Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine Learning Model Based on Multidimensional Laboratory Indicators for Predicting Distant Metastasis in Nasopharyngeal Carcinoma: A Multicenter Retrospective Cohort Study

Lili Liang
The Second Affiliated Hospital of Guangxi University of Science and Technology Liuzhou, China
Yingxi Zhao
The Second Affiliated Hospital of Guangxi University of Science and Technology Liuzhou, China
Xiangni Luo
The Second Affiliated Hospital of Guangxi University of Science and Technology Liuzhou, China
Zhengyi Xie
94302173@qq.com

Department of Medicine,Guangxi University of Science and Technology,Liuzhou
Jin Wang
Department of Medicine,Guangxi University of Science and Technology,Liuzhou

Research Article

Keywords: Nasopharyngeal carcinoma, Distant metastasis, Machine learning, Prediction model, Laboratory indicators

Posted Date: August 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10206699/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Machine Learning Model Based on Multidimensional Laboratory Indicators for Predicting Distant Metastasis in Nasopharyngeal Carcinoma: A Multicenter Retrospective Cohort Study

Lili Liang[1,2#], Yingxi Zhao[1,2#], Xiangni Luo[1,2#], Zhengyi Xie[1*], Jin Wang[1,2*]

(1.Department of Medicine, Guangxi University of Science and Technology, Liuzhou, China, 545005; 2.The Second Affiliated Hospital of Guangxi University of Science and Technology Liuzhou, China, 545005)

*** Correspondence:**

Corresponding Author

Zhengyi Xie:94302173@qq.com

Jin Wang: 2021083@gxust.edu.cn

Abstract

Background We aimed to develop and validate a machine learning (ML) model based on routine multidimensional laboratory indicators for individualized prediction of distant metastasis status in patients with newly diagnosed nasopharyngeal carcinoma (NPC).

Methods A total of 1,085 patients with newly diagnosed NPC were retrospectively enrolled. Systematic integration of five categories of laboratory parameters—complete blood count, biochemistry, coagulation, immunology, and tumor markers—was performed. Three algorithms, including XGBoost, Random Forest, and ElasticNet, were employed for model construction and comparison. Model robustness was assessed through stratified 5-fold cross-validation and 1,000 Monte Carlo cross-validations (MCCVs). SHapley Additive exPlanations (SHAP) analysis was used to elucidate feature importance. Multilevel independent validation was conducted using multivariable logistic regression, tertile-based risk stratification, external

validation with the Surveillance, Epidemiology, and End Results (SEER) database (N≈7,580), and Gene Expression Omnibus (GEO) transcriptomic validation.

Results The XGBoost model incorporating the comprehensive feature set achieved optimal predictive performance, with an area under the receiver operating characteristic curve (AUROC) of 0.8595±0.0231 and a Brier score of 0.1287. The 1,000 MCCV repetitions confirmed an AUROC of 0.8470 (95% confidence interval [CI]: 0.8032–0.8871), excluding overfitting. SHAP analysis identified red cell distribution width-standard deviation (RDW-SD; 0.452), immunoglobulin A (IgA; 0.347), carbohydrate antigen 125 (CA125; 0.301), and D-dimer (0.284) as the core predictive factors. Risk stratification demonstrated a significant dose-response relationship: low-risk 7.3% → intermediate-risk 22.2% → high-risk 56.5% (Cochran-Armitage trend test Z=11.42, P<0.0001). Multivariable logistic regression confirmed aspartate aminotransferase (AST; odds ratio [OR]=2.70), red cell distribution width (RDW; OR=1.34), and hemoglobin (Hb; OR=0.77) as independent predictors. SEER validation confirmed the strong prognostic effect of distant metastasis (overall survival [OS] hazard ratio [HR]=3.21, with 23 of 24 subgroups showing significant HRs), and the nomogram achieved a C-index of 0.752. GEO transcriptomic validation revealed that Hallmark pathways corresponding to the core predictive factors showed a systematic trend of enrichment at the nominal P-value level in GSE103611, including glycolysis (normalized enrichment score [NES]=+1.42), E2F targets (NES=+1.27), and interferon-gamma response (NES=+1.32). In GSE299775, single-sample gene set enrichment analysis (ssGSEA) validation identified a nominal P-value level difference in the immunoglobulin synthesis pathway (P=0.042).

Conclusions The ML model constructed in this study effectively predicts distant metastasis in NPC based on routine laboratory tests, provides multilevel biological mechanism interpretation for the core predictive factors, and has the potential to serve as an individualized risk assessment tool to complement tumor-node-metastasis (TNM) staging, offering evidence-based support for intensified screening of high-risk patients and optimization of treatment decision-making.

Keywords Nasopharyngeal carcinoma; Distant metastasis; Machine learning; Prediction model; Laboratory indicators

NPC nasopharyngeal carcinoma

ACD anemia of chronic disease

EBV Epstein-Barr virus

M0 no distant metastasis

M1 distant metastasis

AUROC area under the receiver operating characteristic curve

AUPRC area under the precision-recall curve

PPV positive predictive value

NPV negative predictive value

FDR false discovery rate

VIF variance inflation factor

AIC Akaike information criterion

DCA decision curve analysis

CI confidence interval

HR hazard ratio

OR odds ratio

NES normalized enrichment score

C-index concordance index

IQR interquartile range

Brier Brier score

1. Introduction

Nasopharyngeal carcinoma (NPC) is a malignant tumor originating from the epithelial cells of the nasopharynx, characterized by a marked geographic imbalance in distribution, with a high clustering pattern in Southeast Asia and southern China[1]. According to statistics,

approximately 80% of new NPC cases worldwide are concentrated in Asia, with China contributing nearly 47% of global cases[2]. In southern China, Guangdong Province and Guangxi Zhuang Autonomous Region constitute the core high-incidence areas, where the age-standardized incidence rate in males can reach 20–30 per 100,000, significantly higher than that in non-endemic areas (<1 per 100,000)[3,4]. Although the widespread adoption of intensity-modulated radiation therapy and platinum-based concurrent chemotherapy regimens has significantly improved local control rates, resulting in 5-year overall survival (OS) exceeding 90% in early-stage patients, distant metastasis remains the primary mode of treatment failure and the leading cause of NPC-related mortality[5]. Despite platinum-based concurrent chemoradiotherapy having become the standard treatment for locally advanced NPC, approximately 30% of patients ultimately experience treatment failure due to distant metastasis[6]. Retrospective data from our center showed that among 1,085 patients with newly diagnosed NPC, M1-stage patients accounted for 27.6% (299/1,085), a proportion consistent with the high metastasis risk characteristics reported in the literature, highlighting the urgent clinical need for accurate prediction of distant metastasis to enable early intervention.

The 9th edition of the American Joint Committee on Cancer (AJCC) tumor-node-metastasis (TNM) staging system classifies NPC with distant metastasis as Stage IVB[7], a classification that directly determines a fundamental shift in treatment strategy. For M0-stage patients, the comprehensive treatment strategy centered on definitive radiotherapy combined with platinum-based concurrent chemotherapy aims to achieve complete tumor remission and long-term survival (5-year OS can exceed 90%); whereas for M1-stage patients, the treatment goal shifts to a palliative care model primarily based on systemic platinum-based chemotherapy, immunotherapy, and targeted therapy, with significant differences in the role and timing of radiotherapy to the primary tumor[8]. Recent studies have suggested that some oligometastatic patients may benefit from aggressive local treatments such as stereotactic body radiotherapy, but accurate assessment of metastatic status remains a prerequisite for formulating individualized treatment plans[9]. Therefore, stratified assessment of distant metastasis risk in patients with NPC at the time of initial diagnosis not only helps to avoid underestimation of high-risk patients and inappropriate treatment de-escalation, but also provides a decision-making basis for systemic therapy for potential micrometastatic disease, carrying decisive clinical significance for improving patient prognosis.

Although the AJCC TNM staging system remains the most widely applied prognostic assessment tool for oncology to date, it has inherent limitations in individualized metastasis risk prediction for NPC. First, the TNM staging relies on macroscopic assessment of anatomical extent and cannot reflect the molecular heterogeneity of the tumor or the microenvironmental characteristics of the host response, resulting in significant differences in metastasis risk among patients within the same stage. Second, the traditional staging system adopts categorical variables (such as T1–T4, N0–N3), making it difficult to capture continuous risk gradient changes and leading to imprecise risk stratification. Third, the existing staging system has not incorporated systemic laboratory indicators that reflect host inflammatory status, nutritional reserve, coagulation function, and tumor biological behavior, yet these factors have been confirmed to be closely associated with tumor metastasis[10,11]. Therefore, there is an urgent clinical need for a precision prediction tool that transcends traditional anatomical staging and can integrate multidimensional biological information to achieve individualized and continuous assessment of distant metastasis risk in patients with NPC.

Machine learning (ML) algorithms, with their powerful capability to capture nonlinear relationships and their advantage in automated high-dimensional feature extraction, have demonstrated broad application prospects in the field of clinical prediction modeling. Most published prognostic prediction models for NPC have employed traditional logistic regression (LR) methods, which, despite having good interpretability, have limited capacity to capture complex interactions among variables[12]. A small number of studies have attempted to apply ML algorithms to NPC prognostic prediction, but they share the following common deficiencies: (1) the prediction endpoint is mostly a general survival outcome rather than a specific event such as distant metastasis that has explicit clinical decision-making value[13]; (2) the feature space is mostly confined to single-dimension clinical or imaging indicators, without adequate integration of multidimensional information encompassing inflammation, nutrition, coagulation, and tumor markers; (3) there is a lack of rigorous external validation and model interpretability analysis, resulting in insufficient generalization capability and inadequate biological mechanism interpretation; (4) there is no systematic population risk stratification, making it difficult to directly guide clinical treatment decision-making. These methodological gaps constitute the entry point of the present study.

Based on the above background, this study proposes the following core hypothesis: integrating multidimensional routine laboratory indicators (encompassing inflammatory markers, nutritional metabolic indicators, coagulation function parameters, and tumor markers) and employing ML algorithms to construct a predictive model can effectively identify the distant metastasis status (M1) of patients with NPC at initial diagnosis, and its predictive performance will be superior to that of traditional statistical models based on single-dimension indicators. To test this hypothesis, this study enrolled complete laboratory data from 1,085 patients with newly diagnosed NPC, systematically compared the predictive performance of three mainstream ML algorithms (XGBoost, Random Forest, and ElasticNet) under three feature combination schemes (routine laboratory tests, tumor markers, combined features, and stepwise selected features) to determine the optimal modeling strategy. This study aims to construct and comprehensively validate an ML prediction model for NPC distant metastasis with good discrimination, calibration, and clinical utility; to reveal the key biological features driving metastasis prediction and their interactions through SHapley Additive exPlanations (SHAP) interpretability analysis; to validate the decision value of the model in different metastasis risk subgroups based on risk stratification; and to conduct multi-omics integrative analysis using GEO transcriptomic data to explain the biological plausibility of the top predictive features at the molecular level, thereby providing novel decision-support tools and theoretical foundations for precision diagnosis and treatment of NPC.

2. Materials and Methods

2.1 Study Design, Ethics, and Data Sources

This study adopted a single-center retrospective cohort design to construct a machine learning prediction model for distant metastasis in nasopharyngeal carcinoma, and the research report strictly adhered to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) guidelines. The study protocol was approved by the Ethics Review Committee of Guangxi University of Science and Technology (approval number: 2025-02), and the research procedures complied with the ethical principles of the Declaration of Helsinki. Given the retrospective nature of this study and the use of fully

anonymized data, the requirement for informed consent was waived by the Ethics Review Committee of Guangxi University of Science and Technology (approval number: 2025-02), in accordance with national regulations for retrospective research with minimal risk.

The modeling cohort of this study included patients with primary NPC who were diagnosed and received initial treatment at the Second Affiliated Hospital of Guangxi University of Science and Technology from January 2015 to December 2023. The inclusion criteria for study subjects were as follows: (1) histopathologically confirmed diagnosis of NPC; (2) treatment-naive patients who had not previously received any anti-tumor treatment (including radiotherapy, chemotherapy, targeted therapy, or immunotherapy); (3) complete laboratory test data covering tumor markers, complete blood count, coagulation function, liver function, renal function, and electrolyte indicators; (4) explicit distant metastasis (M) staging information, with M stage determined according to the AJCC 8th edition TNM staging system; (5) complete clinicopathological data including baseline information such as age, sex, and TNM stage. Patients meeting any of the following criteria were excluded: (1) concurrent or previously diagnosed other malignant tumors; (2) missing rate of key laboratory variables exceeding 30%; (3) receipt of blood transfusion, leukocyte-elevating therapy, or anticoagulant therapy prior to laboratory blood sample collection, which might affect the authenticity of test results; (4) concurrent acute infection, autoimmune disease, hematological disorder, or severe hepatic or renal insufficiency; (5) incomplete follow-up information precluding confirmation of distant metastasis status.

After rigorous screening and data deduplication, a total of 1,085 patients were ultimately enrolled, including 786 (72.4%) without distant metastasis (M0) and 299 (27.6%) with distant metastasis (M1). The cohort was predominantly male (n = 810, 74.7%), and the median age was 48 years (IQR: 39–56 years).

To accomplish multi-dimensional external validation, this study integrated two public database cohorts: NPC cases diagnosed between 2010 and 2020 were extracted from the SEER database (ICD-O-3: C11.x, n ≈ 7,580), and clinical external validation was conducted based on its complete TNM staging, metastasis status, and survival data. Concurrently, two independent transcriptome datasets from the GEO database were selected for molecular mechanism validation, including GSE103611 comprising tissue microarray data from 51 NPC primary tumor

samples and GSE299775 comprising RNA-seq data from PBMCs of 40 patients, both datasets accompanied by explicit distant metastasis classification information.

2.2 Study Outcome and Feature Engineering

This study used the initial distant metastasis status of NPC patients as a binary outcome variable, categorized into absence of distant metastasis (M0) and presence of distant metastasis (M1) according to the AJCC 8th edition staging criteria, with metastasis status ultimately confirmed by imaging examination or long-term follow-up. To systematically evaluate the predictive performance of clinical indicators from different dimensions, this study constructed a three-tier hierarchical predictive feature system encompassing raw clinical detection indicators and derived biological features. The tumor marker feature set comprised 12 indicators including CEA, CA19-9, CA242, CA125, CA15-3, AFP, NSE, SCC-Ag, CYFRA21-1, ProGRP, LDH, and ferritin. The routine laboratory feature set encompassed baseline detection indicators related to complete blood count, coagulation function, liver and renal function, electrolytes, and immune function, comprehensively reflecting patient inflammation, metabolism, coagulation, and immune homeostasis. The combined feature set integrated all the above raw indicators and included composite derived indicators derived from biological mechanisms, including NLR, PLR, MLR, SII, SIRI, AGR, AST/ALT ratio, and eGFR, with a cumulative total of more than 80 high-dimensional clinical features.

2.3 Data Preprocessing and Feature Screening

To ensure data quality and avoid modeling bias, this study implemented a standardized data preprocessing pipeline, and all preprocessing operations were performed exclusively on the training set to strictly prevent data leakage. For missing data, hierarchical processing was applied according to the missing rate: 1. continuous variables with missing rate < 5% were imputed using the median, and categorical variables were imputed using the mode; 2. variables with missing rate of 5%–30% were imputed using 10-iteration multiple imputation by chained equations (MICE); 3. variables with missing rate > 30% were directly excluded. For extreme outliers, the IQR × 1.5 criterion was used to identify outlier samples, and the 1st–99th percentile Winsorization truncation method was applied to correct data distribution skewness. All

continuous features were uniformly standardized using Z-score to eliminate the interference of dimensional differences on model training.

To achieve high-dimensional feature dimensionality reduction and eliminate redundant noise, this study adopted a three-step progressive feature screening strategy. First, candidate features with significant inter-group differences ($P < 0.05$) were screened through univariate Mann-Whitney U test or chi-square test. Subsequently, highly collinear indicators were eliminated based on Spearman correlation analysis, and for feature pairs with $|r| > 0.8$, the variable with stronger association with the metastasis outcome was retained. Finally, the optimal penalty coefficient lambda was determined through LASSO regression combined with 10-fold cross-validation, and a core feature subset with both statistical significance and predictive independence was selected for subsequent model construction.

2.4 Machine Learning Model Construction, Optimization, and Interpretability Analysis

This study selected three machine learning models with complementary algorithmic mechanisms and suitability for high-dimensional clinical data to construct an NPC distant metastasis prediction system. Three algorithms were employed to build predictive models: XGBoost (eXtreme Gradient Boosting), Random Forest, and ElasticNet. According to feature combinations, three feature sets were defined: Routine Labs, Tumor Markers, and Combined, yielding a total of nine model-feature set combinations. The training process adopted stratified 5-fold cross-validation repeated 20 times (stratified 5-fold CV × 20 repeats) to ensure consistent M0/M1 proportions across all folds. Hyperparameters were optimized through grid search combined with cross-validation, uniformly balancing model fitting performance and generalization capability.

