer, with high specificity but low sensitivity[28]. It is typically combined with markers such as CA199 and CEA to improve the detection rate of gastric cancer[29]. Tong et al. found that CA724 can also serve as a prognostic indicator for gastric cancer[30]. In our study, CA724 also emerged as one of the most influential features in both OS and PFS, further confirming the prognostic value of CA724 in gastric cancer. Based on K-means clustering analysis, we utilized core indicators (WBC, PNI, CA724, treatment efficacy and immunotherapy) to classify patients with advanced gastric cancer into two subtypes with significant survival differences.

Subtype 1 (low inflammation, high nutrition) had the better prognosis and therapeutic outcome. Subtype 2 (high inflammation, low nutrition) had the worse prognosis validating the regulatory role of the "inflammation-nutrition-prognosis" axis. The clustering results were verified through distinct radar chart profiles and principal component analysis, demonstrating high stability and clinical discriminative ability. This classification aids in the early identification of high-risk patients for intensified treatment strategies, while also preventing overtreatment in low-risk patients.

Regarding model construction, we compared the performance of seven machine learning models and identified that the XGBoost model exhibited the optimal overall performance, with mean C-indexes of 0.684 and 0.678 for OS and PFS prediction, respectively. To reduce overfitting of the data results, we incorporated specific anti-overfitting code during data analysis and stipulated that a difference greater than 0.05 in the C-index between the training and validation sets indicates model overfitting. This is particularly important in the context of high-dimensional clinical data. Time-dependent ROC curve analysis further confirmed the reliable predictive performance of the XGBoost model. In the training set, the AUC

values for OS at 1, 2, and 3 years were 0.733, 0.740, and 0.780, respectively, while those for PFS at 3, 6, and 12 months were 0.873, 0.726, and 0.695. Even in the independent validation set, the model still maintained good predictive accuracy, indicating that it has potential clinical application value for risk stratification of advanced gastric cancer patients. Although our model does not achieve the same level of performance as multi-omics-based predictive models, it demonstrates clear clinical utility by leveraging only readily available, routine clinical and hematological parameters. Through in-depth mining of these basic data, the model still provides valuable evidence for clinical risk stratification and treatment decision-making.

SHAP analysis can reflect the regulatory mechanisms and interactions of key features on survival outcomes, effectively enhancing the interpretability of the XGBoost model. Our results indicated that all eight core features exhibited significant nonlinear correlations and threshold effects in prognostic prediction. For example, treatment efficacy was identified as the most critical protective factor for PFS, with its high level (effective treatment) showing a significantly negative SHAP value, which could substantially counteract the adverse effects of high-risk features such as elevated CA724. This finding

highlights the importance of evaluating treatment response during first-line therapy and promptly adjusting treatment strategies for patients with poor response, which aligns with the concept of personalized cancer treatment.

This study possesses several strengths: First, we employed a dual-endpoint-driven variable selection method (considering both progression-free survival and overall survival), which effectively avoided the bias associated with relying on a single endpoint, thereby enhancing the reliability of the core prognostic indicators. Second, we systematically compared the performance of seven machine learning models and utilized SHAP analysis to improve the interpretability of the optimal model. Third, we integrated cluster analysis with survival analysis; this stratification system fully considers the heterogeneity among gastric cancer patients, providing a more precise basis for personalized treatment. Fourth, the study pooled multicenter samples, enhancing the model's generalizability and clinical applicability.

However, this study also has some limitations that warrant consideration. First, as a retrospective study, it may be subject to selection bias (e.g., missing data for certain clinical features or patients lost to follow-up). Second, the study population was

limited to patients with advanced gastroesophageal adenocarcinoma; therefore, the results may not be applicable to patients with early-stage gastric cancer, among others. Finally, the current conclusions are inferences based on SHAP analysis and existing research, and thus require further validation through prospective studies.

In conclusion, this study successfully constructed an XGBoost-based survival prediction model for advanced gastric cancer using eight core prognostic features, which showed good predictive performance and generalization ability. SHAP analysis revealed the nonlinear correlations and interactive effects of core features in prognostic regulation, and K-means clustering identified two subtypes of patients with distinct survival outcomes. These findings confirm that the combination of inflammatory markers, nutritional indicators, and clinicopathological characteristics can effectively improve the accuracy of prognostic evaluation in patients with advanced gastric cancer. And provide a novel theoretical basis and practical tool for personalized risk stratification and treatment decision-making.

In the future, we will expand the sample size, incorporate more promising prognostic factors (including tumor mutational

burden [TMB] and circulating tumor DNA [ctDNA]) and prospective multicenter validation to further optimize our model.

Data availability

The data generated in this study are available upon request from the corresponding author.

Authors' Contributions

Guochao Deng, Yanrong Wang, Fangfang Liu, Xue Han contributed equally as co-first authors. Guochao Deng designed this study, analyzed data, visualized data, drafted and revised this manuscript. Yanrong Wang and Xue Han provides primary data support. Fangfang Liu participated jointly in the main data analysis procedures and provided technical support. Yuting Pan, Nan Zhang, Yushan Jia, Mengjiao Fan, Quanli Han and Ru Jia participated in proofreading data and manuscript. Yingtai Chen and G.H Dai are responsible for manuscript quality control and providing amendments. All authors read and approved the final manuscript.

Acknowledgements

Grateful to Yanrong Wang and Xue Han for the efforts in data preparation and collection at the early stage.

Ethics approval and consent to participate

The study was approved by the institutional ethics committees of the Chinese People's Liberation Army General Hospital (approval number: S2021-642-01). This study does not involve animal experiments. All treatments were performed in accordance with institutional guidelines and regulations. Clinical data retrieved electronically from the medical records of PLA General Hospital Registry. Our study obtained informed consent from the study participants, in accordance with the Declaration of Helsinki.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Funding

No Funding.

