International Journal of General Medicine

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Machine Learning-Based Diagnostic Models for Early Gastric Cancer Using Clinical Laboratory Indicators

Runbi Ji[1,2], Ruoyu Yang[1,2], Jun Yao[1], Shenglan Dai[1], Xin Zhu[1], Qiang Ye[1]

[1]The Affiliated People's Hospital of Jiangsu University, Zhenjiang, Jiangsu, 212002, People's Republic of China; [2]Jiangsu Key Laboratory of Medical Science and Laboratory Medicine, School of Medicine, Jiangsu University, Zhenjiang, Jiangsu, 212013, People's Republic of China

Correspondence: Runbi Ji, The Affiliated People's Hospital of Jiangsu University, Zhenjiang, Jiangsu, 212002, People's Republic of China, Tel +86 511 88915575, Fax +86 511 85234387, Email runbiji@163.com

Background: The occurrence of gastric cancer is a complex pathological process leading to multiple abnormalities in clinical laboratory indicators. Machine learning techniques can make it easy to handle millions of variables to make more accurate predictions and diagnoses of diseases.

Methods: Clinical data from gastric cancer patients in a single-center who underwent surgery between 2016 and 2023 were collected. Five machine learning algorithms (extreme gradient boosting, XGBoost; random forest, RF; support vector machine-recursive feature elimination, SVM-RFE; light gradient boosting machine, LGBM; and recursive partitioning, rpart) were utilized to develop diagnostic models. Among the date, 60% were randomly selected to train the models, while the remaining 40% were used for testing. We used the area under the receiver operating characteristic curve (AUROC), F1-score value, sensitivity, and specificity to evaluate the performance of models.

Results: The XGBoost algorithm showed the best performance in gastric cancer diagnosis, with significantly higher area under curve (AUC) (combining blood indicators and pathological parameters, AUC=0.9909) value than other models. Glutathione reductase (GR), carbohydrate antigen 724 (CA724), erythrocytes (RBC), carbohydrate antigen 242 (CA242), and albumin (ALB) contributed the most to the diagnosis. The tumor size were independent risk factors for early gastric cancer.

Conclusion: Machine learning combined blood indicators and pathological parameters could predict gastric cancer risk more accurately. The XGBoost model had the best diagnostic performance. The study provides confirmatory data support for the preclinical implementation of the model.

Keywords: early gastric cancer, machine learning, diagnostic model, clinical laboratory indicators, glutathione reductase

Introduction

Gastric cancer (GC) remains one of the most prevalent malignant tumors worldwide, ranks as the fifth in incidence rate and the fourth in mortality.[1,2] Early gastric cancer (EGC) refers to lesions confined to the mucosal and submucosal, regardless of size or lymph node metastasis. There are often no obvious symptoms in EGC, occasionally accompanied by discomfort similar to chronic gastritis or gastric ulcers. Because of the atypical symptoms, the diagnosis rate of EGC is relatively low.[3–5]

The 5-year survival rate of EGC after surgery (or endoscopic resection) can reach 90% to 95%, much higher than that of advanced gastric cancer (AGC). Early screening and diagnosis are crucial to early management. The methods for early screening and diagnosis include X-ray barium contrast examination, endoscopic examination, serological examination and pathological diagnosis.[6] Upper gastrointestinal X-ray barium contrast examination is radioactive and has a low positive rate, which has gradually been phased out in clinical practice. Gastroscopy and histopathological examination are currently considered as the gold standard for diagnosing GC. But the acceptance of this technology is relatively low for its high cost, limitation of equipment and physician skills, as well as causing discomfort to patients.[3,7] The commonly tumor markers used in serological examination, such as CEA, CA199, CA724, CA125, CA242, have a positive detection

https://doi.org/10.2147/IJGM.S559103
Received: 7 August 2025
Accepted: 14 January 2026
Published: 28 January 2026

rate of only 20%~30% in AGC and <10% in EGC. Their sensitivity and specificity are insufficient.[8,9] Exploring effective screening methods for EGC has significant clinical implications.