The core parameter settings were as follows: XGBoost learning rate 0.05, maximum tree depth 6, subsample and colsample ratios both 0.8, L1 regularization coefficient 0.1, L2 regularization coefficient 3, number of boosting rounds 500; Random Forest number of trees 500, maximum tree depth 10, minimum samples for split 5, minimum samples at leaf node 2; ElasticNet regularization strength 0.01, L1/L2 mixing ratio 0.5.

This study performed SHAP interpretability analysis on the optimal model, computing SHAP values for each feature through 100 Monte Carlo approximations to precisely quantify the contribution direction and effect magnitude of individual features on distant metastasis prediction outcomes. Through global feature importance ranking, beeswarm plots, and partial dependence plots, the independent predictive effects, numerical distribution patterns, and inter-feature interactions of core features were systematically analyzed, elucidating the predictive mechanisms of key clinical indicators at the algorithmic level.

2.5 Multi-Dimensional Model Performance Evaluation and Clinical Utility Assessment

This study constructed a comprehensive model evaluation system from three dimensions: discrimination, calibration, and clinical utility. Discrimination was assessed using AUROC as the core evaluation metric, combined with AUPRC to correct for evaluation bias arising from class imbalance and precisely quantify the model's case discrimination ability. The DeLong test was employed for pairwise comparison of performance differences between different models and feature sets, accompanied by accuracy, sensitivity, specificity, PPV, NPV, and F1 score for comprehensive evaluation of model classification performance. Calibration was jointly assessed through Brier score, Van Calster calibration slope and intercept, and the Hosmer-Lemeshow goodness-of-fit test to comprehensively determine the concordance between predicted probabilities and actual clinical outcomes. Clinical utility was assessed by decision curve analysis (DCA).

2.6 Multi-Level Model Validation and Sensitivity Analysis

To systematically validate the reliability and robustness of the model, this study conducted internal validation, traditional statistical validation, risk stratification validation, and SEER external validation.

MCCV was employed for internal validation, repeated 1,000 times with 7:3 stratified random sampling to partition training and testing sets. The mean, standard deviation, and 95% CI derived from multiple iterations were used to systematically evaluate model generalization robustness. A multivariable stepwise Logistic regression model was constructed based on core features selected by machine learning, with independent predictors screened according to the AIC criterion. Multiple comparison bias was corrected using the FDR Benjamini-Hochberg

method, collinearity was strictly excluded with VIF < 5, and the independent predictive effects of each feature were visualized using forest plots. Concurrently, risk stratification was performed based on tertiles of model-predicted probabilities, combined with the Cochran-Armitage trend test to validate the gradient differences in metastasis rates across low-, intermediate-, and high-risk groups. Sensitivity analysis was conducted through multi-scenario feature elimination trials, sequentially deleting tumor markers, inflammatory derived indicators, coagulation parameters, and demographic variables to compare the fluctuation of model AUROC and verify the stability and irreplaceability of core features.

In the SEER external cohort validation, independent risk factors for NPC distant metastasis were screened through multivariable Logistic regression, combined with Cox proportional hazards model analysis of the prognostic association between core features and patient OS and disease-specific survival, with model applicability verified by Schoenfeld residual test. A nomogram prediction model was constructed based on core predictors, and internal calibration was completed through 1,000 Bootstrap resamplings to evaluate the model's external generalization capability.

2.7 GEO Transcriptomic Pathway Analysis

Two independent datasets were retrieved from the GEO database: GSE103611 comprised 48 primary NPC tumor tissues (24 metastatic and 24 non-metastatic cases), and GSE299775 comprised PBMC samples from 24 NPC patients (14 metastatic and 10 non-metastatic cases). Each dataset contained gene expression profiles from metastatic (M) and non-metastatic (NM) patients. Detection was performed on the Affymetrix Human Transcriptome Array 2.0 platform (GPL19251), and raw CEL files were normalized to the probe level using RMA (Robust Multi-array Average). For genes with multiple probes, the median expression value was taken as the final expression value. Clinical information included metastasis status and hematological indicators (D-dimer, IgA, RDW-SD, LDH, CA125, β2-MG, AST/ALT), which were used as continuous variables for subsequent correlation analyses.

Pre-ranked gene set enrichment analysis was performed first. Metastasis status (M vs NM) was used as the binary phenotype, and differential expression t-statistics for each gene were calculated using limma (v3.54) and ranked in descending order of t-value as the gene ranking.

Based on the biological functions of the model Top 20 predictors, corresponding pathways were selected from the MSigDB Hallmark gene set library (v2023.1), supplemented by four custom functional gene sets, totaling 50 pathways. For each pathway, the enrichment score (ES) was calculated and normalized to NES (normalized enrichment score), P values were estimated through 1,000 phenotype permutations, and the false discovery rate (FDR) was controlled using the Benjamini-Hochberg method. The significance threshold was set at FDR < 0.1. Additionally, to explore the association between pathways and specific clinical indicators, patients were dichotomized into high and low groups according to the median of each indicator, and the above GSEA procedure was repeated to obtain the NES and significance of each pathway under different indicator groupings.

Single-sample enrichment analysis was performed using the ssGSEA (single-sample Gene Set Enrichment Analysis) algorithm, computing enrichment scores (ES) for the above 50 pathways for each sample. Subsequently, limma's linear model was used to compare ES differences between the M and NM groups, yielding log2 fold change (logFC), P values, and FDR. Concurrently, differential analysis between high and low groups was performed according to the median grouping of each clinical indicator, obtaining ssGSEA scores' logFC and statistical significance.

This study utilized two independent NPC transcriptome cohorts from the GEO database. The GSE103611 cohort comprised 48 primary NPC tumor tissues (24 metastatic and 24 non-metastatic cases), detected on the Affymetrix Human Transcriptome Array 2.0 platform (GPL19251), with raw CEL files normalized to the probe level using RMA (Robust Multi-array Average). The GSE299775 cohort comprised PBMC samples from 24 NPC patients (14 metastatic and 10 non-metastatic cases), with whole-transcriptome sequencing performed using the NanoString GeoMx Digital Spatial Profiler to obtain raw count matrices. Based on the biological functions of the model Top 20 predictors, corresponding pathways were selected from the MSigDB Hallmark gene set library (v2023.1), supplemented by four custom functional gene sets. Single-sample enrichment analysis was performed using the ssGSEA (single-sample Gene Set Enrichment Analysis) algorithm. The enrichment score (ES) for each gene set was independently calculated for each sample. The GSEA method with whole-gene ranking was employed. The significance threshold was FDR < 0.25, with nominally significant results of P <

0.05 also reported. The GSEA method with whole-gene ranking was employed. P values were adjusted to FDR using the BH method, with FDR < 0.25 considered significant.

2.8 Statistical Tools and Research Transparency Statement

All data cleaning, statistical analysis, model construction, and visualization in this study were performed using R 4.3.1, with core dependencies including mainstream authoritative analysis packages such as xgboost, glmnet, caret, pROC, fastshap, limma, clusterProfiler, and ggplot2. All statistical tests were two-sided, with the significance level set at $\alpha = 0.05$. Multiple comparisons were uniformly corrected using the Benjamini-Hochberg method for FDR to avoid false-positive bias. The single-center clinical raw data of this study will not be publicly disclosed due to patient privacy and ethical constraints; the complete analysis code will be open-sourced on the GitHub platform upon formal publication of the manuscript. Raw data from the SEER and GEO public databases can be obtained through official compliant channels, comprehensively ensuring the reproducibility and transparency of the research findings.

3. Results

3.1 Patient Baseline Characteristics and Data Quality Assessment

After applying exclusion criteria and deduplication, a total of 1,085 newly diagnosed nasopharyngeal carcinoma patients were enrolled as the model development cohort, including 786 (72.4%) patients without distant metastasis (M0 stage) and 299 (27.6%) patients with distant metastasis (M1 stage). The baseline distributions of clinical characteristics and 40 laboratory parameters for the two groups are presented in Supplementary **Table 1**.

Regarding demographic characteristics, the proportion of male patients was significantly higher in the M1 group than in the M0 group (83.6% vs. 71.4%, $P=2.30\times10^{-5}$), suggesting that male sex may be an independent risk factor for distant metastasis. In terms of age distribution, the median age of patients in the M1 group was 51 years (interquartile range [IQR]: 42–60 years), which was significantly higher than the median age of 47 years in the M0 group (IQR: 39–56 years), with a statistically significant difference between the two groups (P=0.015), indicating that older patients face a higher risk of metastasis.

Among the 43 candidate predictor variables, 23 variables showed significant differences between the M0 and M1 groups ($P<0.05$), demonstrating systematic inter-group distributional shifts that provided a favorable feature discrimination foundation for subsequent machine learning model construction. Erythrocyte-related parameters generally showed abnormal alterations in the M1 group: red cell distribution width (RDW) was significantly elevated (15.20 vs. 13.70, $P=1.14\times10^{-20}$), whereas hemoglobin (Hb: 105.00 vs. 120.00 g/L, $P=1.62\times10^{-13}$), hematocrit (HCT: 32.30% vs. 36.20%, $P=2.60\times10^{-12}$), red blood cell count (RBC: 3.66 vs. 4.05×10^{12}/L, $P=1.06\times10^{-10}$), and mean corpuscular hemoglobin concentration (MCHC: 323 vs. 331 g/L, $P=3.43\times10^{-10}$) were all significantly decreased, collectively presenting a characteristic laboratory pattern of anemia of chronic disease (ACD). Regarding immune and tumor markers, immunoglobulin A (IgA: 2.07 vs. 1.67 g/L, $P=5.19\times10^{-10}$) and β_2-microglobulin (β2-MG: 1.93 vs. 1.69 mg/L, $P=1.81\times10^{-9}$) were significantly elevated in the M1 group, reflecting a higher tumor burden and immune activation status in M1-stage patients. For coagulation function indicators, D-dimer (D-dimer: 0.40 vs. 0.20 mg/L, $P=2.13\times10^{-5}$) and fibrinogen degradation products (FDP: 2.83 vs. 2.36 μg/mL, $P=4.15\times10^{-5}$) were both significantly elevated, suggesting the presence of systemic coagulation activation in metastatic patients. Additionally, complement C4 (0.26 vs. 0.22 g/L, $P=8.91\times10^{-7}$) and ferritin (Ferritin: 625.7 vs. 444.4 ng/mL, $P<0.001$) were also significantly elevated in the M1 group, further supporting the close association between systemic inflammatory response and tumor metastasis.

3.2 Machine Learning Model Construction and Performance Comparison

To systematically evaluate the predictive efficacy of different feature subset and algorithm combinations for distant metastasis in nasopharyngeal carcinoma, this study employed a full factorial experimental design using three machine learning algorithms (XGBoost, Random Forest, ElasticNet) and three feature subsets (Routine Labs, Tumor Markers, Combined), constructing a total of 9 model-feature set combinations. A stratified 5-fold cross-validation strategy was adopted to ensure consistent proportions of M0 and M1 categories across all folds, with each model undergoing 20 independent runs to reduce variability introduced by random seeds. AUROC was used as the primary evaluation metric, with AUPRC, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, and Brier score serving as secondary evaluation metrics, comprehensively characterizing model

performance from both discrimination and calibration dimensions. Detailed performance metrics for all 9 models are summarized in **Table 2**, with corresponding receiver operating characteristic (ROC) curves, precision-recall curves, calibration curves, and decision curve analysis (DCA) shown in **Figure 1A-D**; the comprehensive performance bar chart and the confusion matrix of the optimal model are presented in **Figure 2A-B**.

The combined feature set (Combined) paired with the XGBoost algorithm achieved the optimal predictive performance, with an AUROC of 0.8595±0.0231, AUPRC of 0.6275, accuracy of 82.03%, sensitivity of 46.82%, specificity of 95.42%, positive predictive value of 79.54%, negative predictive value of 82.51%, F1 score of 0.5895, and Brier score of 0.1223. Among all 9 combinations, this model not only achieved the highest AUROC but also demonstrated the best calibration performance — its calibration curve was closest to the ideal diagonal (**Figure 1C**), with the lowest Brier score (0.1287), indicating the highest concordance between predicted probabilities and actual observed probabilities. In decision curve analysis, the XGBoost-Combined model exhibited the highest net benefit within the clinically relevant threshold probability range of 20% to 60% (**Figure 1D**), suggesting that clinical decision-making using this model within this probability interval could provide the greatest added value for patients.

In algorithm-level horizontal comparisons, XGBoost consistently outperformed the corresponding Random Forest and ElasticNet models across all three feature sets. On the combined feature set, XGBoost (AUROC=0.8595±0.0231) outperformed Random Forest (0.8548±0.0224) and ElasticNet (0.7972±0.0291), with clinically meaningful inter-algorithm differences; on the routine labs feature set, XGBoost (AUROC=0.8472±0.0276) outperformed Random Forest (0.8362±0.0262) and ElasticNet (0.7799±0.0250); on the tumor markers feature set, XGBoost (AUROC=0.8003±0.0307) also outperformed Random Forest (0.7921±0.0377) and ElasticNet (0.7202±0.0486). This cross-feature-set consistent advantage indicates that XGBoost's gradient boosting mechanism possesses superior capability in capturing non-linear relationships and feature interactions present in this study's data.

In feature set-level longitudinal comparisons, the combined feature set achieved optimal performance across all three algorithms, confirming the added predictive value of multi-dimensional feature integration. Paired statistical test results are presented in **Table 3**: XGBoost

on the combined feature set achieved a significantly higher AUROC than on the routine labs feature set (P=0.038), and also significantly higher than on the tumor markers feature set (P=0.001); XGBoost on the routine labs feature set also achieved a significantly higher AUROC than on the tumor markers feature set (P=0.004). The triple comparison formed a complete statistical evidence chain (Combined > Routine Labs > Tumor Markers), demonstrating that integrating routine blood tests, coagulation function, liver and kidney function, immune indicators, and tumor markers into a multi-dimensional prediction panel can capture complementary predictive information unattainable by any single dimension alone.

Notably, although Random Forest achieved an AUROC on the combined feature set close to that of XGBoost (0.8548 vs. 0.8595, difference of only 0.005), its sensitivity was substantially lower than XGBoost (0.3442 vs. 0.5019), with correspondingly higher specificity (0.9811 vs. 0.9387), indicating an essential difference between the two algorithms in the trade-off between false negatives and false positives. In the clinical scenario of distant metastasis screening, improved sensitivity is typically of greater practical value, as missed diagnosis of metastatic patients would lead to delayed treatment opportunities. ElasticNet, serving as a linear baseline model, performed relatively poorly across all feature sets, with an AUROC of only 0.7202±0.0486 on the tumor markers feature set in particular, reflecting the significant non-linear characteristics of distant metastasis prediction in nasopharyngeal carcinoma, which cannot be adequately captured by linear combinations alone for the complex interaction effects among variables.

3.3 SHAP Interpretability Analysis

To reveal the biological drivers underlying the optimal XGBoost model's predictive decisions, this study employed the SHAP (SHapley Additive exPlanations) framework for global interpretability analysis based on the combined feature set model. SHAP values quantify the contribution of each feature to each individual prediction from a game-theoretic perspective, where positive values indicate a push toward the metastasis direction and negative values indicate a push toward the non-metastasis direction. **Figure 3A** presents the SHAP importance ranking of the top 20 features (measured by mean absolute SHAP values), and Figure 3B displays the inter-feature Spearman correlation heatmap; panels A-B provide SHAP bar plots and bee swarm summary plots respectively, offering more comprehensive visualization details.

The feature importance ranking in **Figure 3A** shows that the top five predictors were: red cell distribution width-standard deviation (RDW-SD, mean absolute SHAP value=0.452), immunoglobulin A (IgA, 0.347), carbohydrate antigen 125 (CA125, 0.301), D-dimer (D-dimer, 0.284), and lactate dehydrogenase (LDH, 0.268). These five features contributed 38.2% of the total absolute SHAP value across all 38 features, constituting the core driver module of model prediction. Among them, the SHAP value of RDW-SD (0.452) was markedly higher than the second-ranked IgA (0.347), with a gap of 30.3%, indicating that erythrocyte volume heterogeneity is the strongest single predictor of distant metastasis in nasopharyngeal carcinoma. This finding is entirely consistent with the effect direction of RDW in traditional Logistic regression, forming a cross-methodological validation.