References

1. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2020**. *CA Cancer J Clin* 2020, **70**(1):7-30.
2. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A: **Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries**. *CA Cancer J Clin* 2024, **74**(3):229-263.
3. Patel AK, Sethi NS, Park H: **Gastric Cancer: A Review**. *JAMA* 2026, **335**(5):439-450.
4. Chen Y, Jia K, Xie Y, Yuan J, Liu D, Jiang L, Peng H, Zhong J, Li J, Zhang X *et al*: **The current landscape of gastric cancer and gastroesophageal junction cancer diagnosis and treatment in China: a comprehensive nationwide cohort analysis**. *J Hematol Oncol* 2025, **18**(1):42.
5. Zhu H, Shen D, Gan X, Sun D: **An Interpretable Machine Learning Model Based on Inflammatory-Nutritional Biomarkers for Predicting Metachronous Liver Metastases After Colorectal Cancer Surgery**. *Biomedicines* 2025, **13**(7).
6. Zhou M, Gu Q, Zhou M, Yang S, Liu Y, Guan B, Xie B, Han A, Xiang J, Yan D: **Extensive study on**

- the associations of 12 composite inflammatory indices with colorectal cancer risk and mortality a cross-sectional analysis of NHANES 2001-2020.** *Int J Surg* 2025.
7. Chen J, Jin L, Luo R, Zhang X, Chen Y, Han Z, Liu T: **Predictive value of preoperative systemic immune-inflammation index and prognostic nutrition index in patients with epithelial ovarian cancer.** *J Ovarian Res* 2025, **18**(1):45.
 8. Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J: **From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment.** *Cell* 2023, **186**(8):1772-1791.
 9. Bektas M, Burchell GL, Bonjer HJ, van der Peet DL: **Machine learning applications in upper gastrointestinal cancer surgery: a systematic review.** *Surg Endosc* 2023, **37**(1):75-89.
 10. Greener JG, Kandathil SM, Moffat L, Jones DT: **A guide to machine learning for biologists.** *Nat Rev Mol Cell Biol* 2022, **23**(1):40-55.
 11. El Arab RA, Almoosa Z, Alkhunaizi M, Abuadas FH, Somerville J: **Artificial intelligence in hospital infection prevention: an integrative review.** *Front Public Health* 2025, **13**:1547450.
 12. Azodi CB, Tang J, Shiu SH: **Opening the Black Box: Interpretable Machine Learning for Geneticists.** *Trends Genet* 2020, **36**(6):442-455.
 13. Kanaji S, Urakawa N, Harada H, Ikeda T, Koterazawa Y, Aoki T, Sawada R, Otowa Y, Goto H, Hasegawa H *et al*: **Multidisciplinary treatment for advanced gastric cancer.** *Int J Clin Oncol* 2025, **30**(7):1268-1275.
 14. Sato S, Nakayama I, Shitara K: **Pharmacotherapy for previously treated gastric cancer patients: current options and future developments.** *Expert Rev Clin Pharmacol* 2025, **18**(8):607-623.
 15. Correa P, Piazuelo MB: **The gastric precancerous cascade.** *J Dig Dis* 2012, **13**(1):2-9.
 16. Jing Y, Ren M, Li X, Sun X, Xiao Y, Xue J, Liu Z: **The Effect of Systemic Immune-Inflammatory Index (SII) and Prognostic Nutritional Index (PNI) in Early Gastric Cancer.** *J Inflamm Res* 2024, **17**:10273-10287.
 17. Miyamoto R, Inagawa S, Sano N, Tadano S, Adachi S, Yamamoto M: **The neutrophil-to-lymphocyte ratio (NLR) predicts short-term and long-term outcomes in gastric cancer patients.** *Eur J Surg Oncol* 2018, **44**(5):607-612.
 18. Zhang Y, Lin S, Yang X, Wang R, Luo L: **Prognostic value of pretreatment systemic immune-inflammation index in patients with gastrointestinal cancers.** *J Cell Physiol* 2019, **234**(5):5555-5563.
 19. Migita K, Takayama T, Saeki K, Matsumoto S, Wakatsuki K, Enomoto K, Tanaka T, Ito M, Kurumatani N, Nakajima Y: **The prognostic nutritional index predicts long-term outcomes of gastric cancer patients independent of tumor stage.** *Ann Surg Oncol* 2013, **20**(8):2647-2654.
 20. Gou M, Zhang Y, Qu T, Jia R, Wang Z, Dai G: **Prognostic value of the Lung Immune Prognostic Index for HER-2-negative metastatic gastric cancer.** *Biomark Med* 2023, **17**(17):711-721.
 21. Mohri T, Mohri Y, Shigemori T, Takeuchi K, Itoh Y, Kato T: **Impact of prognostic nutritional index on long-term outcomes in patients with breast cancer.** *World J Surg Oncol* 2016, **14**(1):170.
 22. Hu Y, Shen J, Liu R, Feng Z, Zhang C, Ling L, Chen L: **Prognostic value of pretreatment prognostic nutritional index in non-small cell lung cancer: A systematic review and meta-analysis.** *Int J Biol Markers* 2018, **33**(4):372-378.
 23. Luo Z, Zhou L, Balde AI, Li Z, He L, ZhenWei C, Zou Z, Huang S, Han S, Wei Zhou M *et al*:

- Prognostic impact of preoperative prognostic nutritional index in resected advanced gastric cancer: A multicenter propensity score analysis.** *Eur J Surg Oncol* 2019, **45**(3):425-431.
24. Peng P, Chen L, Shen Q, Xu Z, Ding X: **Prognostic Nutritional Index (PNI) and Controlling Nutritional Status (CONUT) score for predicting outcomes of breast cancer: A systematic review and meta-analysis.** *Pak J Med Sci* 2023, **39**(5):1535-1541.
 25. Migita K, Matsumoto S, Wakatsuki K, Kunishige T, Nakade H, Miyao S, Sho M: **Effect of Oral Nutritional Supplementation on the Prognostic Nutritional Index in Gastric Cancer Patients.** *Nutr Cancer* 2021, **73**(11-12):2420-2427.
 26. Thurnham DI: **Interactions between nutrition and immune function: using inflammation biomarkers to interpret micronutrient status.** *Proc Nutr Soc* 2014, **73**(1):1-8.
 27. Li S, Cong X, Gao H, Lan X, Li Z, Wang W, Song S, Wang Y, Li C, Zhang H *et al*: **Correction to: Tumor-associated neutrophils induce EMT by IL-17a to promote migration and invasion in gastric cancer cells.** *J Exp Clin Cancer Res* 2019, **38**(1):177.
 28. Ning S, Wei W, Li J, Hou B, Zhong J, Xie Y, Liu H, Mo X, Chen J, Zhang L: **Clinical significance and diagnostic capacity of serum TK1, CEA, CA 19-9 and CA 72-4 levels in gastric and colorectal cancer patients.** *J Cancer* 2018, **9**(3):494-501.
 29. Yang AP, Liu J, Lei HY, Zhang QW, Zhao L, Yang GH: **CA72-4 combined with CEA, CA125 and CA19-9 improves the sensitivity for the early diagnosis of gastric cancer.** *Clin Chim Acta* 2014, **437**:183-186.
 30. Tong Y, Zhao Y, Shan Z, Zhang J: **CA724 predicts overall survival in locally advanced gastric cancer patients with neoadjuvant chemotherapy.** *BMC Cancer* 2021, **21**(1):4.