Machine learning, a subset of AI, in combination with clinical laboratory data has shown great promise in aiding the diagnosis, prediction, monitoring, and prognosis of diseases.[10] Compared with traditional regression methods, machine learning algorithms are characterized by their superior performance in predicting results within large databases.[3,11,12] For example, Wang et al constructed multiple models using 50 features from the blood routine and biochemical detection data for the diagnosis of various circulation system diseases.[13] At present, serological examination for tumor diagnosis mainly focuses on the detection of tumor markers, with less attention paid to routine testing items. Blood routine examination is often used for diagnosis of hematological diseases. Biochemical indicators are commonly used for evaluations such as liver and kidney function. There are few reports exploring the relationship between conventional test indicators and EGC. The occurrence of gastric cancer is a complex pathological process involving multiple factors and steps. It is unknown whether this persistent pathological state can be predicted early through a combination of conventional laboratory indicators.

Therefore, this study intends to use machine learning to explore the potential relationship between 75 conventional testing indicators and EGC, establish a diagnostic model composed of routine testing indicators, evaluating its feasibility in early diagnosis of gastric cancer, and to provide clinicians with specialized insights for diagnosis and disease prevention.

Materials and Methods

Data Collection and Preprocessing

After deleting incomplete clinical data, we obtained 73 blood routine, biochemical and tumor marker indicators, as well as their age and gender, totally 75 variables of 1652 patients who came for the Affiliated People's Hospital of Jiangsu University for treatment. In addition to the control samples, we also collected pathological information from patients with early gastric cancer (EGC) and advanced gastric cancer (AGC), including tumor distribution area, differentiation degree, histological type, lymph node involvement, invasion depth, tumor size, and TNM stage. For the classification indicators of gender and pathological information, we transformed them into dummy variables using the "dummyVars" function in subsequent analysis. In addition, we used the R package "missForest" to fill in missing information. We used R package "tableone" to statistically analyze the gender, age, and 73 indicators of all patients. Binary logistic regression was used to identify risk factors for early gastric cancer. The R package "autoReg" was used to calculate the odds ratios (ORs) and 95% confidence intervals (CIs) for each indicator, and establish different prediction models.

Model Training and Evaluation

All samples were randomly divided into a training set and a validation set according to the principle of 6:4. We used five machine learning methods, including extreme gradient boosting (XGBoost), random forest (RF), support vector machine-recursive feature elimination (SVM-RFE), light gradient boosting machine (LGBM), and recursive partitioning (rpart) to construct diagnostic models for 75 variables. During the training process, 10 cross validations were conducted for each model to maintain its stability, and random search was used to select the best hyperparameters. We ranked the importance of each feature in the model after adjusting the optimal parameters for 5 machine learning methods, and divided the 75 variables into 5 levels (A-E) based on their importance from high to low. We selected features that were at the (A, B, C) level among at least three methods as the filtered features. In the validation set, the F1 score, area under the receiver operating characteristic curve (AUROC), sensitivity and specificity were used to evaluate each model comprehensively. We also compared the performance differences of different prediction models, and conducted difference tests. DCA curve and calibration curve were used to assess models. We will upload all the code used in this analysis to GitHub (https://github.com/Bon-jour/Identification-of-Early-Gastric-Cancer-Markers.git) for further discussion.

Results

Clinical Baseline Characteristics

A total of 1652 patients were included in the study (male/female: 934/718), with average age of 59.96±13.92 years old. According to their states of illnesses, they were divided into three groups including healthy controls (n=616), early

gastric cancer (EGC) patients (n=509), and advanced gastric cancer (AGC) patients (n=527) (Figure 1). The clinical baseline characteristics and 75 variables are summarized (Supplementary Table 1). And the abbreviations of the variables are presented (Supplementary Table 1).