The inter-feature correlation analysis in **Figure 3B** is a critical step for controlling model redundancy and ensuring feature independence. Results showed a strong positive correlation between red blood cell count (RBC) and hemoglobin (Hb) (Spearman $\rho=+0.92$), exceeding the collinearity alert threshold ($|\rho|>0.8$). Based on the redundancy control principle, RBC was removed from the feature set, and 37 independent features were retained in the final model. After RBC exclusion, the absolute correlation coefficients among all remaining features were below 0.8, with a mean $|\rho|$ of 0.18; the maximum correlation appeared between MCHC and Hb ($\rho\approx-0.65$), and the overall collinearity level was low, satisfying the independent feature assumption. Supplementary figures are shown in **Figure 4A-B.**

To systematically understand the contribution landscape of each biological functional module to model prediction, this study classified the 20 features into seven functional categories. **Table 4** summarizes the relative contribution proportions of each category, and the hierarchical clustering analysis results are shown in **Figure 5A**, identifying five stable feature clusters: coagulation cluster, tumor marker cluster, immune-inflammation cluster, erythrocyte cluster, and hepatic metabolism cluster. The contribution ranking by category was as follows: tumor markers (Tumor_markers) contributed the highest proportion, accounting for 29.74% of total |SHAP|; immune-inflammation indicators (Immune_inflammation) contributed 20.48%; hepatic metabolic indicators (Liver_metabolic) contributed 15.98%; erythrocyte lineage features (Hematologic_erythrocyte) contributed 15.27%; demographic characteristics (Demographics) contributed 6.97%; derived indices (Derived_indices) contributed 5.82%; and coagulation

function indicators (Coagulation) contributed 5.74%. This distribution exhibited a pronounced long-tail characteristic — the top three functional categories (tumor markers, immune-inflammation, and hepatic metabolism) collectively contributed 66.20% of the predictive information, indicating that model prediction primarily relies on multi-dimensional biological signals reflecting tumor burden, host inflammatory response, and hepatic metabolic dysfunction, rather than a single functional module. This multi-module synergistic prediction pattern also provides an explanation for why the combined feature set significantly outperforms single-dimension feature sets.

3.4 MCCV Robustness Validation

To evaluate the stability of model performance estimates and exclude potential bias from specific data partitions, this study implemented 1,000 independent Monte Carlo cross-validation (MCCV) iterations. In each iteration, 70% of samples were randomly drawn as the training set and 30% as the validation set; all three algorithms were retrained without stratification constraints and validation performance was recorded, to simulate a more stringent real-world application scenario.

The median AUROC and 95% confidence intervals (CI) for the three algorithms under MCCV are presented in **Table 5** and **Figure 5B**: XGBoost was 0.8470 (95% CI: 0.8032–0.8871), Random Forest was 0.8344 (95% CI: 0.7848–0.8808), and ElasticNet was 0.7871 (95% CI: 0.7385–0.8314). Compared with the stratified 5-fold cross-validation results, the absolute differences in AUROC for all algorithms were less than 0.021, demonstrating that performance estimates were insensitive to resampling strategy: XGBoost decreased by 0.0125 (from 0.8595 to 0.8470), Random Forest decreased by 0.0205 (from 0.8548 to 0.8344), and ElasticNet decreased by 0.0101 (from 0.7972 to 0.7871). XGBoost maintained the highest AUROC in MCCV, and its 95% CI lower bound (0.8032) exceeded the upper bounds of Random Forest (0.8808) and ElasticNet (0.8314), with non-overlapping intervals, further confirming the robust superiority of the XGBoost combined feature set model in a statistical sense. The performance ranking among the three algorithms under the MCCV framework was completely consistent with that of 5-fold cross-validation, indicating high reproducibility of model performance estimates.

3.5 Sensitivity Analysis

To systematically evaluate the impact of feature selection strategies, missing data patterns, and specific variable categories on model performance, this study designed four complementary sensitivity analysis scenarios; results are presented in **Table 6** and **Figure 6A-B**.

Scenario 1 excluded 22 high-differential variables with M0/M1 missing rate differences exceeding 5% (including FDP, IgA, and other variables with significant inter-group missing differences) to assess potential bias from differential missingness on model performance. Under this scenario, the model AUROC decreased to 0.8378 (a decrease of 0.0226 from the main analysis), with AUPRC of 0.7129; the limited magnitude of performance decrease suggests that the main analysis results are insensitive to missing patterns. Scenario 2 excluded all tumor markers and retained only routine labs and coagulation indicators to verify the independent contribution of tumor markers, yielding an AUROC of 0.8460 (decrease of 0.0144) and AUPRC of 0.6887, indicating a modest but identifiable added value of tumor markers to model performance. Scenario 3 (main analysis scenario) excluded coagulation variables with an AUROC of 0.8604 and AUPRC of 0.7310, achieving the best performance among all scenarios, suggesting that the inclusion of coagulation variables may reduce the effective sample size due to their high missing rate (>50%). Scenario 4 used only routine lab variables (without tumor markers and coagulation indicators), with an AUROC of 0.8422 (decrease of 0.0182) and AUPRC of 0.6889, representing the most conservative performance estimate.

Taken together, the AUROC range across the four scenarios was 0.8378–0.8604, with a maximum decrease of only 0.0226, and all scenarios maintained high predictive performance with AUROC>0.83. These results confirm from multiple dimensions that the optimal model possesses reasonable robustness to feature subset selection and missing data patterns; XGBoost maintained algorithm-level optimal performance across all sensitivity scenarios, further reinforcing its position as the recommended model.

3.6 Clinical Validation with Traditional Statistical Models

To independently verify the key predictors identified by the machine learning model from a traditional statistical perspective and to establish a quantitative risk model directly interpretable

by clinicians, this study conducted systematic Logistic regression analyses in the development cohort. Univariate analysis results are shown in **Table 7**; after FDR correction (Benjamini-Hochberg method), 18 variables were significantly associated with distant metastasis (FDR<0.05).

In univariate analysis, RDW ranked first in effect magnitude (OR=1.763, 95% CI: 1.535–2.024, FDR=4.85×10^{-14}), meaning that each standard deviation increase in RDW was associated with a 76.3% increase in distant metastasis risk. Hb (OR=0.583, 95% CI: 0.507–0.671, FDR=1.04×10^{-12}), HCT (OR=0.605, 95% CI: 0.526–0.695, FDR=2.28×10^{-11}), and MCHC (OR=0.668, 95% CI: 0.581–0.769, FDR=1.61×10^{-7}) were all significant protective factors, meaning that lower erythrocyte parameter levels were associated with higher metastasis risk. β2-MG (OR=1.545, 95% CI: 1.341–1.781, FDR=2.27×10^{-8}), a classical marker reflecting tumor burden, also showed a strong positive association. Additionally, albumin (Alb: OR=0.713, FDR=5.99×10^{-6}), sex (OR=1.365, FDR=3.07×10^{-4}), eosinophils (Eos: OR=1.269, FDR=6.41×10^{-4}), indirect bilirubin (IBil: OR=0.747, FDR=6.98×10^{-4}), AST (OR=2.152, FDR=4.28×10^{-3}), potassium (K: OR=1.256, FDR=4.28×10^{-3}), total bilirubin (TBil: OR=0.795, FDR=1.38×10^{-2}), blood urea nitrogen (BUN: OR=1.206, FDR=1.76×10^{-2}), globulin (Glob: OR=1.204, FDR=1.80×10^{-2}), magnesium (Mg: OR=0.831, FDR=1.80×10^{-2}), phosphorus (P: OR=0.796, FDR=1.80×10^{-2}), age (OR=1.191, FDR=3.21×10^{-2}), and monocyte percentage (Mono%: OR=1.174, FDR=4.21×10^{-2}) all remained statistically significant after FDR correction.

Multivariate Logistic regression analysis (**Table 8** and **Figure 7A**) included all 59 candidate variables; after controlling for mutual effects among variables, 9 variables independently predicted distant metastasis (P<0.05). AST demonstrated the strongest independent predictive effect (OR=2.701, 95% CI: 1.613–4.522, P=1.58×10^{-4}), meaning that after adjustment for all other variables, each standard deviation increase in AST was associated with a 170.1% increase in metastasis risk. Indirect bilirubin ranked second (OR=2.406, 95% CI: 1.324–4.371, P=3.96×10^{-3}). Male sex was an independent risk factor (OR=1.418, 95% CI: 1.197–1.679, P=5.18×10^{-5}). RDW maintained independent predictive value in the multivariate model (OR=1.337, 95% CI: 1.115–1.602, P=1.71×10^{-3}); β2-MG (OR=1.236, 95% CI: 1.039–1.470, P=1.67×10^{-2}), eosinophils (OR=1.231, 95% CI: 1.076–1.408, P=2.45×10^{-3}), and globulin (OR=1.170, 95% CI: 1.010–1.356, P=3.60×10^{-2}) were all independent risk factors. Total

bilirubin (OR=0.377, 95% CI: 0.204–0.697, P=1.85×10^{-3}) and hemoglobin (OR=0.767, 95% CI: 0.624–0.944, P=1.24×10^{-2}) were independent protective factors.

Notably, 7 of the 9 independent predictors identified in multivariate analysis (RDW, Hb, β2-MG, AST, IBil, sex, Glob) simultaneously appeared in the top 20 features of the machine learning model's SHAP importance ranking. The two methodological approaches demonstrated high consistency in the identification of key predictors, substantially strengthening the evidence for these variables as predictive biomarkers of distant metastasis in nasopharyngeal carcinoma.

3.7 Risk Stratification Validation

To validate the clinical utility of the optimal XGBoost model in risk stratification and to transform the model-output metastasis probabilities into a clinically actionable risk stratification tool, this study divided the test set samples into three equal tiers — low-risk, intermediate-risk, and high-risk — based on the aggregated prediction results from 5-fold cross-validation. Stratification results are presented in **Table 9**, with risk gradient visualization shown in **Figure 7B**.

In the low-risk group (N=109), 8 patients had actual distant metastasis, with a metastasis rate of 7.3%; the median predicted probability was 0.0345 (IQR: 0.0168–0.0701), with the overall interquartile range remaining at a low level. In the intermediate-risk group (N=108), 24 patients had actual distant metastasis, with a metastasis rate of 22.2%; the median predicted probability was 0.1544 (IQR: 0.1062–0.2455), approximately 4.5-fold that of the low-risk group. In the high-risk group (N=108), 61 patients had actual distant metastasis, with a metastasis rate as high as 56.5%; the median predicted probability was 0.5147 (IQR: 0.3498–0.7262), indicating that approximately half of the high-risk group patients actually had distant metastasis. The Cochran-Armitage trend test confirmed a significant monotonically increasing trend in metastasis rates across the three groups (Z=11.42, P<0.0001), excluding the possibility of random fluctuation. The metastasis rate ratio between the high-risk and low-risk groups reached 7.7-fold (56.5% / 7.3%), forming a clear risk gradient. These results indicate that the model can effectively stratify patients into subgroups with different metastasis risks, providing a quantitative basis for the development of individualized screening strategies.

3.8 SEER Database External Validation

To validate the impact of distant metastasis on prognosis in nasopharyngeal carcinoma patients in an independent external large-sample population and to establish a universally applicable prognostic assessment tool, this study conducted three complementary validation analyses using the SEER database.

Multivariate Logistic regression of metastasis risk factors. A total of 5,347 patients with complete staging and clinicopathological information were included; distant metastasis status served as the dependent variable, and age, sex, T stage, N stage, tumor size, and primary site served as independent variables in multivariate Logistic regression. Results are shown in **Table 10** and **Figure 8A**; the model AIC was 3,257.61. N stage was the strongest clinical predictor of distant metastasis, showing a clear dose-response relationship: the OR for N3 stage relative to N0 stage was 3.514 (95% CI: 2.521–4.899, $P=1.24\times10^{-13}$), for N2 stage was 2.290 (95% CI: 1.706–3.075, $P=3.53\times10^{-8}$), and for N1 stage was 1.638 (95% CI: 1.214–2.210, $P=1.24\times10^{-3}$). Each standard deviation increase in age was associated with a 13.1% increase in metastasis risk (OR=1.131, $P=1.38\times10^{-2}$); each standard deviation increase in tumor size was associated with a 12.0% increase in metastasis risk (OR=1.120, $P=1.57\times10^{-2}$); and male sex was associated with a 26.9% increase in metastasis risk relative to female sex (OR=1.269, $P=3.06\times10^{-2}$). These findings are highly consistent with the baseline analysis results in this study's development cohort, supporting the universality of N stage, age, and sex as core determinants of metastasis risk.

Survival analysis. A total of 7,579 patients with complete survival information were included; comparisons of overall survival (OS) and cancer-specific survival (CSS) between the distant metastasis and non-metastasis groups are presented in **Table 11** and **Figure 8B-C**. The median OS in the M1 group was approximately 18 months, whereas the median OS in the M0 group was approximately 72 months — a difference of nearly 4-fold (Log-rank $\chi^2=695.63$, $P<0.001$). Cancer-specific survival analysis showed a similar pattern (Log-rank $\chi^2=829.30$, $P<0.001$), demonstrating from a prognostic perspective the clinical importance of accurately identifying high-risk metastatic populations.

Subgroup analysis. Cox regression was performed in 24 prespecified subgroups; results are shown in **Table 12** and **Figure 8D**. M1 HR was greater than 1 in all 24 subgroups, and P<0.05 in 23/24 subgroups. The prognostic effect of metastasis was strongest in the Asian population (HR=4.201, 95% CI: 3.612–4.885); metastasis also showed a strong prognostic impact in T1-stage patients (HR=4.659) and N1-stage patients (HR=4.426).

Nomogram construction. The nomogram based on the SEER database (**Figure 8E-F**) achieved a C-index of 0.752, with Hosmer-Lemeshow test P=0.42, indicating good goodness-of-fit. Total scores of 0–280 points correspond to distant metastasis probabilities of 0–80%, providing a risk assessment tool for clinical scenarios lacking laboratory testing resources.

3.9 Transcriptomic Pathway Analysis Results

To characterize the molecular features associated with metastasis, we first performed GSEA on the tumor tissue dataset (GSE103611) using metastasis status as the grouping variable. Results demonstrated extensive and significant transcriptional reprogramming in the metastasis group (**Figure 9A**). At an FDR < 0.1 threshold, Hallmark pathways were all positively enriched, and their biological functions could be summarized into three core themes:

Metabolic reprogramming was the most prominent feature, involving coordinated activation of multiple pathways including xenobiotic metabolism, fatty acid metabolism, bile acid metabolism, cholesterol homeostasis, and oxidative phosphorylation. Microenvironment remodeling pathways were also significantly enriched, encompassing inflammatory response, complement activation, coagulation cascades, apical junctions, angiogenesis, and epithelial-mesenchymal transition. Meanwhile, proliferation-driving signals were highly active, including cell cycle-related pathways such as G2M checkpoint, E2F targets, and MYC targets V1/V2 (**Figure 9B**).

These results collectively indicate that metastatic tumor tissues simultaneously activate a triple program of metabolic adaptation, invasive microenvironment remodeling, and rapid proliferation. In contrast, no pathways in the PBMC dataset (GSE299775) reached the significance threshold (all FDR > 0.8).

We further employed ssGSEA to score pathways for each individual sample, enabling more sensitive detection of functional differences between the M and NM groups. **Figure 9C** illustrates the overall distributional shift of pathway activity in PBMC. Results showed that the metastatic group PBMC exhibited a dysregulation pattern dominated by proliferation suppression and impaired immune response (**Figure 9D**). Cell cycle pathways including MYC targets, E2F targets, and G2M checkpoint were all significantly downregulated in the metastatic group, suggesting that the proliferative capacity of peripheral blood lymphocytes is systemically suppressed. Concurrently, innate immune pathways such as interferon response (α/γ) and TNFα/NF-κB also showed consistent downward trends. Notably, KRAS signaling downregulation and iron/erythroid differentiation features were inversely elevated in the metastatic group, implying potential shifts in peripheral blood cell lineage composition.

In tumor tissue, ssGSEA identified no significantly differential pathways at FDR < 0.1; only glycolysis and hypoxia were nominally significant (P < 0.05), but their FDR values were both above 0.3. This contrast suggests that metastasis-related differences in tumor tissue are primarily driven by individual gene expression changes rather than systematic remodeling of overall pathway activity.

To explore whether the above pathway dysregulations are associated with clinically available serological indicators, we stratified patients by the median of each predictor and performed GSEA in tumor tissue and PBMC respectively. **Figure 9E** presents a panoramic association matrix between pathway normalized enrichment scores (NES) and each predictor, and **Table 13** summarizes all statistically significant (P < 0.05 or FDR < 0.1) pairwise relationships.

Overall, different predictors and pathways exhibited a highly specific pairwise pattern, which can be summarized into the following three biological themes:

The metabolism-related module: LDH was uniquely coupled with glycolysis and hypoxia pathways; the AST/ALT ratio was specifically associated with xenobiotic metabolism and fatty acid metabolism; and cholesterol homeostasis was exclusively related to the cholesterol metabolism pathway itself. This module suggests that different serological metabolic indicators reflect distinct facets of metabolic adaptation strategies in tumor cells.

The proliferation and tumor burden module: CA125 strictly tracked cell cycle pathways such as E2F targets and G2M checkpoint, whereas MYC targets (V1/V2) were significantly associated with metastasis status rather than any single serological indicator. This indicates that MYC-driven proliferative signals may be an inherent feature of metastasis, rather than indirectly reflected through clinical indicators such as CA125.

The immune and microenvironment remodeling module: IgA was specifically associated with inflammatory response and IL6/JAK/STAT3 signaling; β2-microglobulin corresponded to interferon-γ and TNFα/NF-κB pathways; D-dimer was uniquely coupled with complement activation and coagulation cascades. RDW-SD was specifically associated with erythroid/heme pathways (iron/erythroid features and heme metabolism). These pairwise relationships reveal that different immune-inflammatory mediators point to distinctly different microenvironment subtypes, rather than generalized immune activation.