Table 1. Baseline Characteristics of Train, Internal and External Datasets

Variable	Dataset					P-value ²
	N	Overall ¹	Train set ¹	Internal set ¹	External set ¹	
Gender	982					0.028
Male		624 (63.5%)	338 (60.4%)	154 (65.3%)	132 (71.0%)	
Female		358 (36.5%)	222 (39.6%)	82 (34.7%)	54 (29.0%)	
Age	982	59.16 ± 11.77	59.50 ± 11.86	58.98 ± 11.98	58.34 ± 11.21	0.256
ECOG	982					0.215
0-1		882 (89.8%)	496 (88.6%)	213 (90.3%)	173 (93.0%)	

Variable	N	Dataset				P-value ²
		Overall ¹	Train set ¹	Internal set ¹	External set ¹	
2		100 (10.2%)	64 (11.4%)	23 (9.7%)	13 (7.0%)	
Local.site	982					<0.001
Gastroesophageal junction		292 (29.7%)	169 (30.2%)	69 (29.2%)	54 (29%)	
Stomach		642 (65.4%)	391 (69.8%)	167 (70.8%)	84 (45.2%)	
Unknown		48 (4.9%)	0 (0%)	0 (0%)	48 (25.8%)	
Differentiation	982					<0.001
Poor		753 (76.7%)	465 (83.0%)	190 (80.5%)	98 (52.7%)	
Moderate/Well		229 (23.3%)	95 (17.0%)	46 (19.5%)	88 (47.4%)	
Lauren	982					0.352
Diffuse		110 (11.2%)	63 (11.3%)	27 (11.4%)	20 (10.8%)	
Mixed		61 (6.2%)	37 (6.6%)	18 (7.6%)	6 (3.2%)	
Intestinal		56 (5.7%)	37 (6.6%)	12 (5.1%)	7 (3.8%)	
Unknown		755 (76.9%)	423 (75.5%)	179 (75.8%)	153 (82.3%)	
Her2	982					<0.001
Negative		410 (41.8%)	219 (39.1%)	90 (38.1%)	101 (54.3%)	
Low expression(+)		168 (17.1%)	116 (20.7%)	47 (19.9%)	5 (2.7%)	
Median expression (2+, fish-)		117 (11.9%)	70 (12.5%)	22 (9.3%)	25 (13.4%)	
Positive (2+, fish+/3+)		46 (4.7%)	19 (3.4%)	7 (3.0%)	20 (10.8%)	
Unknown		241 (24.5%)	136 (24.3%)	70 (29.7%)	35 (18.8%)	
MMR	982					<0.001
pMMR		583 (59.4%)	380 (67.9%)	145 (61.4%)	58 (31.2%)	
dMMR		38 (3.9%)	19 (3.4%)	10 (4.2%)	9 (4.8%)	
Unknown		361 (36.8%)	161 (28.8%)	81 (34.3%)	119 (64%)	
PDL1	982					<0.001

Variable	N	Dataset				P-value ²
		Overall ¹	Train set ¹	Internal set ¹	External set ¹	
CPS≤1		239 (24.3%)	138 (24.6%)	59 (25%)	42 (22.6%)	
CPS>1		299 (30.4%)	187 (33.4%)	83 (35.2%)	29 (15.6%)	
Unknown		444 (45.2%)	235 (42%)	94 (39.8%)	115 (61.8%)	
Liver,n (%)	982					0.109
Absent		615 (62.6%)	360 (64.3%)	151 (64.0%)	104 (55.9%)	
Present		367 (37.4%)	200 (35.7%)	85 (36.0%)	82 (44.1%)	
Lung,n (%)	982					0.028
Absent		855 (87.1%)	494 (88.2%)	210 (89.0%)	151 (81.2%)	
Present		127 (12.9%)	66 (11.8%)	26 (11.0%)	35 (18.8%)	
Peritoneum,n (%)	982					<0.001
Absent		729 (74.2%)	404 (72.1%)	162 (68.6%)	163 (87.6%)	
Present		253 (25.8%)	156 (27.9%)	74 (31.4%)	23 (12.4%)	
Lymph.nodes	982					<0.001
Absent		139 (14.2%)	39 (7%)	23 (9.7%)	77 (41.4%)	
Present		843 (85.8%)	521 (93%)	213 (90.3%)	109 (58.6%)	
Metastasis.number,n (%)	982					0.311
≤3		885 (90.1%)	507 (90.5%)	207 (87.7%)	171 (91.9%)	
>3		97 (9.9%)	53 (9.5%)	29 (12.3%)	15 (8.1%)	
Surgery	982					<0.001
None		574 (58.5%)	358 (63.9%)	137 (58.1%)	79 (42.5%)	
Pre-chemotherapy surgery		286.0 (29.1%)	150.0 (26.8%)	67.0 (28.4%)	69.0 (37.1%)	
Post-chemotherapy surgery		122.0 (12.4%)	52.0 (9.3%)	32.0 (13.6%)	38.0 (20.4%)	
Immunotherapy,n (%)	982					<0.001

Variable	N	Dataset				P-value ²
		Overall ¹	Train set ¹	Internal set ¹	External set ¹	
First-line	982	787 (80.1%)	492 (87.9%)	206 (87.3%)	89 (47.8%)	<0.001
Subsequent-line		195 (19.9%)	68 (12.1%)	30 (12.7%)	97 (52.2%)	
Chemotherapy,n (%)						
Oxaliplatin plus fluorouracil		591 (60.2%)	387 (69.1%)	166 (70.3%)	38 (20.4%)	
Abraxane-based		243 (24.7%)	132 (23.6%)	52 (22%)	59 (31.7%)	
Other		148 (15.1%)	41 (7.3%)	18 (7.6%)	89 (47.8%)	
Anti-angiogenic drug,n (%)	982					0.019
None	982	835 (85.0%)	487 (87.0%)	202 (85.6%)	146 (78.5%)	<0.001
Yes		147 (15.0%)	73 (13.0%)	34 (14.4%)	40 (21.5%)	
Efficacy,n (%)	982					
PD	982	207 (21.1%)	86 (15.4%)	38 (16.1%)	83 (44.6%)	<0.001
Non-PD		775 (78.9%)	474 (84.6%)	198 (83.9%)	103 (55.4%)	
HB,Mean ± SD	982	109.47 ± 42.43	108.80 ± 49.85	106.64 ± 35.49	115.06 ± 19.84	0.343
LN,Mean ± SD	982	1.43 ± 0.68	1.47 ± 0.77	1.43 ± 0.52	1.29 ± 0.50	<0.001
WBC,Mean ± SD	982	12.43 ± 25.81	13.98 ± 29.01	13.08 ± 27.13	6.92 ± 4.04	0.788
LDH,Mean ± SD	982	215.98 ± 184.22	209.36 ± 192.20	217.51 ± 169.65	233.99 ± 177.00	0.234
NLR,Mean ± SD	982	3.45 ± 2.99	3.27 ± 2.89	3.44 ± 2.76	4.01 ± 3.48	<0.001
PLR,Mean ± SD	982	199.38 ± 160.74	190.15 ± 115.37	195.02 ± 120.35	232.70 ± 277.47	0.042
LIPI	982					0.158
0	982	544 (55.4%)	306 (54.6%)	125 (53%)	113 (60.8%)	
1		369 (37.6%)	212 (37.9%)	90 (38.1%)	67 (36.0%)	
2		69 (7.0%)	42 (7.5%)	21 (8.9%)	6 (3.2%)	