Specifically, the proportion of males is significantly higher than that of females (P<0.001) in both EGC (57.8%) and AGC (73.2%) groups (Figure 2A). Traditional tumor markers cannot distinguish EGC patients from healthy individuals, and only AFP, CEA, CA199, CA242 and CA724 have significant statistical significance in the AGC group compared to both the control group and the EGC group (Figure 2B). The expression of inflammatory indicators, such as CRP are increased in GC patients. Especially in the AGC group, the average level of CRP (8.6 ± 17.9 mg/L) is about twice that of normal individuals (4.7 ± 5.4 mg/L) (Figure 2C). The expression of anemia indicators (RBC, HCT, HGB) gradually decreases, while RDW gradually increases, from healthy individuals to the EGC and then the AGC group (Figure 2D). Liver function indicators, such as TP, ALB, GLB and PAB, which reflect the nutritional status of the body, significantly reduce in GC patients (Figure 2E). The expression of liver injury markers (ADA, AFU, ALP and GR) increased in GC patients (Figure 2F). The expression of eGFR showed decrease, while CREA increased in GC patients (Figure 2G). The risk factors for coronary heart disease, the reduction of ApoA1, are also common in patients with GC (Figure 2H).

Analysis of Risk or Protective Factors for EGC

We defined the control group as 0 (no event occurred) and the EGC group as 1 (event occurred group). The results of multivariable logistic regression were summarized (Table 1). After multiple tests and corrections (Supplementary Table 2), 4 indicators could serve as protective factors. They were ApoA1 (OR=0.11, P<0.001), eGFR (OR=0.89, P<0.001), AFU (OR=0.91, P=0.017) and CREA (OR = 0.95, P< 0.001). Another 3 indicators could serve as risk factors for EGC. They were male gender (OR=4.06, P<0.001), age (OR=1.08, P<0.001), and GR (OR=1.05, P=0.029). In summary, we found that in