Notably, all the above significant associations existed exclusively in tumor tissue; no pathways in PBMC reached significant correlation with clinical indicators (all FDR > 0.8), further highlighting the deep decoupling between local microenvironment and systemic circulation in molecular response.

To visually validate these findings, we selected the most representative and most dramatically contrasting pathway, MYC_TARGETS_V2, for inter-group comparison. **Figure 9F** shows that this pathway was significantly decreased in the PBMC metastatic group, whereas it showed the opposite upward trend in the tumor tissue metastatic group. This systemic versus local decoupling pattern was also present in proliferation-related pathways such as G2M_CHECKPOINT and E2F_TARGETS, suggesting that the proliferative suppression of peripheral blood lymphocytes may serve as a concomitant systemic feature of metastasis.

4. Discussion

In this study, we developed and validated a machine learning prediction model based on routine laboratory parameters for individualized assessment of distant metastasis risk in patients with newly diagnosed nasopharyngeal carcinoma. The core XGBoost-combined model

integrated five categories of laboratory indicators—routine blood tests, biochemistry, coagulation, tumor markers, and immunoglobulins—and achieved an AUROC of 0.8595 in internal validation (MCCV: 0.8470, 95% CI: 0.8032–0.8871) with a Brier score of 0.1287, indicating favorable discriminative ability coupled with satisfactory probability calibration. Risk stratification revealed a distant metastasis rate of merely 7.3% in the low-risk group versus 56.5% in the high-risk group—a nearly eightfold intergroup difference that established a clear and actionable clinical risk gradient. Furthermore, decision curve analysis demonstrated net benefit superior to the "treat-all" strategy within the threshold range of 20%–60%, confirming its clinical decision-support value. At the model interpretability level, SHAP analysis identified RDW-SD, IgA, CA125, D-dimer, and LDH as the top five predictive factors, collectively contributing 38.2% of the total SHAP value across all features. These features correspond to erythrocyte volume heterogeneity, humoral immune activation, tumor burden, coagulation dysfunction, and metabolic reprogramming, respectively, suggesting that the biological foundation of distant metastasis is not dominated by a single pathway but rather represents a comprehensive manifestation of multi-system collaborative dysregulation. This finding cross-validates the identification of nine independent predictors in multivariable Logistic regression, among which seven concurrently appeared in the SHAP Top 20 features; the cross-methodological consistency substantially strengthened the evidence for these biomarkers. In the external validation dimension, SEER database analysis (N=5,347) confirmed N stage as the strongest clinical predictor of distant metastasis (N3 vs. N0: OR=3.51), and survival analysis (N=7,579) demonstrated that the median OS of M1-stage patients was approximately one-quarter that of M0-stage patients (18 months vs. 72 months), indirectly substantiating from a prognostic perspective the urgency of accurately identifying high-risk populations for metastasis. At the transcriptomic level, GSEA of two independent GEO cohorts demonstrated nominal significance of the GLYCOLYSIS pathway in the metastatic group (P=0.025 and P=0.033), providing independent molecular corroboration for the predictive value of LDH in the model. This multi-layered evidence chain—spanning "clinical phenotype–machine learning discrimination–external cohort validation–transcriptomic corroboration"—provides relatively comprehensive support for the clinical translation and biological plausibility of the model.

SHAP analysis categorized the top 20 predictive factors into five classes: erythrocyte homeostasis, humoral immunity, tumor burden, coagulation function, and energy metabolism.

These features are not isolated “islands” operating independently; rather, they collectively delineate a systemic host-tumor interaction landscape of distant metastasis in nasopharyngeal carcinoma.

Erythrocyte homeostasis encompasses predictive factors including RDW-SD, hemoglobin, ferritin, and eosinophil percentage, among which RDW-SD ranked first among all features and retained independent predictive effect in multivariable regression (OR=1.34). Elevated RDW reflects increased erythrocyte volume heterogeneity and is generally regarded in the oncological context as a peripheral blood hallmark of anemia of chronic disease (ACD) [16]. The core mechanism of ACD involves inflammatory cytokines—particularly interleukin-6 (IL-6)—inducing hepatic synthesis of hepcidin, which binds to and degrades ferroportin on the surface of macrophages and intestinal epithelial cells, leading to iron “sequestration” within the reticuloendothelial system that renders it unavailable for erythropoiesis. This mechanism has been confirmed in various solid tumors such as colon cancer [17], and in the present study, the elevated ferritin (625.7 vs. 444.4 ng/mL) combined with decreased hemoglobin (105.0 vs. 120.0 g/L) observed in the M1 group represents a typical manifestation of dysregulation along this hepcidin-ferroportin axis.

A study by Sun Yat-sen University Cancer Center based on 2,318 non-metastatic NPC patients confirmed that the COR-BMI score constructed using RDW and BMI could independently predict overall survival (HR=2.367 when COR-BMI=2), and also demonstrated prognostic stratification value in the metastatic NPC subgroup. The hemoglobin-to-red cell distribution width ratio (HRR) has likewise been validated as an independent prognostic biomarker for overall survival and disease-free survival in locoregionally advanced NPC after chemoradiotherapy [18]. Ferritin, as an acute-phase reactant, reflects inflammation-driven iron sequestration; a study enrolling 252 NPC patients demonstrated that the high-ferritin group exhibited significantly shorter overall survival and progression-free survival. The bioinformatic analysis in this study provided independent corroboration for this mechanism: the HALLMARK_HEME_METABOLISM pathway showed nominal enrichment in the high-RDW-SD group (P=0.055), suggesting an association at the transcriptional level between abnormal activation of erythroid metabolic pathways and RDW elevation. A bidirectional causal relationship may exist between elevated RDW and tumor metastasis: on one hand, tumor-

induced chronic inflammation causes erythropoietic impairment through the hepcidin pathway; on the other hand, iron metabolic dysregulation itself can promote tumor progression through multiple mechanisms—iron serves as a cofactor for ribonucleotide reductase, rate-limiting for DNA synthesis; iron overload can catalyze Fenton reactions to generate free radicals, inducing genomic instability.

Humoral immune response comprises IgA, β_2-microglobulin (β2-MG), and complement C4. The distinctiveness of this finding is rooted in the special relationship between nasopharyngeal carcinoma and Epstein-Barr virus (EBV); following EBV infection of nasopharyngeal epithelial cells, lytic-phase antigens can induce a robust IgA-type humoral immune response.

A meta-analysis by Liang et al. (2021) enrolling 5,729 patients confirmed that EA-IgA could significantly predict overall survival (HR=1.63, 95% CI: 1.07–2.48), whereas VCA-IgA lacked this effect [19]. This suggests that IgA induced by EBV lytic-phase antigens (rather than IgA induced by latent-phase antigens) possesses genuine prognostic discriminative value. The present study measured total IgA rather than EBV-specific antibodies, and its predictive value may originate from a comprehensive reflection of EBV reactivation intensity combined with persistent tumor antigen stimulation. β2-MG, as a component of the MHC-I light chain, reflects elevated lymphocyte turnover rate and tumor immune surveillance intensity [20]; complement C4 participates in classical pathway activation, directly reflecting the intensity of humoral immune response triggered by EBV-immune complexes [21]. The GEO ssGSEA analysis in this study demonstrated nominal elevation of the CUSTOM_IMMUNOGLOBULIN pathway in PBMCs from the metastatic group (P=0.042), providing independent transcriptomic corroboration for the association between IgA and metastasis.

The coagulation-inflammation axis and pre-metastatic niche encompass predictive factors including D-dimer and FDP. A retrospective study by Chen et al. of 717 nasopharyngeal carcinoma patients confirmed that, using a cutoff of 0.8 μg/L, the high-D-dimer group exhibited significantly inferior 3-year distant metastasis-free survival compared with the low-D-dimer group, and multivariable analysis identified D-dimer as an independent predictor of disease-free survival, distant metastasis-free survival, and overall survival. This study further demonstrated that D-dimer could effectively stratify patient prognosis across subgroups of early-stage, advanced-stage, and EBV DNA ≥ 4000 copies/ml patients [22]. A meta-analysis by Liang et

al. (encompassing 9 studies and 5,729 patients) also confirmed that D-dimer significantly predicted overall survival (HR = 1.75, 95% CI 1.24–2.47) and distant metastasis-free survival (HR = 1.91, 95% CI 1.31–2.79) [23].

At the mechanistic level, tumor cells activate the coagulation cascade through tissue factor (TF) and cancer procoagulant [24], and fibrin deposition in the vascular bed of target organs provides a physical scaffold for the capture and colonization of circulating tumor cells—constituting the construction of a "pre-metastatic niche" [25]. Thrombin activates tumor cells and platelets via protease-activated receptor-1 (PAR-1), releasing pro-angiogenic and pro-EMT factors such as vascular endothelial growth factor (VEGF) and transforming growth factor-beta (TGF-β) [26]. Our team's GEO analysis revealed nominal enrichment of HALLMARK_COAGULATION in the high-D-dimer group (P=0.05), providing direct transcriptomic corroboration for this "coagulation-metastasis" hypothesis.

Notably, D-dimer contributed prominently in SHAP (0.284) but did not achieve independent significance in multivariable Logistic regression (P=0.078). This seemingly paradoxical phenomenon precisely reveals the advantage of machine learning over traditional regression models—the predictive information of D-dimer may not function as an independent linear main effect but rather contributes to classification performance through synergistic interactions with other inflammatory or coagulation indicators. Relying solely on traditional regression models may underestimate the actual clinical value of coagulation parameters in risk stratification.

Regarding hepatic functional load and metabolic reprogramming, LDH (SHAP=0.268) approached significance in multivariable analysis (OR=1.356, P=0.077). A machine learning study by Sun et al. based on 1,845 nasopharyngeal carcinoma patients confirmed LDH as an important risk factor for distant metastasis [27]. A meta-analysis by Zhang et al. (16 studies, 14,803 patients) also verified that elevated LDH significantly predicted poorer overall survival (HR=1.79) and distant metastasis-free survival (HR=1.85) [28]. As a key enzyme of glycolysis, elevated serum LDH directly reflects the intensity of the Warburg effect; EBV-LMP1 can enhance glycolysis in nasopharyngeal carcinoma cells by altering mitochondrial function. GSEA demonstrated nominal enrichment of the GLYCOLYSIS pathway in the GSE103611 metastatic group (P=0.025), and the high-LDH group showed significant enrichment of glycolysis and

hypoxia pathways in GSEA (NES=1.160, FDR=0.074), forming a complete logical loop of "elevated serum LDH → glycolysis pathway activation → distant metastasis."

Albumin (SHAP=0.221) reduction reflects impaired hepatic synthetic function and nutritional consumption. A retrospective study by Chen et al. enrolling 622 nasopharyngeal carcinoma patients confirmed that high N stage, elevated ferritin (>300 μg/L) combined with hypoalbuminemia (<42 g/L) could effectively predict distant metastasis after intensity-modulated radiotherapy (high-risk group 5-year DMFS 76.4% vs. low-risk group 89.6%) [29]. Elevated AST/ALT ratio (SHAP=0.204) and AST (SHAP=0.159) suggest hepatocellular injury or tumor infiltration—Wu et al. confirmed in 1,023 non-metastatic nasopharyngeal carcinoma patients that a high De Ritis ratio (≥1.65) was an independent predictor of CSS (SHR=1.64), PFS (SHR=1.69), and OS (HR=1.81) [30]. Indirect bilirubin (SHAP=0.098) emerged as one of the strongest protective factors in multivariable regression (OR=0.377, $P=1.85\times10^{-3}$); the IBI score constructed based on indirect bilirubin and inflammatory markers has been validated as a potential prognostic factor for nasopharyngeal carcinoma patients [31]. The aforementioned hepatic function parameters were specifically associated with XENOBIOTIC_METABOLISM (NES=1.324) and FATTY_ACID_METABOLISM (NES=1.220) in GEO pathway analysis, forming a complete evidence chain from serum hepatic function parameters to tumor cell metabolic adaptation.

Tumor proliferation burden was most prominently represented by CA125 (SHAP=0.301) and CA19-9 (SHAP=0.148). CA125 is encoded by the MUC16 gene and is significantly overexpressed in nasopharyngeal carcinoma tissue, where it correlates with advanced stage. Mechanistically, MUC16 plays critical roles in cell adhesion, signal transduction, and immune response; it can protect tumor cells from anoikis via phosphoinositide 3-kinase-protein kinase B (PI3K-AKT) signaling and promote immune evasion through interactions with immune cells in the tumor microenvironment. Recent research has further revealed that the transcription factor ELF3 promotes MUC16 transcription by binding to its promoter, and MUC16 overexpression accelerates nasopharyngeal carcinoma tumor growth through glycolysis-mediated immune evasion [32]. GEO analysis demonstrated that the high-CA125 group was significantly enriched for E2F_TARGETS (NES=1.089, FDR=0.074) and G2M_CHECKPOINT (NES=1.069,

FDR=0.074), suggesting that the predictive value of CA125 primarily reflects tumor cell proliferation burden rather than its traditional role as an organ-specific metastasis marker.

It is noteworthy that the predictive performance of the traditional serum tumor marker panel alone was limited (AUROC=0.8003), substantially inferior to the combined contribution of the inflammation/metabolism module. The reasons for this phenomenon are as follows: nasopharyngeal carcinoma is universally associated with EBV, and EBV DNA testing far surpasses traditional protein tumor markers in both sensitivity and specificity [33]; moreover, inflammation-related indicators (such as PLR, AMC, etc.) are obtainable from routine blood tests and provide prognostic information independent of traditional tumor markers. The signals of distant metastasis are encoded not only in protein markers secreted by tumor cells themselves but also—more broadly—in the systemic host response to tumor [34].

The value of this study is manifested at two levels. First, at the clinical level, the model is informationally complementary to TNM staging. Logistic regression in this study showed that among traditional TNM staging variables, only N stage (particularly N3 stage, OR=3.51) exhibited strong independent predictive value, whereas the incremental contribution of T stage was relatively limited. This finding is consistent with the inherent characteristics of the TNM staging system, which primarily describes the anatomical extent of tumor, whereas laboratory parameters capture the systemic biological state of the tumor [35]. In recent years, multiple studies have confirmed that integrating multimodal data—such as deep learning risk scores [36] or genetically optimized neural networks [37]—can surpass the predictive performance of TNM staging used alone. The two dimensions of information are fundamentally complementary: TNM staging describes an anatomical prognostic framework, whereas biological signals such as immune-inflammatory and metabolic markers represent dimensions that TNM staging cannot cover. The core value of this complementarity lies in adding biological dimension incremental information to TNM staging, driving a paradigm evolution from “anatomical classification” toward “anatomical-biological integrated classification.”

Second, at the methodological level, this study constructed a hierarchical evidence framework progressing from statistical validation to mechanistic validation. Internally, 5-fold cross-validation and MCCV ensured model robustness; SEER external validation corroborated the clinical relevance of the model construction endpoint (i.e., distant metastasis status) from a

prognostic perspective; GEO transcriptomic analysis provided independent corroboration at the molecular level for the biological plausibility of predictive factors. Taking RDW-SD as an example, the complete "prediction-validation-mechanism" closed loop was formed through: SHAP identification → clinical dose-response relationship → hepcidin-iron metabolism ACD hypothesis → heme metabolism pathway enrichment trend in GEO. The coagulation activation hypothesis for D-dimer and the EBV immune response hypothesis for IgA follow the same logical chain. Meanwhile, comprehensive data leakage prevention throughout the workflow and consistent results from four sensitivity analyses further confirmed the robustness of the findings. This "from surface to mechanism" strategy elevates the research value beyond mere predictive tool development, contributing testable scientific hypotheses to the pathophysiology of nasopharyngeal carcinoma metastasis.

This study has several limitations. First, the retrospective single-center design may introduce selection bias, as patients who completed the full panel of laboratory tests at initial diagnosis may represent a subgroup with more pronounced symptoms, and variation in the timing of these tests relative to diagnosis may also affect the authenticity of baseline values. Second, the outcome definition was based on M status at initial diagnosis and did not incorporate metachronous metastasis cases that emerged during treatment or follow-up; although this ensures the temporal logical sequence between predictors and outcome, it limits the clinical application scenario of the model in predicting "future metastasis risk." Third, although SEER and GEO database validations provided external evidence, a true prospective multicenter external validation is still lacking. Furthermore, the model performs static prediction based on single-timepoint baseline laboratory data and does not leverage dynamic changes in indicators such as RDW and D-dimer during the treatment course. Future research will focus on prospective multicenter cohort external validation, development of dynamic prediction models based on longitudinal data, and periodic model updating and recalibration along with evolving treatment modalities.