Variable	N	Dataset				P-value ²
		Overall ¹	Train set ¹	Internal set ¹	External set ¹	
PAR,Mean ± SD	982	6.67 ± 3.13	6.73 ± 3.25	6.71 ± 3.14	6.45 ± 2.73	0.746
HALP,Mean ± SD	982	29.45 ± 24.61	30.45 ± 28.29	28.44 ± 19.73	27.70 ± 17.11	0.917
PNI,Mean ± SD	982	44.88 ± 6.21	44.94 ± 6.36	44.64 ± 5.94	45.02 ± 6.11	0.538
CEA,Mean ± SD	982	124.95 ± 877.18	79.65 ± 767.36	140.82 ± 933.89	241.18 ± 1,083.00	<0.001
CA199,Mean ± SD	982	870.64 ± 5,096.82	915.26 ± 5,122.06	868.51 ± 6,220.55	738.99 ± 3,049.74	0.264
CA724,Mean ± SD	982	59.60 ± 194.32	60.76 ± 209.37	50.12 ± 183.95	68.15 ± 156.82	<0.001

¹ Categorical variables are presented as n (%), continuous variables as mean ± standard deviation (Mean ± SD)

² Comparisons among three cohorts: Chi-square test or Fisher’s exact test for categorical variables; Kruskal-Wallis H test for continuous variables.

Table2. C-indexes of seven machine learning survival models for OS and PFS in training, internal, and external validation sets

C-index of overall survival models					C-index of progression free survival models			
Train set	Internal set	External set	Average		Train set	Internal set	External set	Average

XGBoost	0.721	0.688	0.642	0.683	0.701	0.667	0.677	0.678
GBM	0.706	0.677	0.658	0.680	0.694	0.659	0.673	0.672
LASSO	0.658	0.681	0.663	0.668	0.658	0.653	0.708	0.673
CoxBoost	0.658	0.681	0.663	0.667	0.658	0.653	0.693	0.668
plsRcox	0.658	0.680	0.662	0.667	0.659	0.654	0.684	0.666
survival-SVM	0.769	0.613	0.616	0.674	0.784	0.634	0.458	0.625
SuperPC	0.637	0.662	0.673	0.657	0.608	0.602	0.548	0.586

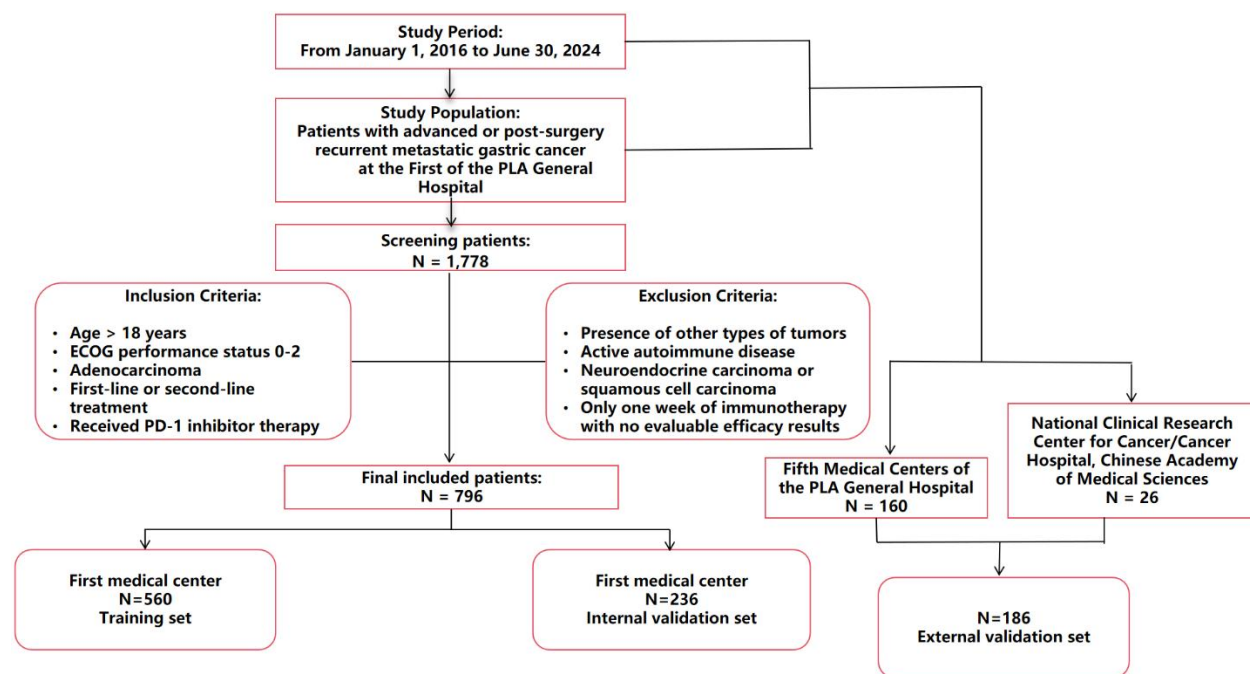


Figure1.Flowchart of patient selection process

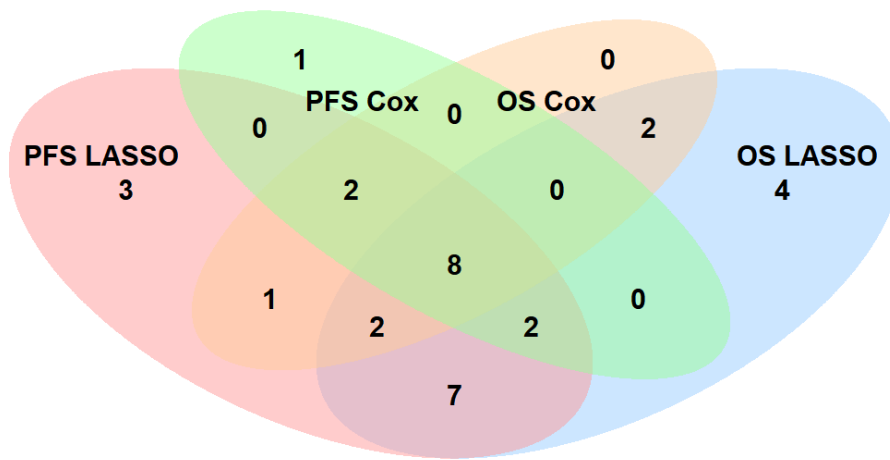


Figure2. Venn diagram of prognostic features from four screening strategies: Univariate-Cox OS, Univariate-Cox PFS, LASSO-Cox OS, LASSO-Cox PFS.