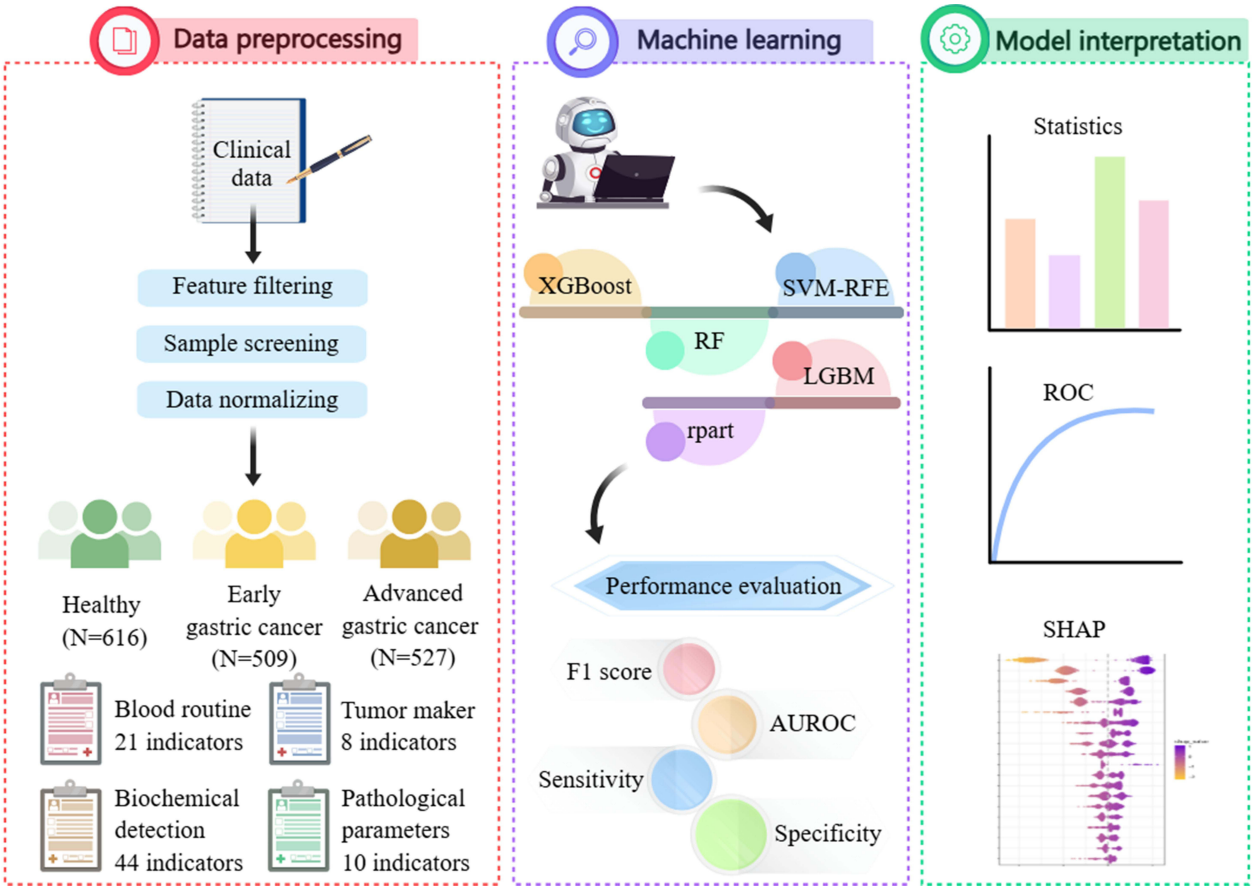

Figure 1 Study design.