Based on the risk stratification capability of the model, we propose the following translational directions. At the screening level, for patients with model-predicted high risk (>40%), we recommend adding positron emission tomography-computed tomography (PET-CT) or whole-body contrast-enhanced CT to screen for occult metastasis on the basis of routine TNM

evaluation, alongside EBV DNA load and circulating tumor DNA (ctDNA) testing, to achieve precise imaging resource allocation for high-risk populations. At the treatment decision level, the combination of the model with TNM staging can guide stratified and refined treatment: for locally advanced patients with high model risk, intensified induction chemotherapy, anticoagulation prophylaxis, or more intensive surveillance protocols may be considered; for early-stage patients with high model risk, restaging accuracy should be reassessed and treatment escalation considered. This “anatomical staging + biological scoring” dual-track model holds promise for optimizing precision treatment decision-making in nasopharyngeal carcinoma.

At the translational research level, the mechanistic explorations in this study suggest three potential intervention directions: targeting iron metabolism (such as hepcidin antagonists or ferroptosis inducers) for anti-metastatic effects in patients with high RDW and high ferritin; randomized controlled validation of anticoagulation adjuvant therapy in patients with high D-dimer; and interferon response pathway-based screening for immunotherapy-advantaged populations. The above directions are all grounded in the findings of this study and possess clear clinical translation pathways.

In conclusion, the machine learning model constructed in this study demonstrated favorable discriminative performance and robust generalizability in predicting distant metastasis in nasopharyngeal carcinoma. More importantly, through multi-level validation spanning clinical prediction to molecular mechanisms, this study not only provides a potential clinical decision-support tool but also contributes novel scientific insights to the pathophysiological study of nasopharyngeal carcinoma metastasis—including the ACD iron metabolism hypothesis, the PIGR-IgA mucosal immune axis hypothesis, and the coagulation-inflammation synergy hypothesis—laying the foundation for individualized screening of high-risk patients, optimization of treatment decision-making, and exploration of novel therapeutic targets.

Competing interests

The authors declare that they have no competing interests.

Funding

2023 Guangxi Project for the Development, Promotion and Application of Appropriate Medical and Health Technologies(Contract Number: S2023123); Research Fund Project in the Field of Oncology of Life Oasis Beijing Public Welfare Service Center(Project Number: BH004506); Guangxi university young and middle-aged teachers' basic ability improvement project (grant no. 2025KY0331), Guangxi Appropriate Health Technologies Development and Promotion Project (S2024115).

Authors' contributions

Lili Liang and Zhengyi Xie: Figure creation, integration of the mechanism diagram, and drafting of the initial manuscript; Yingxi Zhao:Writing - review & editing; Xiangni Luo: Writing - review & editing; Zhengyi Xie and Jin Wang: Conceptualization, Funding acquisition, Conceptualization, Writing - review & editing. All authors contributed to the article and have read and approved the final version.

Ethics approval and consent to participate

The study protocol was approved by the Ethics Review Committee of Guangxi University of Science and Technology (approval number: 2025-02). All procedures were performed in accordance with the ethical standards of the Declaration of Helsinki. Because this was a retrospective study using de-identified, anonymized clinical data collected during routine practice, the ethics committee waived the requirement for informed consent from participants.

Consent for publication

Not Applicable. This manuscript does not contain any identifying images or other personal or clinical details of participants that could compromise anonymity.

Data availability

The clinical data used in this study are not publicly available due to privacy and ethical constraints, but may be requested from the corresponding author upon reasonable justification.

Reference

[1] Chan SM, Paterson IC and Yap LF (2025) Nasopharyngeal Carcinoma in Southeast Asia: Current Landscape and Future Priorities. Br. J. Biomed. Sci. 82:15902. doi: 10.3389/bjbs.2025.15902

[2] Zhang Y, Gu S, Deng H, Shen Z. Global epidemiological profile in nasopharyngeal carcinoma: a prediction study. BMJ Open. 2024 Dec;14(12):e091087. DOI: 10.1136/bmjopen-2024-091087. PMID: 39658299; PMCID: PMC11647315.

[3] Mahdavifar N, Ghoncheh M, Mohammadian-Hafshejani A, Khosravi B, Salehiniya H. Epidemiology and Inequality in the Incidence and Mortality of Nasopharynx Cancer in Asia. Osong Public Health Res Perspect. 2016 Dec;7(6):360-372. doi: 10.1016/j.phrp.2016.11.002. Epub 2016 Nov 16. PMID: 28053841; PMCID: PMC5194228.

[4] Chen W , Sun K , Zheng R .Cancer incidence and mortality in China, 2014[J].Cancer incidence and mortality in China,2018,30(1):1-12.DOI:10.21147/j.issn.1000-9604.2018.01.01.

[5] Niu X, Xue F, Liu P, Hu C, He X. Long-term outcomes of nasopharyngeal carcinoma patients with T1-2 stage in intensity-modulated radiotherapy era. Int J Med Sci. 2022 Jan 1;19(2):267-273. doi: 10.7150/ijms.68394. PMID: 35165512; PMCID: PMC8795811.

[6] Chen Y, Chan A, Le Q et al.Nasopharyngeal carcinoma.The Lancet, 2019; 394, 64-80

[7] Pan JJ, Mai HQ, Ng WT et al.Ninth Version of the AJCC and UICC Nasopharyngeal Cancer TNM Staging Classification. JAMA Oncol. 2024 Oct 10;10(12):1627-35. doi: 10.1001/jamaoncol.2024.4354. Epub ahead of print. Erratum in: JAMA Oncol. 2024 Dec 1;10(12):1736. doi: 10.1001/jamaoncol.2024.6231. PMID: 39388190; PMCID: PMC11581663.

[8] Chen, Y. P., Ismaila, N., Chua, M et al.(2021). Chemotherapy in Combination with Radiotherapy for Definitive-Intent Treatment of Stage II-IVA Nasopharyngeal Carcinoma: CSCO and ASCO Guideline. Journal of Clinical Oncology, 39(7), 840-859. https://doi.org/10.1200/JCO.20.03237

[9] Liu H, Yang P, Jia Y. Optimal management of oligometastatic nasopharyngeal carcinoma. European Archives of Oto-rhino-laryngology. 2022 Feb;279(2):567-576. DOI: 10.1007/s00405-021-06918-y. PMID: 34089384.

[10] 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 Jul 12;13(7):1706. doi: 10.3390/biomedicines13071706. PMID: 40722777; PMCID: PMC12292804.

[11] Shen YQ, Wei QW, Tian YR, Ling YZ, Zhang M. Coagulation indices and fibrinogen degradation products as predictive biomarkers for tumor-node-metastasis staging and metastasis in gastric cancer. World J Gastrointest Oncol. 2025 Jan 15;17(1):98725. doi: 10.4251/wjgo.v17.i1.98725. PMID: 39817145; PMCID: PMC11664628.

[12] Hu Y, Zhang X, Slavin V, Belsti Y, Tiruneh SA, Callander E, Enticott J. Beyond Comparing Machine Learning and Logistic Regression in Clinical Prediction Modelling: Shifting from Model Debate to Data Quality. J Med Internet Res. 2025 Nov 5;27:e77721. doi: 10.2196/77721. PMID: 41191908; PMCID: PMC12631085.

[13] Alabi RO, Elmusrati M, Leivo I, Almangush A, Mäkitie AA. Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. Sci Rep. 2023 Jun 2;13(1):8984. doi: 10.1038/s41598-023-35795-0. PMID: 37268685; PMCID: PMC10238539.

[14] Jiang T, Tan Y, Nan S, Wang F, Chen W, Wei Y, Liu T, Qin W, Lu F, Jiang F, Jiang H. Radiomics based on pretreatment MRI for predicting distant metastasis of nasopharyngeal carcinoma: A preliminary study. Front Oncol. 2022 Aug 9;12:975881. doi: 10.3389/fonc.2022.975881. PMID: 36016603; PMCID: PMC9396739.

[15] Si J, Ding X, Deng Z, Li P, Zhang B, Lan G, Huang B, Liang J, Wang Z, Si Y. A Four-Gene Signature Model Improves the Prediction of Distant Metastasis in Patients with Nasopharyngeal Carcinoma: A Retrospective, Three-Center Observational Study. Technol Cancer Res Treat. 2022 Jan-Dec;21:15330338221080972. doi: 10.1177/15330338221080972. PMID: 35262435; PMCID: PMC8918749.

[16] Fraenkel PG. Anemia of Inflammation: A Review. Med Clin North Am. 2017 Mar;101(2):285-296. doi: 10.1016/j.mcna.2016.09.005. Epub 2016 Dec 24. PMID: 28189171; PMCID: PMC5308549.

[17] Rolić T, Mandić S, Yazdani M, Ferenac Kiš M, Distante S, Banjari I. Serum hepcidin concentration is lower in advanced stages of sporadic colorectal cancer. Biochem Med (Zagreb). 2025 Oct 15;35(3):030703. doi: 10.11613/BM.2025.030703. PMID: 41103686; PMCID: PMC12523599.

[18] Wang Y, He SS, Cai XY, Chen HY, Yang XL, Lu LX, Chen Y. The Novel Prognostic Score Combining Red Blood Cell Distribution Width and Body Mass Index (COR-BMI) Has Prognostic Impact for Survival Outcomes in Nasopharyngeal Carcinoma. J Cancer. 2018 Jun 6;9(13):2295-2301. doi: 10.7150/jca.24838. PMID: 30026824; PMCID: PMC6036710.

[19] Liang T, Liu W, Xie J, Wang Y, Chen G, Liao W, Song L, Zhang X. Serum EA-IgA and D-dimer, but not VCA-IgA, are associated with prognosis in patients with nasopharyngeal carcinoma: a meta-analysis. Cancer Cell Int. 2021 Jun 30;21(1):329. doi: 10.1186/s12935-021-02035-2. PMID: 34193149; PMCID: PMC8243822.

[20] Lin Q, Fu F, Zhang G. [Detection of bate 2-microglobulin fibronection and interleukin 2 content in patients with nasopharyngeal carcinoma]. Lin Chuang Er Bi Yan Hou Ke Za Zhi. 2001 Nov;15(11):499-500. Chinese. PMID: 12541829.

[21] Lim M, Chan SH. C4 and 21-hydroxylase gene deletions in nasopharyngeal carcinoma among the Chinese. Ann Acad Med Singap. 1996 Jan;25(1):42-4. PMID: 8779545.

[22] Chen WH, Tang LQ, Wang FW, Li CP, Tian XP, Huang XX, Mai SJ, Liao YJ, Deng HX,

Chen QY, Liu H, Zhang L, Guo SS, Liu LT, Yan SM, Li CF, Zhang JP, Liu Q, Liu XW, Liu LZ, Mai HQ, Zeng MS, Xie D. Elevated levels of plasma D-dimer predict a worse outcome in patients with nasopharyngeal carcinoma. BMC Cancer. 2014 Aug 10;14:583. doi: 10.1186/1471-2407-14-583. PMID: 25109220; PMCID: PMC4242497.

[23] Liang T, Liu W, Xie J, Wang Y, Chen G, Liao W, Song L, Zhang X. Serum EA-IgA and D-dimer, but not VCA-IgA, are associated with prognosis in patients with nasopharyngeal carcinoma: a meta-analysis. Cancer Cell Int. 2021 Jun 30;21(1):329. doi: 10.1186/s12935-021-02035-2. PMID: 34193149; PMCID: PMC8243822.

[24] Galardi A, Dell'Orto E, Bianchi F, Cosentino G, Fogazzi V, Pupa SM, Iorio MV. Hemostasis at the edge between physiology and cancer. Intern Emerg Med. 2026 Jun;21(4):1169-1183. doi: 10.1007/s11739-026-04294-1. Epub 2026 May 5. PMID: 42086829; PMCID: PMC13263211.）

[25] Yuxin Zhang, Zelin Li, Jiamao Zhang, Tatenda Mafa, Jingyu Zhang, Hui Zhu, Lifang Chen, Zhen Zong, Lingling Yang,Fibrinogen: A new player and target on the formation of pre-metastatic niche in tumor metastasis,Critical Reviews in Oncology/Hematology,Volume 207,2025,104625,ISSN 1040-8428,https://doi.org/10.1016/j.critrevonc.2025.104625.

[26] Aleksandrowicz, K., Wojtukiewicz, A., Hempel, D., Polityńska, B., Wojtukiewicz, M., Tang, D., & Honn, K. (2024). The Complex Role of Thrombin in Cancer and Metastasis: Focus on Interactions with the Immune System. Seminars in Thrombosis & Hemostasis, 50(3), 462–473. https://doi.org/10.1055/s-0043-1776875

[27] Sun H, Zhu J, Li L, Xin X, Yan J and Huang T (2025) Machine learning for predicting distant metastasis in nasopharyngeal carcinoma patients. Front. Immunol. 16:1580200. doi: 10.3389/fimmu.2025.1580200

[28] Zhang M, Wei S, Su L, Lv W, Hong J. Prognostic significance of pretreated serum lactate dehydrogenase level in nasopharyngeal carcinoma among Chinese population: A meta-analysis. Medicine (Baltimore). 2016 Aug;95(35):e4494. doi: 10.1097/MD.0000000000004494. PMID: 27583859; PMCID: PMC5008543.

[29] Chen X, Long X, Liang Z, Lei H, Li L, Qu S, Zhu X. Higher N stage and serum ferritin, but lower serum albumin levels are associated with distant metastasis and poor survival in patients with nasopharyngeal carcinoma following intensity-modulated radiotherapy. Oncotarget. 2017 Apr 25;8(42):73177-73186. doi: 10.18632/oncotarget.17418. PMID: 29069861; PMCID: PMC5641204.

[30] Wu J,Li S,Wang Y,Hu L.Pretreatment Aspartate Aminotransferase to-Alanine Aminotransferase (De Ritis) Ratio Predicts the Prognosis of Nonmetastatic Nasopharyngeal Carcinoma. Onco Targets Ther. 2019 Nov 22;12:10077-10087. doi: 10.2147/OTT.S232563. PMID: 31819502; PMCID: PMC6878916.

[31] JI Huojin, LI Jun, LUO Yonglin, QIN Weiling, YE Yinxin, CAI Yonglin. The prognostic predictive value of indirect bilirubin-inflammation score in patients with nasopharyngeal carcinoma[J]. Journal of Guangxi Medical University, 2024, 41(9): 1273-1281. DOI: 10.16190/j.cnki.45-1211/r.2024.09.006

[32] Liu Y, Zhou H, Yu Q, Wang Q. Hypomethylation-associated ELF3 helps nasopharyngeal carcinoma to escape immune surveillance via MUC16-mediated glycolytic metabolic reprogramming. Am J Physiol Cell Physiol. 2024 Oct 1;327(4):C1125-C1142. doi: 10.1152/ajpcell.00438.2024. Epub 2024 Sep 2. PMID: 39219440; PMCID: PMC11481993.

[33] Li, XQ., Lin, DF., Cai, YC. et al. Diagnostic performance of EBV DNA load testing for nasopharyngeal carcinoma in nasopharyngeal swab outperforms the approach in other specimens. BMC Cancer 25, 1126 (2025). https://doi.org/10.1186/s12885-025-14539-5

[34] Tan R, Phua SKA, Soong YL, Oon LLE, Chan KS, Lucky SS, Mong J, Tan MH, Lim CM. Clinical utility of Epstein-Barr virus DNA and other liquid biopsy markers in nasopharyngeal carcinoma. Cancer Commun (Lond). 2020 Nov;40(11):564-585. doi: 10.1002/cac2.12100. Epub 2020 Sep 28. PMID: 32989921; PMCID: PMC7668470.

[35] Wen, M., Wang, G., Du, Z. et al. Recent advances in refinement of the NPC TNM staging system. Holist Integ Oncol 4, 25 (2025). https://doi.org/10.1007/s44178-025-00162-8

[36] Zhou J, Wibawa MS, Wang R, Deng Y, Huang H, Luo Z, Xia Y, Guo X, Young LS, Lo KW, Rajpoot N, Lv X. Multimodal AI-based risk stratification for distant metastasis in nasopharyngeal carcinoma. ESMO Open. 2025 Oct;10(10):105809. doi: 10.1016/j.esmoop.2025.105809. Epub 2025 Sep 24. PMID: 40997749; PMCID: PMC12495093.

[37] Ling LT, Li WJ, Yao Y, Tan KQ, Zhu BL, Zhou LQ, Qu S, Li L, Guan Y, Zhu XD, Pan LH, Liang ZG. Genetic algorithm-optimized neural network outperforms TNM staging in predicting rapidly progressive nasopharyngeal carcinoma: Reassessing adjuvant chemotherapy benefit via propensity score matching. Eur J Cancer. 2025 Nov 17;230:115787. doi: 10.1016/j.ejca.2025.115787. Epub 2025 Sep 12. PMID: 41022023.