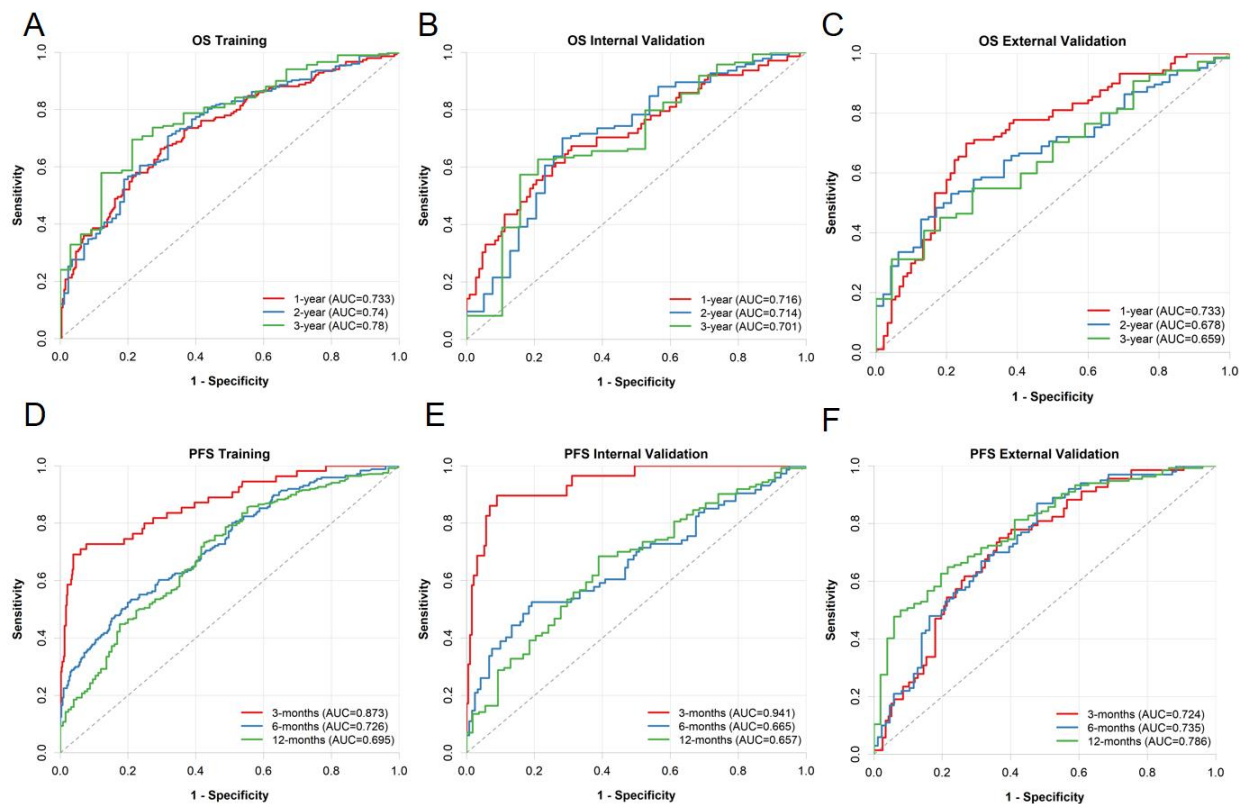


Figure3A-F. Time-dependent receiver operating characteristic (ROC) curves of the XGBoost-based prognostic model for predicting overall survival (OS) and progression-free survival (PFS). (A–C) ROC curves for 1-, 2-, and 3-year OS

prediction in the training (A), internal validation (B), and external validation (C) sets. (D–F) ROC curves for 3-, 6-, and 12-month PFS prediction in the training (D), internal validation (E), and external validation (F) sets.

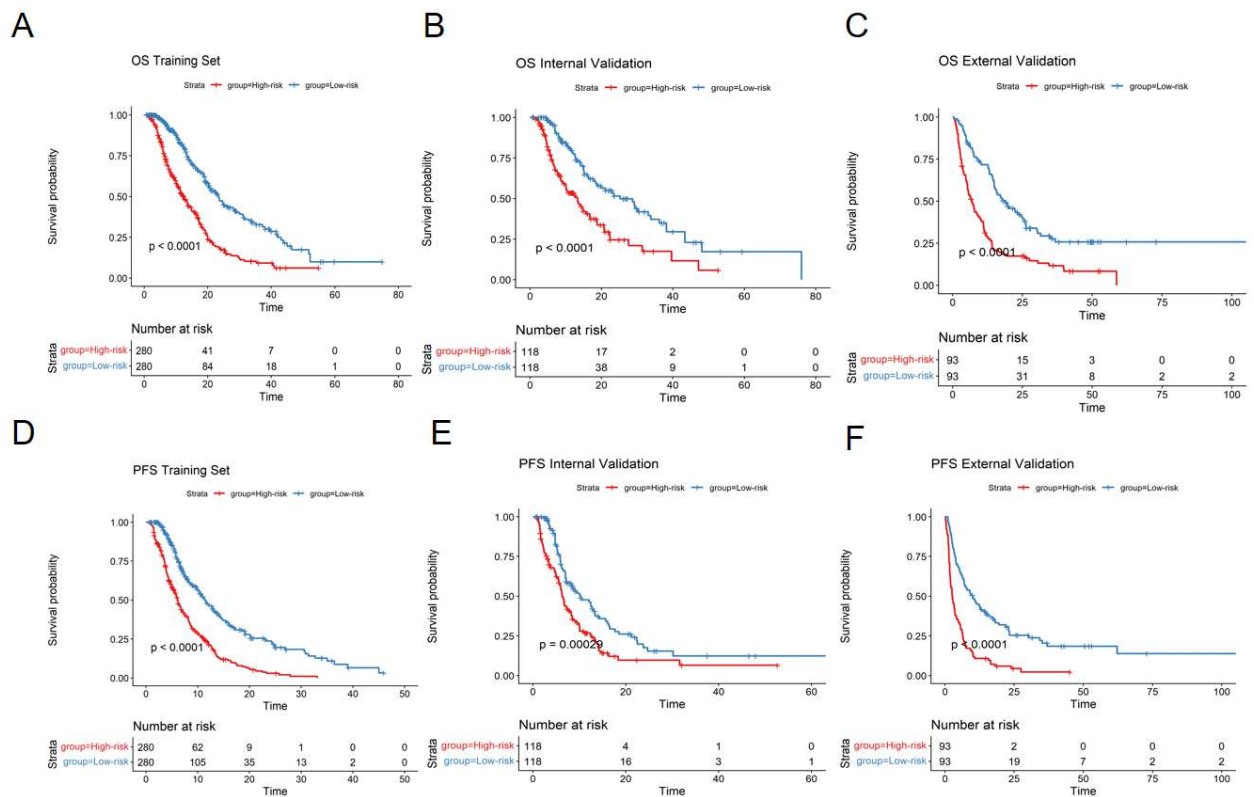


Figure 4A-F. Kaplan-Meier survival curves for overall survival (OS) and progression-free survival (PFS) stratified by risk group in different cohorts. (A–C) Kaplan-Meier curves for OS in the training (A), internal validation (B), and external validation (C) cohorts. (D–F) Kaplan-Meier curves for PFS in the training (D), internal validation (E), and external validation (F) cohorts.