https://doi.org/10.2147/IJGM.S559103

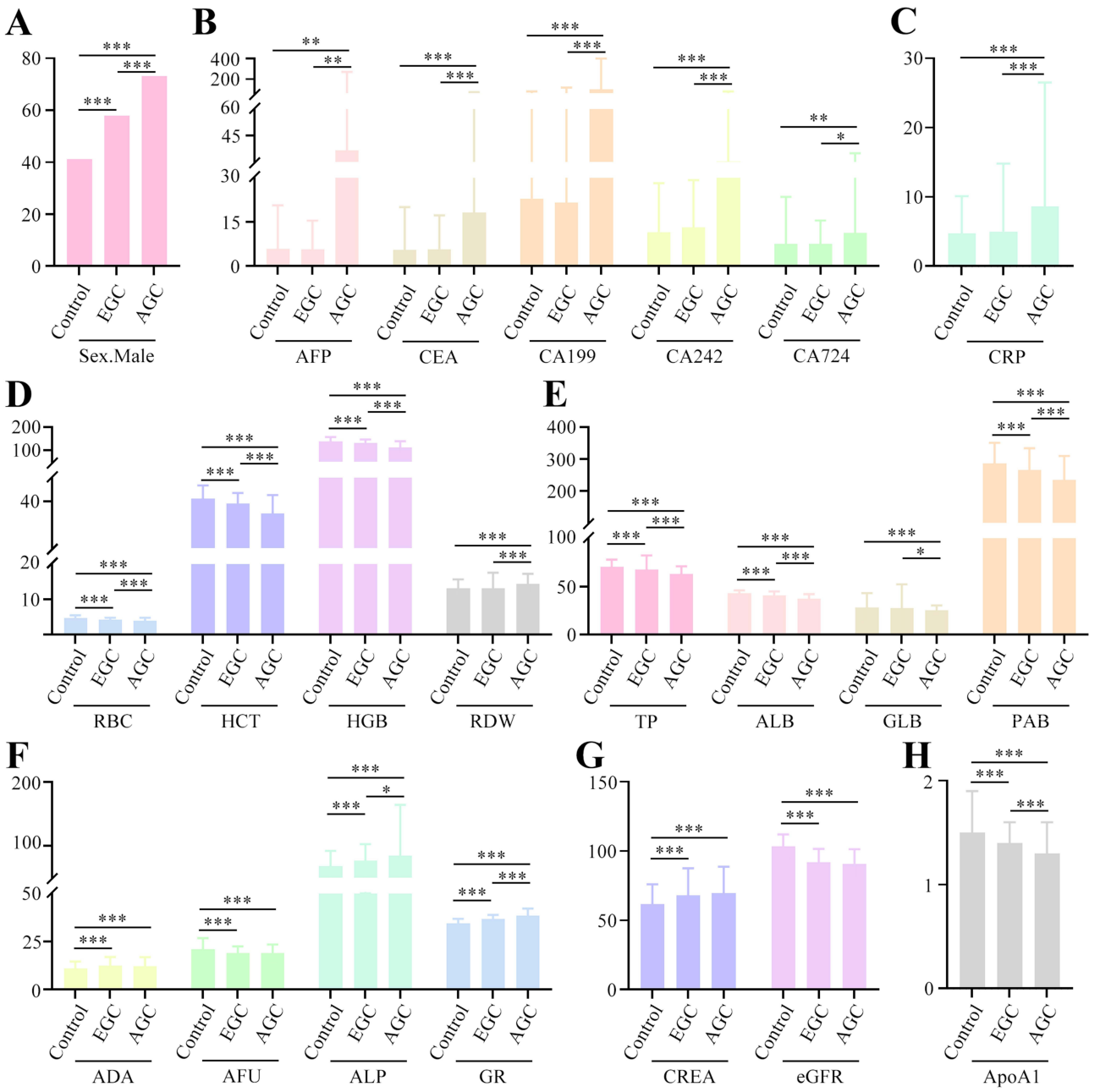

Figure 2 Identification of differential laboratory indicators with consistent trends in healthy controls, early and advanced gastric cancer patients. (**A**) Gender. (**B**) Traditional tumor markers. (**C**) Inflammatory indicators CRP. (**D**) Anemia indicators. (**E**) Liver function indicators. (**F**) Liver injury indicators. (**G**) Renal function indicators. (**H**) Blood lipid indicators. *P < 0.05; **P < 0.01; ***P < 0.001.

addition to age and gender, 5 indicators such as AFU, ApoA1, eGFR, CREA and GR in blood examination were expected to become risk or protective factors for EGC (Table 1).

Machine Learning Models for EGC Diagnosis

Due to the difficulty in explaining the relationship between some indicators obtained by traditional logistic regression statistical methods and GC, we consider using machine learning methods to further explore clinical data in order to obtain relatively accurate conclusions. We ranked the importance of 75 variables by using five machine learning methods (XGBoost, RF, SVM-RFE, LGBM, and rpart). To reduce the number of redundant features, we divided the 75 variables into A, B, C, D, and E levels based on their importance distribution. Those variables that were at A or B levels in at least among three methods