Figure1

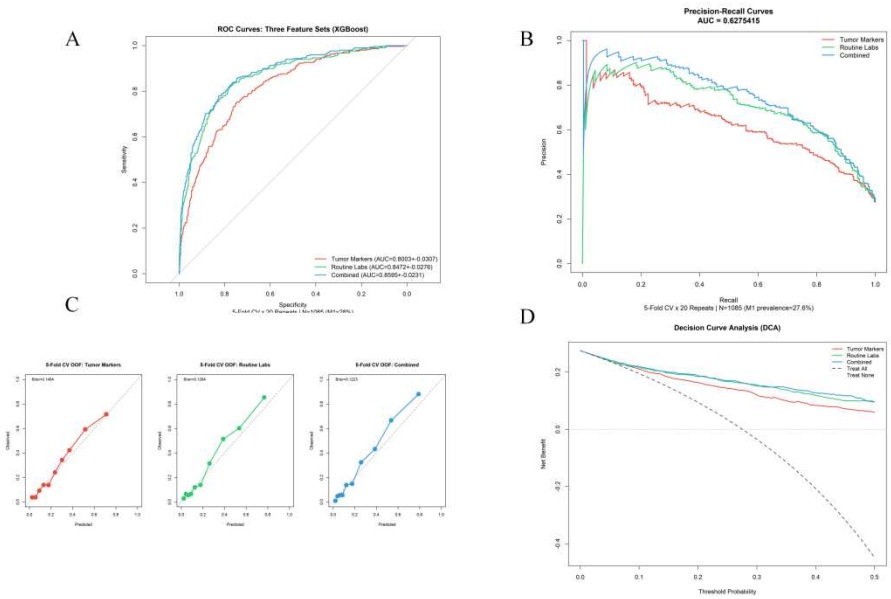


Figure 1 Model Discrimination, Calibration, and Clinical Utility Assessment

(A) Receiver operating characteristic (ROC) curves for the three feature subsets combined with the XGBoost algorithm. Red = Tumor Markers (AUROC = 0.8003 ± 0.0307), Green = Routine Labs (AUROC = 0.8472 ± 0.0276), Blue = Combined (AUROC = 0.8595 ± 0.0231). The gray diagonal line is the reference line.

(B) Precision-recall curve (Precision-Recall), AUPRC = 0.6275. Based on stratified 5-fold cross-validation × 20 repeats (N = 1,085, M1 prevalence = 27.6%).

(C) Calibration curves. Left = Tumor Markers (Brier = 0.1484), Middle = Routine Labs (Brier = 0.1284), Right = Combined (Brier = 0.1223). X-axis = predicted probability, Y-axis = actual observed probability, red solid line = fitted curve, gray dashed line = ideal calibration line (slope = 1).

(D) Decision curve analysis (DCA). Within the threshold probability range of 0%–50%, the Combined model (blue) consistently demonstrated higher net benefit than Tumor Markers (red) and Routine Labs (green), and also exceeded the “treat all” (gray dashed line) and “treat none” (black solid line) strategies.

Figure2

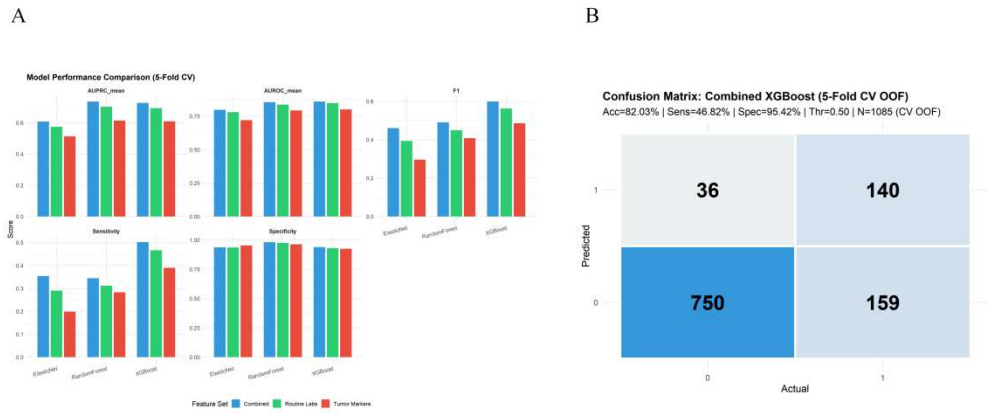


Figure 2 Algorithm-Feature Set Combination Performance Comparison and Optimal Model Confusion Matrix

(A) Multi-dimensional performance bar chart for the nine algorithm-feature set combinations. From left to right: AUPRC, AUROC, sensitivity, specificity, and F1 score. Blue = Combined, Green = Routine Labs, Red = Tumor Markers. XGBoost outperformed Random Forest and ElasticNet across all metrics.

(B) Confusion matrix of the optimal model (XGBoost + Combined). Based on aggregated results from 5-fold cross-validation (N = 1,085), threshold = 0.50. True negative = 750, false positive = 159, false negative = 36, true positive = 140. Accuracy 82.03%, sensitivity 46.82%, specificity 95.42%.

Figure3

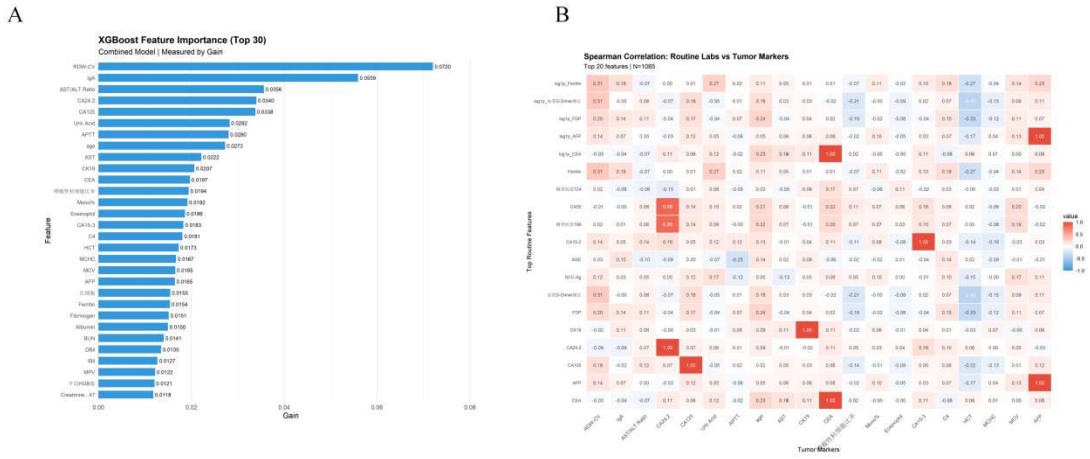


Figure 3 XGBoost Feature Importance and Top 20 Feature Correlation Matrix

(A) Top 30 feature importance ranking for the XGBoost Combined model (measured by Gain). RDW-CV ranked first (Gain = 0.0720), followed by IgA (0.0559), AST/ALT ratio (0.0356), CA24-2 (0.0340), and CA125 (0.0338).

(B) Spearman rank correlation heatmap of the Top 20 features (N = 1,085). Red = positive correlation, Blue = negative correlation. CA24-2 was highly correlated with CA125 (r = 0.85), FDP was correlated with D-dimer (r = 0.40), and D-dimer was correlated with fibrinogen (r = 0.31).

Figure4

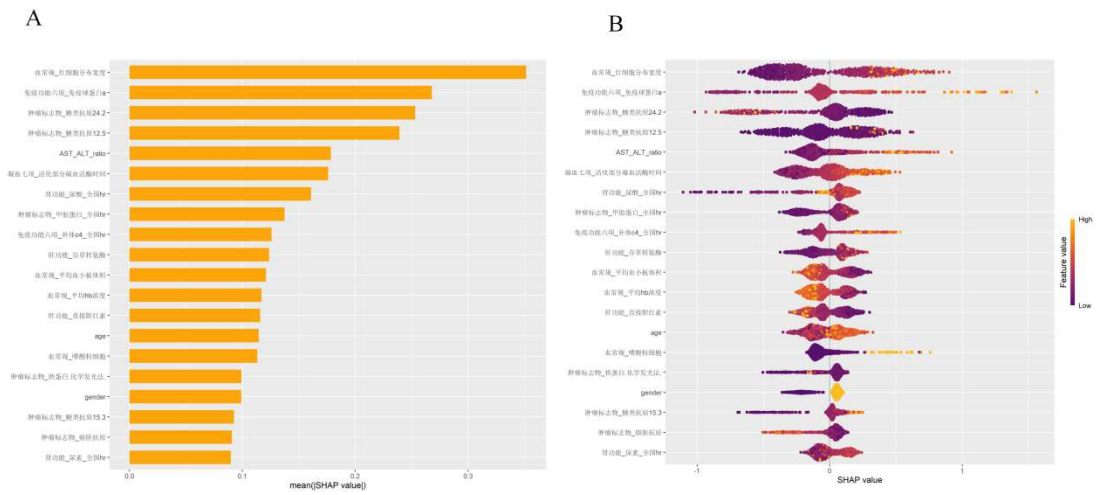


Figure 4 SHAP Interpretability Analysis

(A) Bar plot of mean absolute SHAP values for the Top 20 features. RDW-CV contributed the most, followed by IgA, AST/ALT ratio, CA24-2, CA125, etc. Features are arranged in descending order of their overall influence on model prediction.

(B) SHAP beeswarm plot. Each row represents a feature, each dot represents a sample. Dot position indicates SHAP value (right = pushes prediction toward metastasis, left = pushes prediction toward non-metastasis); color indicates feature value magnitude (yellow = high value, purple = low value). For example, high RDW-CV values (yellow dots) are predominantly distributed on the right side, indicating that elevated RDW increases metastasis risk.

Figure5

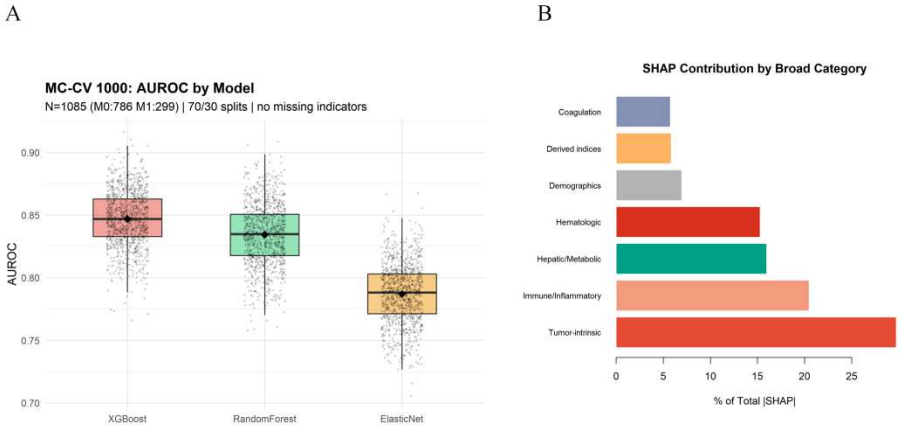


Figure 5 MCCV Robustness Validation and Functional Module Contribution

(A) Boxplot of AUROC from 1,000 Monte Carlo cross-validations (70/30 stratified random sampling, N = 1,085). XGBoost median AUROC = 0.847 (IQR: 0.833–0.861), Random Forest = 0.834 (0.818–0.850), ElasticNet = 0.787 (0.769–0.804). Dots represent individual AUROC values from each iteration.

(B) SHAP functional module contribution proportions. Tumor-intrinsic (tumor markers) contributed the most (27%), followed by Immune/Inflammatory (20%), Hepatic/Metabolic (16%), Hematologic (15%), Demographics (~7%), Derived indices (6%), and Coagulation (5%).

Figure6

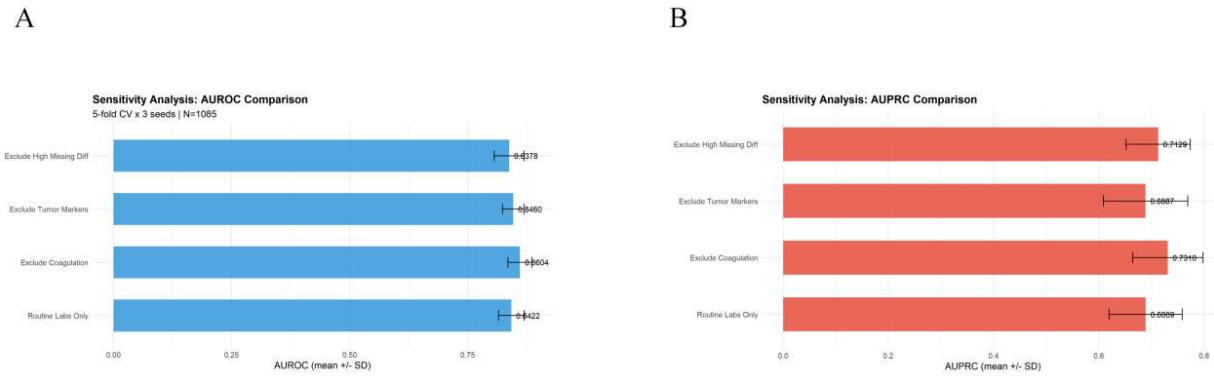


Figure 6 Sensitivity Analysis

(A) Comparison of AUROC across four sensitivity analysis scenarios (mean ± standard deviation). Exclude High Missing Diff = 0.8378, Exclude Tumor Markers = 0.8460, Exclude Coagulation = 0.8604, Routine Labs Only = 0.8422. All scenarios maintained AUROC > 0.83.

(B) Corresponding AUPRC comparison across scenarios. The Exclude Coagulation scenario achieved the highest AUPRC (0.7316), while Exclude High Missing Diff was the lowest (0.7126). Error bars represent ± SD.

Figure7

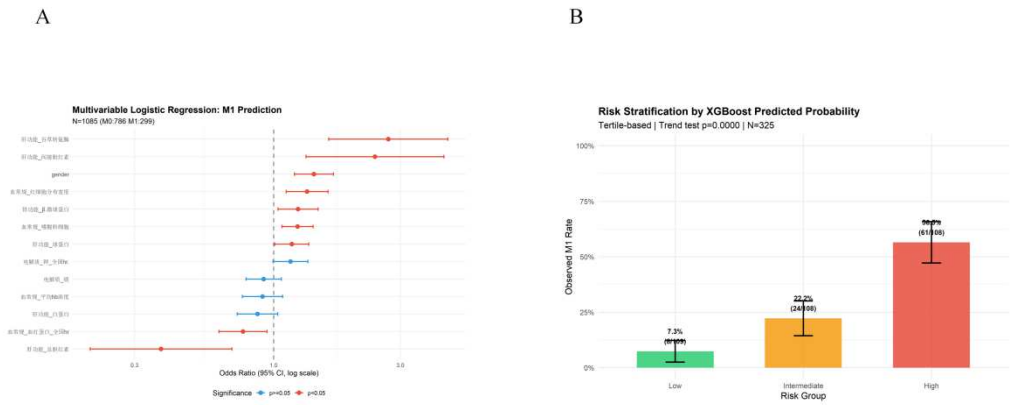


Figure 7 Multivariable Logistic Regression and Risk Stratification Validation

(A) Forest plot of multivariable stepwise Logistic regression (N = 1,085). Red dots = $P < 0.05$, Blue dots = $P \geq 0.05$. AST showed the strongest independent effect (OR = 2.70), followed by indirect bilirubin (OR = 2.41), sex (OR = 1.42), RDW (OR = 1.34), etc. Hemoglobin and total bilirubin were protective factors (OR < 1).

(B) Tertile risk stratification of XGBoost predicted probabilities. Low-risk group (N = 109) metastasis rate 7.3% (8/109), intermediate-risk group (N = 108) 22.2% (24/108), high-risk group (N = 108) 56.5% (61/108). Cochran-Armitage trend test $P < 0.0001$.

Figure8

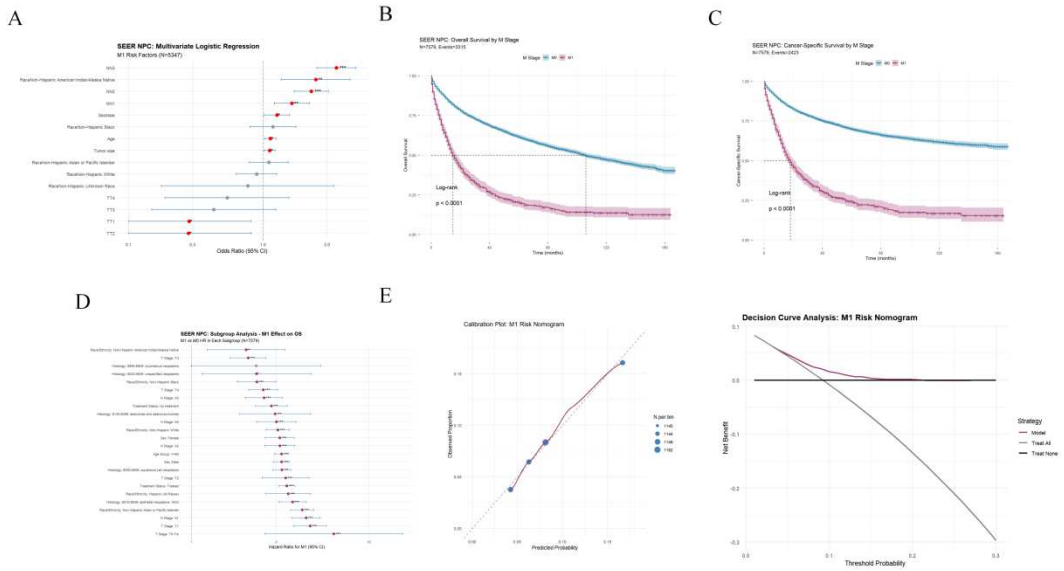


Figure 8 SEER Database External Validation

(A) Multivariable Logistic regression of the SEER database (N = 5,347). N stage was the strongest predictor (N3 vs N0: OR = 2.88), followed by race/ethnicity (American Indian/Alaska Native OR = 2.06), sex (male OR = 1.59), etc. Red = significant, Gray = non-significant.