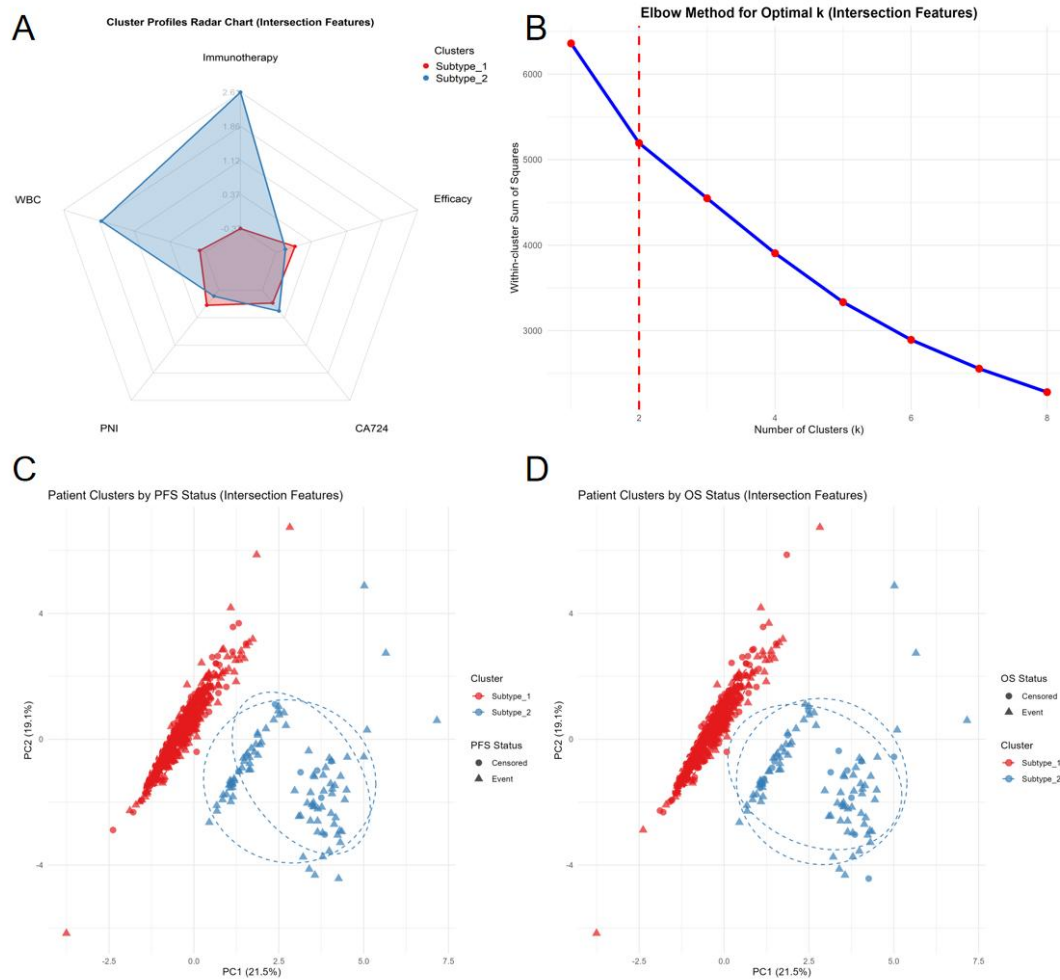


Figure 5A-D. Identification and characterization of molecular subtypes using K-means clustering based on prognostic features. (A) Elbow method plot showing the within-cluster sum of squares (WSS) versus the number of clusters, with the optimal $k = 2$ determined at the "elbow" point. (B) Radar chart showing the standardized expression patterns of the top 5 most discriminative features (WBC, PNI, CA724, treatment efficacy, and immunotherapy status) between the two clusters (Subtype 1 vs. Subtype 2). (C-D) PCA plots visualizing patient clustering patterns according to (C) progression-free survival (PFS) status and (D) overall survival (OS) status.

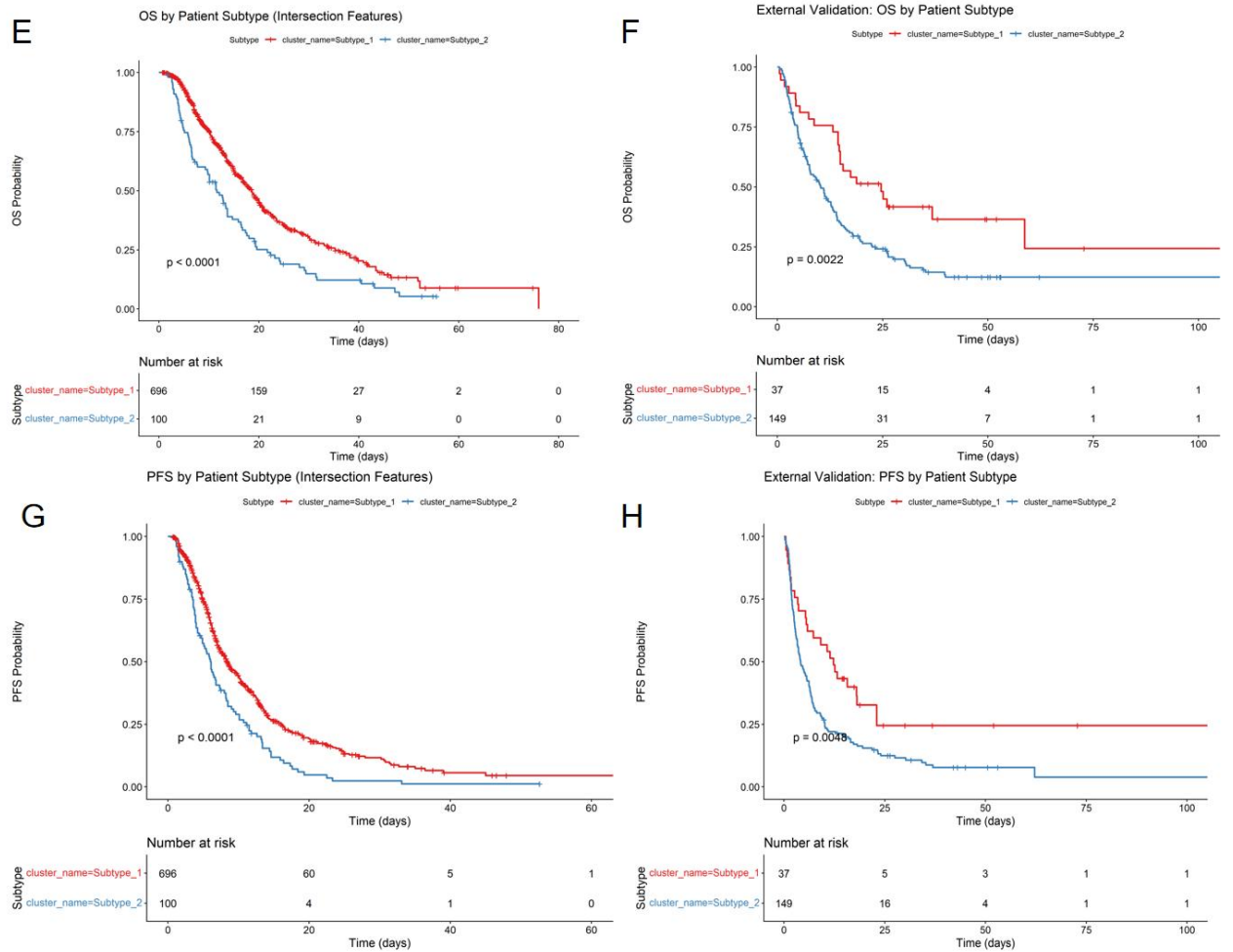


Figure 5E-H. Kaplan-Meier survival curves comparing overall survival (OS) and progression-free survival (PFS) between the two patient subtypes. (E, F) OS curves for the discovery cohort (E) and external validation cohort (F). (G, H) PFS curves for the discovery cohort (G) and external validation cohort (H).

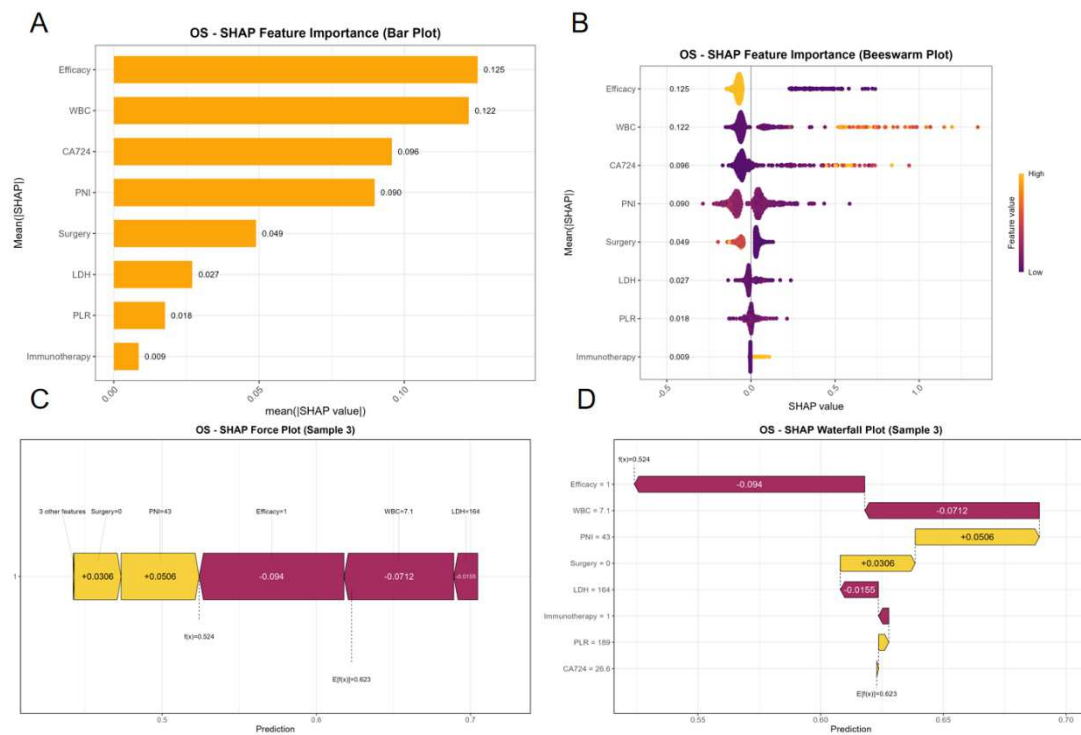


Figure 6A-D. SHAP analysis was performed to quantify and visualize the contribution of each feature to the model's predictions for OS. (A) Bar plot showing the mean absolute SHAP values for OS prediction. (B) Beeswarm plot displaying SHAP values for OS prediction. Each point represents a patient, colored by feature value, showing the direction and magnitude of each feature's effect on model output. (C) SHAP force plot illustrating feature contributions to the predicted OS risk for a representative patient. (D) Waterfall plot decomposing the SHAP values for the same patient, quantifying the contribution of each feature to the final prediction score of OS. Positive SHAP values indicate increased risk of mortality, while negative SHAP values represent protective effects against mortality.

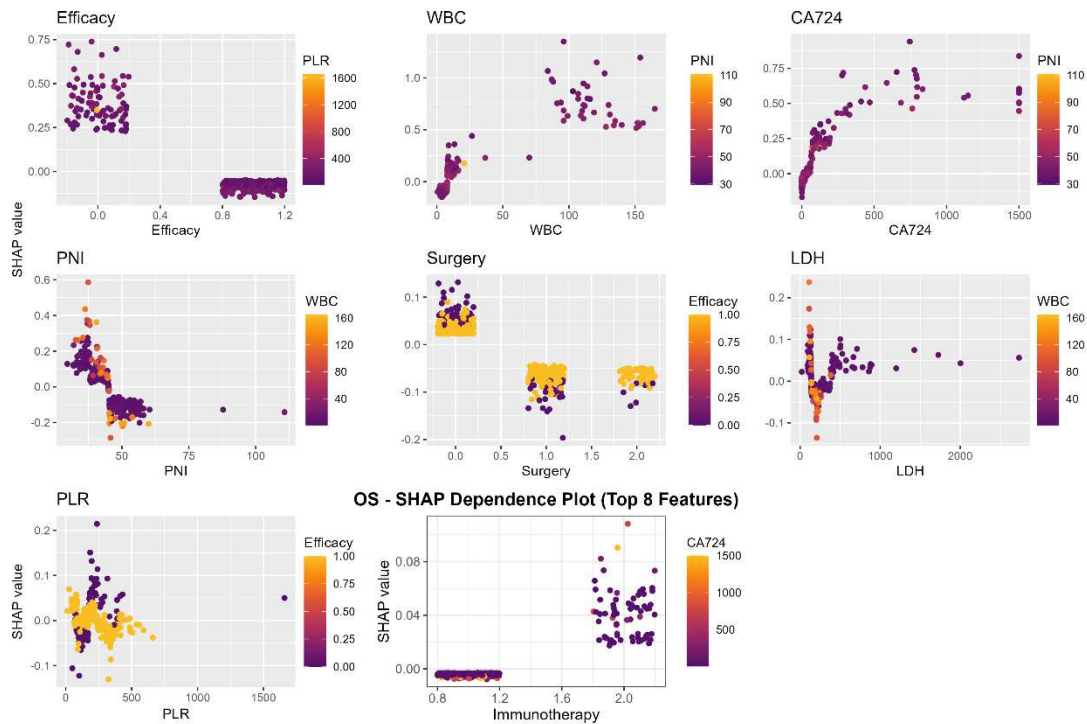


Figure6E. Nonlinear effects and feature interactions in the prognostic model revealed by SHAP dependence plots for OS prediction.