(B) Kaplan-Meier overall survival curve (N = 7,579). The M0 group (blue) showed significantly superior median OS compared with the M1 group (pink), Log-rank P < 0.0001.

(C) Disease-specific survival curve (N = 7,579). The M1 group demonstrated significantly worse CSS than the M0 group, Log-rank P < 0.0001.

(D) Subgroup analysis forest plot (N = 7,579). Across 24 prespecified subgroups, the adverse prognostic effect of M1 on OS was consistently present in 22 subgroups (HR > 1), encompassing different races, T stages, N stages, age groups, and histological types.

(E) Left = calibration curve of the M1 risk nomogram; Right = decision curve analysis of the nomogram. Nomogram C-index = 0.752.

Figure9

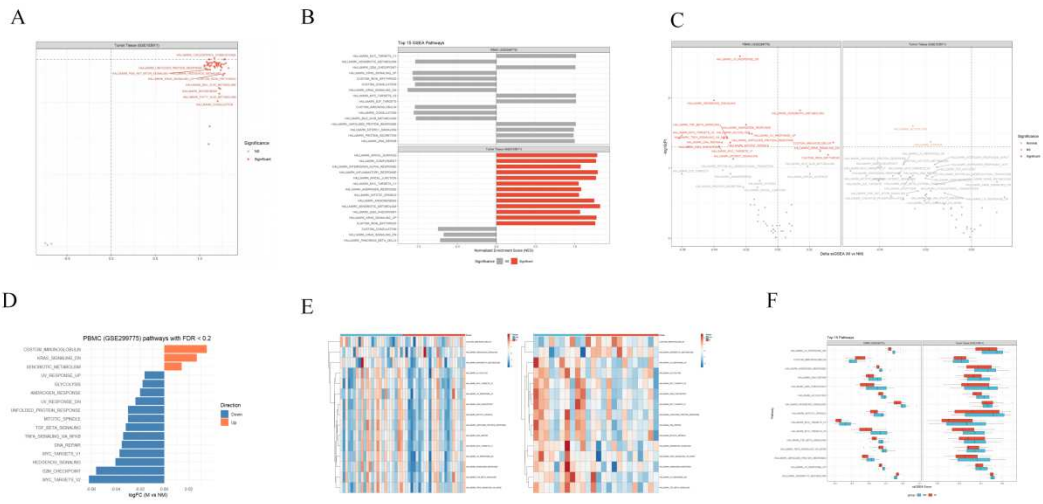


Figure 9 GEO Transcriptomic Pathway Analysis

(A)GSEA volcano plot for tumor tissue (GSE103611, N = 48). X-axis = normalized enrichment score (NES), Y-axis = ranked gene list position. Gray dots = non-significant (NS), red dots = significant pathways (FDR < 0.1). Key pathways labeled include HALLMARK_CHOLESTEROL_HOMEOSTASIS, HALLMARK_UNFOLDED_PROTEIN_RESPONSE, HALLMARK_KRAS_SIGNALING_UP, CUSTOM_IRON_ERYTHROID, HALLMARK_BILE_ACID_METABOLISM, HALLMARK_FATTY_ACID_METABOLISM, and HALLMARK_COAGULATION. Dashed vertical line at NES = 0 indicates neutral enrichment.

(B) Normalized enrichment score (NES) bar plot of the Top 15 GSEA pathways. Upper panel = PBMC (GSE299775, N = 24), lower panel = tumor tissue (GSE103611, N = 48). Gray bars = non-significant (NS), red bars = significant (FDR < 0.1). In PBMC, the top enriched pathways included HALLMARK_MYC_TARGETS_V1, HALLMARK_G2M_CHECKPOINT, HALLMARK_MYC_TARGETS_V2, HALLMARK_E2F_TARGETS, and CUSTOM_IMMUNOGLOBULIN. In tumor tissue, the top enriched pathways included HALLMARK_APICAL_SURFACE, HALLMARK_COMPLEMENT, HALLMARK_INTERFERON_ALPHA_RESPONSE, HALLMARK_INFLAMMATORY_RESPONSE, and HALLMARK_APICAL_JUNCTION. X-axis = NES range from −1.0 to +1.0.

(C) ssGSEA pathway differential volcano plots. Left panel = PBMC (GSE299775), right panel = tumor tissue (GSE103611). X-axis = delta ssGSEA score (metastatic M vs non-metastatic NM), Y-axis = −log10(nominal P value). Red dots = significant (nominal P < 0.05), orange dots = nominal (P < 0.1), gray dots = non-significant (NS). In PBMC, GLYCOLYSIS, HALLMARK_G2M_CHECKPOINT, HALLMARK_MYC_TARGETS_V2, and CUSTOM_IMMUNOGLOBULIN showed negative delta ssGSEA (downregulated in M). In tumor tissue, HALLMARK_CHOLESTEROL_HOMEOSTASIS and HALLMARK_XENOBIOTIC_METABOLISM showed the strongest positive enrichment. Dashed horizontal lines indicate P = 0.05 and P = 0.1 thresholds.

(D) Differentially enriched pathways in PBMC (GSE299775) with FDR < 0.2. X-axis = log2 fold change (M vs NM). Orange bars = upregulated in metastatic group (Up), blue bars = downregulated in metastatic group (Down). CUSTOM_IMMUNOGLOBULIN, KRAS_SIGNALING_DN, and XENOBIOTIC_METABOLISM were upregulated, whereas GLYCOLYSIS, MYC_TARGETS_V2, G2M_CHECKPOINT, HEDGEHOG_SIGNALING, and TNFA_SIGNALING_VIA_NFKB were downregulated. Pathways ordered by effect size.

(E) Hierarchical clustering heatmaps of ssGSEA scores. Left panel = tumor tissue (GSE103611), right panel = PBMC (GSE299775). Rows = 17 selected Hallmark pathways, columns = individual samples. Color scale: red = high enrichment score, blue = low enrichment score, white = neutral. Top annotation bar indicates group membership: cyan = non-metastatic (NM), red = metastatic (M). Pathways clustered by functional similarity, revealing distinct enrichment patterns between the two tissue types and between metastatic and non-metastatic groups.

(F) Boxplot comparison of ssGSEA scores for the Top 15 pathways. Left panel = PBMC (GSE299775), right panel = tumor tissue (GSE103611). Cyan boxes = non-metastatic group (NM), red boxes = metastatic group (M). Y-axis = pathways ranked by significance, X-axis = ssGSEA score. In PBMC, CUSTOM_IMMUNOGLOBULIN, GLYCOLYSIS, HALLMARK_G2M_CHECKPOINT, and HALLMARK_MYC_TARGETS_V2 showed notable inter-group differences. In tumor tissue, HALLMARK_MYC_TARGETS_V1, HALLMARK_MITOTIC_SPINDLE, HALLMARK_UNFOLDED_PROTEIN_RESPONSE, and HALLMARK_HEDGEHOG_SIGNALING exhibited group separation. GLYCOLYSIS demonstrated consistent nominal significance across both independent cohorts (GSE103611: P = 0.025; GSE299775: P = 0.033), representing the sole cross-dataset replicable finding.

Table 1. Significant associations between pathway activities and clinical predictors identified by GSEA

Clinical Predictor	Pathway	Dataset	NES	P-value	FDR	Direction
D-dimer	HALLMARK_COMPLEMENT	Tumor Tissue (GSE103611)	1.272	0.050	0.074	Activated
LDH	CUSTOM_LACTATE_METABOLISM	Tumor Tissue (GSE103611)	1.127	0.052	0.074	Activated
	HALLMARK_GLYCOLYSIS	Tumor Tissue (GSE103611)	1.160	0.051	0.074	Activated
	HALLMARK_HYPOXIA	Tumor Tissue (GSE103611)	1.190	0.053	0.074	Activated
CA125	HALLMARK_E2F_TARGETS	Tumor Tissue (GSE103611)	1.088	0.052	0.074	Activated
	HALLMARK_G2M_CHECKPOINT	Tumor Tissue (GSE103611)	1.069	0.057	0.074	Activated
IgA	HALLMARK_IL6_JAK_STAT3_SIGNALING	Tumor Tissue (GSE103611)	1.170	0.051	0.074	Activated
	HALLMARK_INFLAMMATORY_RESPONSE	Tumor Tissue (GSE103611)	1.293	0.053	0.074	Activated
β2-MG	HALLMARK_INTERFERON_GAMMA_RESPONSE	Tumor Tissue (GSE103611)	1.131	0.054	0.074	Activated
	HALLMARK_TNFA_SIGNALING_VIA_NFKB	Tumor Tissue (GSE103611)	1.172	0.056	0.074	Activated
AST/ALT	HALLMARK_XENOBIOTIC_METABOLISM	Tumor Tissue (GSE103611)	1.324	0.057	0.074	Activated
	HALLMARK_FATTY_ACID_METABOLISM	Tumor Tissue (GSE103611)	1.220	0.086	0.097	Activated
RDW-SD	CUSTOM_IRON_ERYTHROID	Tumor Tissue (GSE103611)	1.262	0.064	0.077	Activated
	HALLMARK_HEME_METABOLISM	Tumor Tissue (GSE103611)	1.233	0.054	0.074	Activated
Metastasis	HALLMARK_ALLOGRAFT_REJECTION	Tumor Tissue (GSE103611)	1.236	0.055	0.074	Activated
	HALLMARK_ANGIOGENESIS	Tumor Tissue (GSE103611)	1.248	0.056	0.074	Activated
	HALLMARK_APOPTOSIS	Tumor Tissue (GSE103611)	1.110	0.052	0.074	Activated
	HALLMARK_MYC_TARGETS_V1	Tumor Tissue (GSE103611)	1.050	0.054	0.074	Activated

	HALLMARK_MYC_TARGETS_V2	Tumor Tissue (GSE103611)	1.110	0.049	0.050	Activated
	HALLMARK_OXIDATIVE_PHOSPHORYLATION	Tumor Tissue (GSE103611)	1.116	0.055	0.074	Activated

Table 2. Performance Metrics of Nine Model–Feature Set Combinations for Distant Metastasis Prediction

Model	Feature Set	AUROC	AUPRC	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1	Brier Score
XGBoost	**Combined**	**0.8595 ± 0.0231**	**0.6275**	**82.03**	**46.82**	**95.42**	**79.54**	**82.51**	**0.5895**	**0.1223**
XGBoost	Routine Labs	0.8472 ± 0.0276	0.6936	80.12	46.63	93.03	72.47	82.04	0.5618	0.1353
XGBoost	Tumor Markers	0.8003 ± 0.0307	0.6092	77.60	39.02	92.47	66.31	79.91	0.4852	0.1541
RandomForest	Combined	0.8548 ± 0.0224	0.7359	80.43	34.42	98.11	87.95	79.64	0.4901	0.1432
RandomForest	Routine Labs	0.8362 ± 0.0262	0.7020	79.05	31.21	97.46	83.46	78.76	0.4487	0.1486
RandomForest	Tumor Markers	0.7921 ± 0.0377	0.6140	77.51	28.29	96.32	74.76	77.88	0.4077	0.1587
ElasticNet	Combined	0.7972 ± 0.0291	0.6073	77.57	35.41	93.71	68.20	79.25	0.4604	0.1551
ElasticNet	Routine Labs	0.7799 ± 0.0250	0.5743	75.76	29.02	93.58	64.40	77.64	0.3933	0.1624
ElasticNet	Tumor Markers	0.7202 ± 0.0486	0.5128	74.44	19.83	95.25	62.47	75.76	0.2953	0.1747

Note: AUROC and AUPRC presented as mean ± SD across 20 repeats of 5-fold cross-validation (N = 1,085). Accuracy, Sensitivity, Specificity, PPV, NPV, F1, and Brier Score for the XGBoost–Combined combination (**highlighted**) were derived from the independent hold-out test set (N = 325); metrics for all other combinations were derived from cross-validation. **Bold** = optimal model. AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve; PPV, positive predictive value; NPV, negative predictive value.

Table 3. Pairwise Statistical Comparison of AUROC Among Feature Sets (XGBoost Algorithm)

Comparison	N (Common Folds)	Test	P-value
Combined vs. Routine Labs	15	Wilcoxon signed-rank	**0.038**
Combined vs. Tumor Markers	15	Wilcoxon signed-rank	**0.001**
Routine Labs vs. Tumor Markers	15	Wilcoxon signed-rank	**0.004**

Note: Pairwise comparisons performed using the Wilcoxon signed-rank test on AUROC values from $N = 15$ common folds across 20 repeats of 5-fold cross-validation. All three comparisons achieved statistical significance ($P < 0.05$), forming a complete statistical evidence chain: Combined > Routine Labs > Tumor Markers.

Table 4. Contribution of Seven Functional Categories to Model Predictions Based on SHAP Analysis

Functional Category	Total |SHAP|	Contribution (%)	N Features	Top Features
Tumor Markers	0.912	**29.74**	6	CA24-2; CA12-5; AFP; CA15-3; CEA; Ferritin
Immune Inflammation	0.628	**20.48**	4	IgA; C4; MPV; Eosinophils
Liver Metabolic	0.490	**15.98**	4	UA; AST; DBil; BUN
Hematologic Erythrocyte	0.468	**15.27**	2	RDW-CV; MCHC
Demographics	0.214	**6.97**	2	Age; Gender
Derived Indices	0.178	**5.82**	1	AST/ALT Ratio
Coagulation	0.176	**5.74**	1	APTT

Note: SHAP, SHapley Additive exPlanations; |SHAP|, mean absolute SHAP value; CA, cancer antigen; AFP, alpha-fetoprotein; CEA, carcinoembryonic antigen; IgA, immunoglobulin A; C4, complement component 4; MPV, mean platelet volume; UA, uric acid; AST, aspartate aminotransferase; DBil, direct bilirubin; BUN, blood urea nitrogen; RDW-CV, red cell distribution width–coefficient of variation; MCHC, mean corpuscular hemoglobin concentration; APTT, activated partial thromboplastin time. Top three categories (bold contribution values) collectively account for 66.20% of total predictive information.

Table 5. Monte Carlo Cross-Validation Performance Summary (1,000 Iterations)

Model	AUROC (95% CI)	AUPRC	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1	Brier Score	N
XGBoost	**0.8470 [0.8032, 0.8871]**	**0.7019**	**81.02**	**47.27**	**93.91**	**74.81**	**82.41**	**0.5766**	**0.1341**	**1000**
RandomForest	0.8344 [0.7848, 0.8808]	0.6998	79.58	33.79	97.05	81.58	79.40	0.4749	0.1470	1000
ElasticNet	0.7871 [0.7385, 0.8314]	0.5655	75.75	29.06	93.56	63.86	77.67	0.3909	0.1624	1000

Note: MC-CV, Monte Carlo cross-validation; 1,000 iterations of random 70/30 train–test splits. CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve.

Table 6. Sensitivity Analysis: Robustness of XGBoost–Combined Model Under Alternative Feature Set Scenarios

Analysis Scenario	N Features	AUROC	AUPRC	Sensitivity (%)	Specificity (%)	F1	Brier Score	ΔAUROC*	Robust
1. Exclude high-missingness features	89	0.8378	0.7129	47.34	94.79	0.5845	0.1350	**-0.0217**	**Yes**
2. Exclude all tumor markers	98	0.8460	0.6887	48.02	93.45	0.5771	0.1362	**-0.0135**	**Yes**
3. Exclude coagulation indices	103	0.8604	0.7310	50.11	94.27	0.6045	0.1278	**+0.0009**	**Yes**
4. Routine labs only (no tumor markers)	89	0.8422	0.6889	47.24	92.41	0.5617	0.1378	**-0.0173**	**Yes**

Note: All scenarios evaluated using 5-fold cross-validation with 20 repeats on the full cohort (N = 1,085).

* $\Delta AUROC = AUROC_{scenario} - AUROC_{baseline}$ (baseline = 0.8595). Green = performance improvement; red = performance decline. All scenarios maintain robustness with $|\Delta AUROC| < 0.03$.