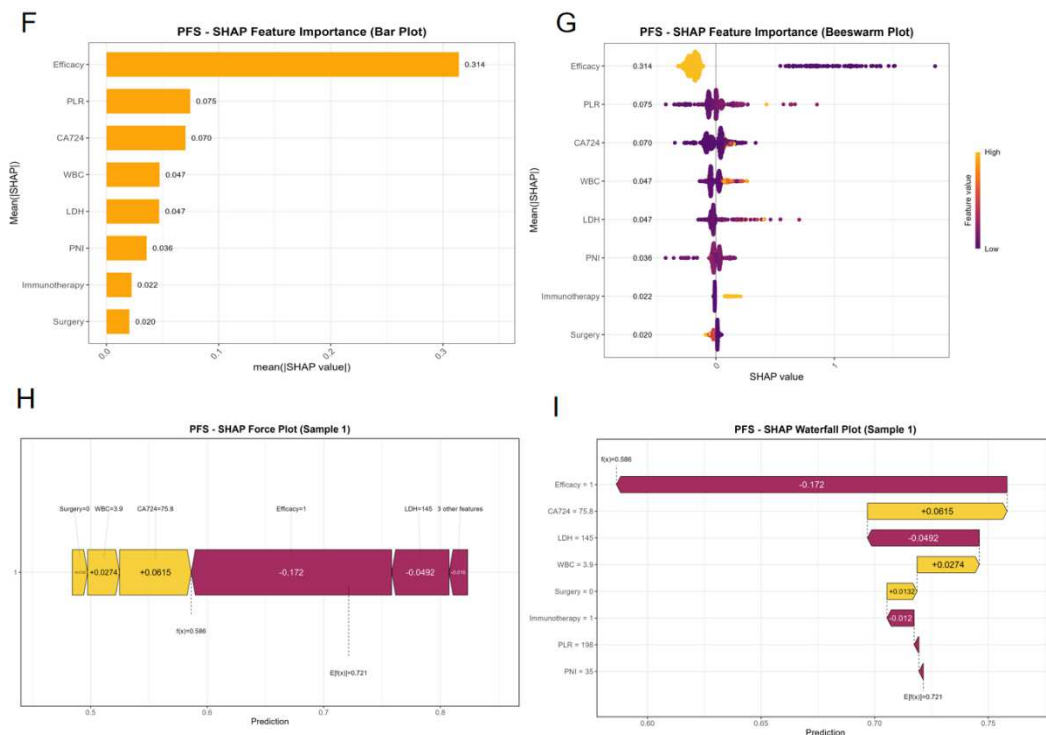


Figure6F-I. SHAP analysis was performed to quantify and visualize the contribution of each feature to the model's predictions for PFS. (F) Bar plot showing the mean absolute SHAP values for PFS prediction. (G) Beeswarm plot displaying SHAP values for PFS prediction. Each point represents a patient, colored by feature value,

showing the direction and magnitude of each feature's effect on model output. (H) SHAP force plot illustrating feature contributions to the predicted PFS risk for a representative patient. (I) Waterfall plot decomposing the SHAP values for the same patient, quantifying the contribution of each feature to the final prediction score of PFS. Positive SHAP values indicate increased risk of disease progression, while negative SHAP values represent protective factors against progression.

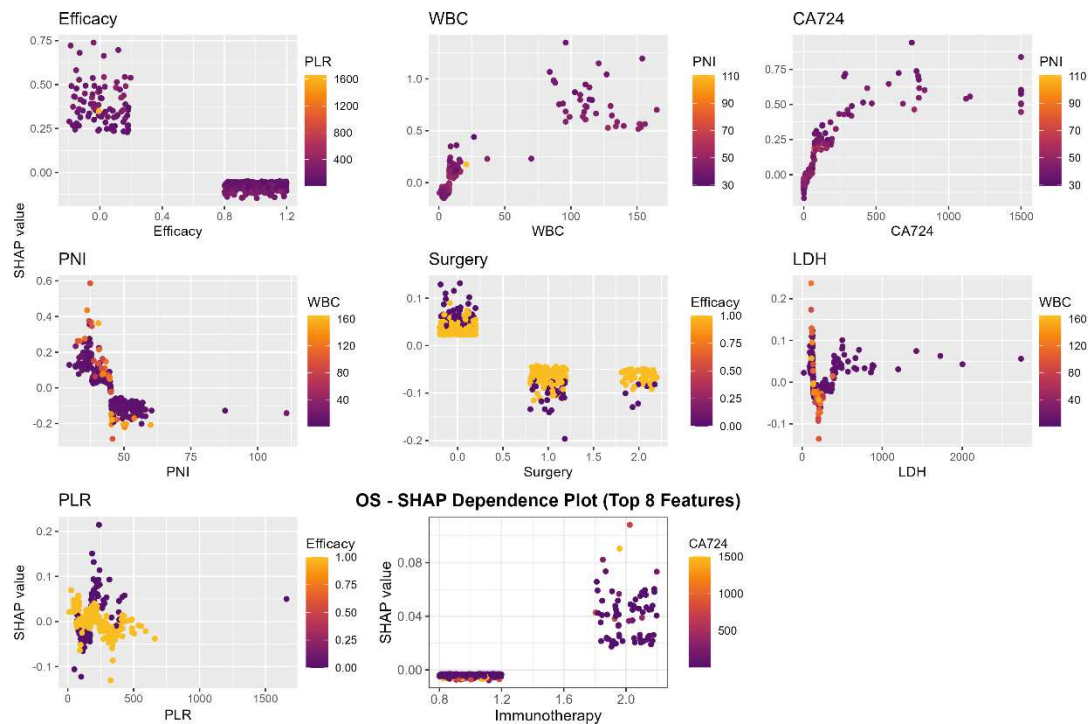


Figure6J. Nonlinear effects and feature interactions in the prognostic model revealed by SHAP dependence plots for PFS prediction.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [FigureS.doc](#)
- [TableS1.docx](#)
- [TableS2.docx](#)