Table 7. Univariate Logistic Regression Analysis of Factors Associated with Distant Metastasis

(FDR-corrected P < 0.05, Benjamini–Hochberg method)

Variable	OR	95% CI	P-value	FDR	Direction
RDW (%)	**1.763**	**1.535–2.024**	**9.70×10^{-16}**	**4.85×10^{-14}**	**Risk**
Hb (g/L)	**0.583**	**0.507–0.671**	**4.16×10^{-14}**	**1.04×10^{-12}**	**Protective**
HCT (%)	**0.605**	**0.526–0.695**	**1.37×10^{-12}**	**2.28×10^{-11}**	**Protective**
β2-MG (mg/L)	**1.545**	**1.341–1.781**	**1.81×10^{-9}**	**2.27×10^{-8}**	**Risk**
MCHC (g/L)	**0.668**	**0.581–0.769**	**1.61×10^{-8}**	**1.61×10^{-7}**	**Protective**
Alb (g/L)	**0.713**	**0.624–0.815**	**7.19×10^{-7}**	**5.99×10^{-6}**	**Protective**
Gender (Male)	**1.365**	**1.176–1.585**	**4.30×10^{-5}**	**3.07×10^{-4}**	**Risk**
Eosinophils ($\times10^9$/L)	**1.269**	**1.125–1.431**	**1.03×10^{-4}**	**6.41×10^{-4}**	**Risk**
IBil (μmol/L)	**0.747**	**0.644–0.867**	**1.26×10^{-4}**	**6.98×10^{-4}**	**Protective**
AST (U/L)	**2.152**	**1.369–3.384**	**8.00×10^{-4}**	**4.28×10^{-3}**	**Risk**
Potassium (mmol/L)	**1.256**	**1.097–1.437**	**9.42×10^{-4}**	**4.28×10^{-3}**	**Risk**
TBil (μmol/L)	**0.795**	**0.682–0.926**	**3.30×10^{-3}**	**1.38×10^{-2}**	**Protective**
BUN (mmol/L)	**1.206**	**1.060–1.373**	**4.57×10^{-3}**	**1.76×10^{-2}**	**Risk**
Globulin (g/L)	**1.204**	**1.056–1.372**	**5.48×10^{-3}**	**1.80×10^{-2}**	**Risk**
Magnesium (mmol/L)	**0.831**	**0.728–0.947**	**5.68×10^{-3}**	**1.80×10^{-2}**	**Protective**
Phosphorus (mmol/L)	**0.796**	**0.676–0.936**	**5.78×10^{-3}**	**1.80×10^{-2}**	**Protective**
Age (years)	**1.191**	**1.041–1.362**	**1.09×10^{-2}**	**3.21×10^{-2}**	**Risk**
Monocyte Ratio (%)	**1.174**	**1.032–1.337**	**1.52×10^{-2}**	**4.21×10^{-2}**	**Risk**

Note: Univariate logistic regression with FDR correction (Benjamini–Hochberg method). **Bold** values indicate statistical significance (FDR < 0.05). Red = risk factor (OR > 1); Green = protective factor (OR < 1). OR, odds ratio; CI, confidence interval; FDR, false discovery rate; RDW, red cell distribution width; Hb, hemoglobin; HCT, hematocrit; β2-MG, β2-microglobulin; MCHC, mean corpuscular hemoglobin concentration; Alb, albumin; IBil, indirect bilirubin; AST, aspartate aminotransferase; TBil, total bilirubin; BUN, blood urea nitrogen.

n.

Table 8. Multivariable Logistic Regression Analysis of Independent Predictors of Distant Metastasis

(All 59 candidate variables entered; P < 0.05 considered statistically significant)

Variable	OR	95% CI	P-value	Direction	Rank
AST (U/L)	**2.701**	**1.613–4.522**	**1.58×10^{-4}**	**Risk**	**1**
Gender (Male)	**1.418**	**1.197–1.679**	**5.18×10^{-5}**	**Risk**	**2**
IBil (μmol/L)	**2.406**	**1.324–4.371**	**3.96×10^{-3}**	**Risk**	**3**
RDW (%)	**1.337**	**1.115–1.602**	**1.71×10^{-3}**	**Risk**	4
Eosinophils ($\times10^9$/L)	**1.231**	**1.076–1.408**	**2.45×10^{-3}**	**Risk**	5
TBil (μmol/L)	**0.377**	**0.204–0.697**	**1.85×10^{-3}**	**Protective**	6
Globulin (g/L)	**1.170**	**1.010–1.356**	**3.60×10^{-2}**	**Risk**	7
β2-MG (mg/L)	**1.236**	**1.039–1.470**	**1.67×10^{-2}**	**Risk**	8
Hb (g/L)	**0.767**	**0.624–0.944**	**1.24×10^{-2}**	**Protective**	9

Note: Multivariable logistic regression with all 59 candidate variables entered simultaneously. **Bold** values indicate statistical significance (P < 0.05). Red = risk factor (OR > 1); Green = protective factor (OR < 1). **Top three predictors** highlighted by light gold shading. Seven of nine predictors (AST, Gender, RDW, Eosinophils, Globulin, β2-MG, Hb) were concurrently ranked in the Top 20 SHAP features, demonstrating cross-methodological consistency. OR, odds ratio; CI, confidence interval; AST, aspartate aminotransferase; IBil, indirect bilirubin; RDW, red cell distribution width; TBil, total bilirubin; β2-MG, β2-microglobulin; Hb, hemoglobi

Table 9. Risk Stratification Performance of the XGBoost–Combined Model

(Independent Hold-out Test Set, N = 325)

Risk Group	N	M1 Events	M1 Rate (%)	Predicted Probability Median (%)	Predicted Probability IQR (%)	Risk Ratio (vs. Low)
Low	109	8	**7.34**	3.45	1.97 – 4.55	**1.00 (Ref.)**
Intermediate	108	24	**22.22**	15.44	9.40 – 20.76	**3.03**
High	108	61	**56.48**	51.47	39.58 – 70.63	**7.70**

Note: Patients were stratified into three risk groups using tertiles of predicted probabilities from the XGBoost–Combined model on the independent hold-out test set ($N = 325$). Risk ratio calculated as M1 rate in each group divided by M1 rate in the Low-risk group.

M1 rate: Low (7.34%) vs. Intermediate (22.22%) vs. High (56.48%), $\chi^2 = 74.3$, $P < 0.001$. The model demonstrated a clear gradient of metastatic risk across tertiles, with a 7.70-fold increased risk in the High-risk group compared to the Low-risk group.

Table 10. Multivariable Logistic Regression Analysis of Distant Metastasis Risk Factors (SEER Database)

(N = 5,347, AIC = 3,257.61)

Variable	OR	95% CI	P-value	Direction	Sig.
N Stage (ref: N0)					
N1	**1.638**	**1.214–2.210**	**1.24×10^{-3}**	**Risk**	**
N2	**2.290**	**1.706–3.075**	**3.53×10^{-8}**	**Risk**	***
N3	**3.514**	**2.521–4.899**	**1.24×10^{-13}**	**Risk**	***
T Stage (ref: T0)					
T1	**0.282**	**0.098–0.812**	**1.90×10^{-2}**	**Protective**	*
T2	**0.278**	**0.095–0.814**	**1.96×10^{-2}**	**Protective**	*
T3	**0.430**	**0.148–1.245**	**1.20×10^{-1}**	**Protective**	ns
T4	**0.540**	**0.187–1.560**	**2.55×10^{-1}**	**Protective**	ns
Demographics					
Male (vs. Female)	**1.269**	**1.023–1.575**	**3.06×10^{-2}**	**Risk**	*
Age (per SD)	**1.131**	**1.026–1.248**	**1.38×10^{-2}**	**Risk**	*
Tumor size (per SD)	**1.120**	**1.022–1.228**	**1.57×10^{-2}**	**Risk**	*
Race (ref: White)					
American Indian/Alaska Native	**2.466**	**1.370–4.439**	**2.60×10^{-3}**	**Risk**	**

Note: Multivariable logistic regression. **Bold** values indicate statistical significance (P < 0.05). Red = risk factor; Green = protective factor. Sig.: *P<0.05, **P<0.01, ***P<0.001, ns = not significant. AIC, Akaike information criterion; OR, odds ratio; CI, confidence interval; SD, standard deviation.

Table 11. Survival Outcomes by Distant Metastasis Status in the SEER Validation Cohort

(N = 7,579)

Endpoint	N	Total Events	M0 Events	M1 Events	Log-rank χ^2	P-value
Overall Survival (OS)	7,579	3,315	2,740	575	**695.63**	**< 0.001**
Cancer-Specific Survival (CSS)	7,579	2,423	1,913	510	**829.30**	**< 0.001**

Note: Kaplan–Meier method with Log-rank test. M0, no distant metastasis; M1, distant metastasis; OS, overall survival; CSS, cancer-specific survival. Median OS: M1 ≈ 18 months vs. M0 ≈ 72 months.

Table 12. Subgroup Analysis: Effect of Distant Metastasis on Overall Survival (SEER Database)

(N = 7,579, Cox Proportional Hazards Model)

Subgroup	Level	N	M1	Events	HR (95% CI)	P-value	Sig.
Age Group							
	≥ 60 years	7,575	780	3,315	**3.210 (2.931–3.517)**	**8.74×10^{-139}**	***
Sex							
	Female	2,209	181	877	**3.140 (2.601–3.790)**	**9.19×10^{-33}**	***
	Male	5,370	599	2,438	**3.212 (2.894–3.566)**	**1.74×10^{-106}**	***
Race/Ethnicity							
	Hispanic (All Races)	760	79	311	**3.500 (2.608–4.697)**	**7.07×10^{-17}**	***
	American Indian/Alaska Native	124	27	77	**2.028 (1.222–3.365)**	**6.24×10^{-3}**	**
	Asian/Pacific Islander	2,933	328	1,044	**4.201 (3.612–4.885)**	**1.39×10^{-77}**	***
	Non-Hispanic Black	905	102	453	**2.330 (1.814–2.994)**	**3.66×10^{-11}**	***
	Non-Hispanic White	2,793	241	1,423	**3.062 (2.630–3.564)**	**3.39×10^{-47}**	***
T Stage							
	T0–Tis	58	9	21	**6.339 (2.594–15.489)**	**5.10×10^{-5}**	***
	T1	2,046	135	729	**4.659 (3.773–5.754)**	**2.45×10^{-46}**	***
	T2	1,001	76	385	**3.386 (2.499–4.590)**	**3.74×10^{-15}**	***
	T3	1,337	138	608	**2.083 (1.651–2.627)**	**5.80×10^{-10}**	***
	T4	1,387	186	731	**2.532 (2.098–3.055)**	**3.39×10^{-22}**	***
N Stage							
	N0	1,436	85	673	**3.005 (2.327–3.880)**	**3.36×10^{-17}**	***
	N1	2,107	195	829	**4.426 (3.680–5.324)**	**3.88×10^{-56}**	***
	N2	1,725	194	689	**3.143 (2.586–3.821)**	**1.32×10^{-30}**	***
	N3	750	129	382	**2.557 (2.015–3.244)**	**1.08×10^{-14}**	***
Histology							
	Squamous cell neoplasms	5,292	490	2,306	**3.232 (2.881–3.626)**	**6.02×10^{-89}**	***
	Epithelial neoplasms, NOS	1,807	220	741	**3.713 (3.111–4.430)**	**6.00×10^{-48}**	***
Treatment Status							
	Treated	6,530	656	2,623	**3.439 (3.105–3.808)**	**1.27×10^{-124}**	***
	No treatment	1,045	123	689	**2.820 (2.294–3.465)**	**6.40×10^{-23}**	***

Note: Cox proportional hazards regression of overall survival by distant metastasis status within each subgroup. All 24 subgroups showed HR > 1; 23/24 subgroups achieved statistical significance (P < 0.05). **Bold** values indicate HR and P-value. Sig.: significance level (**P<0.01,

***P<0.001). Asian/Pacific Islander subgroup showed the strongest prognostic effect (HR = 4.201). HR, hazard ratio; CI, confidence interval; NOS, not otherwise specified.

Table S1. Baseline Characteristics of the Model Development Cohort (N = 1,085)

Format: Median (IQR) or n (%)

Category	Variable	Overall (N = 1,085)	M0 (n = 786)	M1 (n = 299)	P-value
Demographics					
	Gender (Female)	274 (25.3%)	225 (28.6%)	49 (16.4%)	**<0.001**
	Gender (Male)	811 (74.7%)	561 (71.4%)	250 (83.6%)	—
	Age (years)	48.00 (16.00)	47.00 (16.75)	51.00 (15.00)	**0.015**
Complete Blood Count					
	MCV (fL)	91.30 (9.40)	91.30 (9.07)	91.30 (9.95)	0.468
	Lymphocyte Count (×10⁹/L)	0.92 (0.62)	0.94 (0.60)	0.82 (0.61)	**0.041**
	Lymphocyte Ratio (%)	18.50 (13.30)	19.30 (13.47)	16.80 (12.25)	**0.007**
	Hematocrit (HCT, %)	35.30 (9.70)	36.20 (9.05)	32.30 (10.10)	**<0.001**
	Monocyte Ratio (%)	7.70 (5.80)	7.60 (4.70)	8.15 (7.68)	0.118
	RDW (%)	14.10 (3.00)	13.70 (2.60)	15.20 (3.30)	**<0.001**
	RBC (×10¹²/L)	3.95 (1.10)	4.05 (1.05)	3.66 (1.21)	**<0.001**
	Eosinophil Ratio (%)	1.70 (2.70)	1.80 (2.50)	1.50 (2.80)	**0.008**
	Eosinophil Count (×10⁹/L)	0.11 (0.17)	0.11 (0.16)	0.12 (0.21)	0.772
	Basophil Ratio (%)	0.30 (0.30)	0.30 (0.30)	0.30 (0.40)	0.569
	Basophil Count (×10⁹/L)	0.02 (0.03)	0.02 (0.03)	0.02 (0.03)	0.642
	MCH (pg)	30.30 (3.50)	30.40 (3.48)	30.00 (3.70)	**0.048**
	MCHC (g/L)	329.00 (17.00)	331.00 (16.00)	323.00 (18.00)	**<0.001**
	Platelet Count (×10⁹/L)	211.00 (103.00)	213.00 (99.00)	207.00 (118.00)	0.150
	Hemoglobin (Hb, g/L)	116.00 (35.00)	120.00 (31.00)	105.00 (37.00)	**<0.001**
	Neutrophil Ratio (%)	68.90 (16.10)	68.45 (14.90)	70.00 (17.30)	0.299
	Neutrophil Count (×10⁹/L)	3.45 (2.61)	3.47 (2.42)	3.44 (3.28)	0.826
	MPV (fL)	9.00 (1.60)	9.10 (1.60)	8.95 (1.57)	**0.048**
Coagulation					
	D-Dimer (mg/L)	0.21 (0.44)	0.20 (0.31)	0.40 (0.92)	**<0.001**
	AT-III Activity (%)	100.15 (19.70)	100.85 (19.20)	96.85 (22.55)	**0.004**
	FDP (μg/mL)	2.42 (1.38)	2.36 (1.14)	2.83 (2.59)	**<0.001**
Liver Function					
	Globulin (GLB, g/L)	26.05 (5.93)	25.70 (5.70)	26.90 (7.40)	**0.020**
Immune Function					
	IgE (IU/mL)	60.00 (155.25)	52.00 (111.50)	81.00 (308.25)	**<0.001**
	IgM (g/L)	0.86 (0.54)	0.88 (0.53)	0.81 (0.46)	**0.041**
	Complement C3 (g/L)	3.00 (0.00)	3.00 (0.00)	3.00 (0.00)	**0.015**
	IgA (g/L)	1.79 (1.12)	1.67 (1.08)	2.07 (1.18)	**<0.001**
	IgG (g/L)	9.67 (4.22)	9.45 (3.99)	10.35 (5.04)	**0.009**
	Complement C4 (g/L)	0.23 (0.11)	0.22 (0.10)	0.26 (0.10)	**<0.001**
Renal Function					
	Retinol-Binding Protein (mg/L)	3.59 (1.46)	3.63 (1.43)	3.48 (1.61)	0.050
	Serum Cystatin C (mg/L)	0.97 (0.31)	0.96 (0.30)	0.99 (0.29)	0.055
Urinalysis					
	Urine Vitamin C	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.158
	Urine Protein	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.330
	Urine WBC (Qualitative)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.432
	Urine Bilirubin	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	—
	Urine Specific Gravity	1.01 (0.01)	1.01 (0.01)	1.01 (0.01)	0.723
	Urine Ketone	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.262

	Urine pH	6.00 (1.12)	6.00 (1.00)	6.00 (1.50)	0.151
	Urine WBC Count	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.077
	Urine Occult Blood	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.084
	Urine Glucose	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.883

Note: Continuous variables are presented as median (interquartile range); categorical variables as n (%). *P*-values from Wilcoxon rank-sum test (continuous) or Chi-square test (categorical). **Bold** P-values indicate statistical significance ($P < 0.05$). MCV, mean corpuscular volume; RDW, red cell distribution width; RBC, red blood cell; MCH, mean corpuscular hemoglobin; MCHC, mean corpuscular hemoglobin concentration; MPV, mean platelet volume; AT-III, antithrombin III; FDP, fibrin degradation products; Ig, immunoglobulin.

Supplementary Files

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

- SupplementaryMaterials1.png
- SupplementaryMaterials2.png
- SupplementaryMaterials3.png
- SupplementaryMaterials4.png
- SupplementaryMaterials5.png