https://doi.org/10.2147/IJGM.S559103

Table 1 The Logistic Regression Results of Multivariable

Variables	0 (N=616)	1 (N=509)	OR (Univariable)	OR (Multivariable)
Age	48.4±13.0	67.3±8.8	1.17 (1.15–1.19, *P* <0.001)	1.08 (1.04–1.11, *P*<0.001)***
Sex-Female (%)	362 (58.8%)	215 (42.2%)		
Sex-Male (%)	254 (41.2%)	294 (57.8%)	1.95 (1.54–2.47, *P*<0.001)	4.06 (2.45–6.72, *P*<0.001)***
AFP (ng/mL)	5.8±14.8	5.6±9.8	1.00 (0.99–1.01, *P*=0.764)	
CEA (ng/mL)	5.4±14.6	5.7±11.5	1.00 (0.99–1.01, *P*=0.772)	
CA125 (U/mL)	20.8±109.1	18.9±11.3	0.99 (0.98–1.00, *P*=0.006)	0.98 (0.96–1.00, *P*=0.056)
CA153 (U/mL)	8.1±3.5	7.6±3.1	0.95 (0.92–0.99, *P*=0.014)	0.96 (0.89–1.04, *P*=0.320)
CA199 (U/mL)	22.9±56.7	21.6±95.1	1.00 (1.00–1.00, *P*=0.784)	
CA242 (U/mL)	11.5±16.7	13.1±16.1	1.01 (1.00–1.01, *P*=0.105)	
CA50 (U/mL)	10.5±15.6	12.7±14.5	1.01 (1.00–1.02, *P*=0.020)	1.00 (0.98–1.02, *P*=0.914)
CA724 (U/mL)	7.5±16.0	7.5±8.0	1.00 (0.99–1.01, *P*=0.978)	
CRP (mg/L)	4.7±5.4	5.0±9.8	1.01 (0.99–1.02, *P*=0.413)	
WBC ($10^{\wedge}9$/L)	6.2±3.2	5.7±1.8	0.88 (0.82–0.94, *P*<0.001)	1.20 (0.86–1.68, *P*=0.288)
BASO_Count ($10^{\wedge}9$/L)	0.0±0.0	0.0±0.1	0.00 (0.00–0.01, *P*<0.001)	2.91 (0.43–19.79, *P*=0.274)
BASO_Percent (%)	0.0±0.4	0.2±4.5	1.03 (0.95–1.12, *P*=0.475)	
EO_Count ($10^{\wedge}9$/L)	0.1±1.5	0.1±0.4	0.95 (0.73–1.22, *P*=0.674)	
EO_Percent (%)	0.0±0.0	0.0±0.0	0.00 (0.00–0.67, *P*=0.184)	
GRAN_Count ($10^{\wedge}9$/L)	3.7±3.1	3.8±1.6	1.01 (0.96–1.06, *P*=0.665)	
GRAN_Percent (%)	1.0±11.5	1.4±16.3	1.00 (0.99–1.01, *P*=0.698)	
LYM_Count (10^{9}/L)	1.6±0.9	1.5±0.6	0.22 (0.18–0.28, *P*<0.001)	0.61 (0.25–1.50, *P*=0.280)
LYM_Percent (%)	0.4±1.3	0.3±0.1	0.00 (0.00–0.00, *P*<0.001)	0.18 (0.00–54.63, *P*=0.559)
MONO_Count ($10^{\wedge}9$/L)	0.4±0.4	0.3±0.1	0.04 (0.02–0.10, *P*<0.001)	0.34 (0.01–13.13, *P*=0.560)
MONO_Percent (%)	0.1±0.2	0.1±0.0	0.00 (0.00–0.00, *P*<0.001)	0.00 (0.00–168.32, *P*=0.133)
RBC ($10^{\wedge}12$/L)	4.7±0.8	4.3±0.5	0.35 (0.27–0.45, *P*<0.001)	1.12 (0.04–33.26, *P*=0.949)
RDW (%)	13.1±2.6	13.2±4.4	1.01 (0.98–1.05, *P*=0.445)	
HCT (%)	41.2±5.6	39.1±4.5	0.92 (0.89–0.94, *P*<0.001)	0.95 (0.59–1.54, *P*=0.847)
HGB (g/L)	138.7±18.7	130.8±16.4	0.97 (0.97–0.98, *P*<0.001)	1.01 (0.89–1.14, *P*=0.919)
MCH (Pg)	32.0±24.9	30.9±12.9	1.00 (0.99–1.00, *P*=0.405)	
MCHC (g/L)	330.2±30.7	334.0±12.2	1.01 (1.00–1.02, *P*=0.017)	1.01 (0.95–1.06, *P*=0.847)
MCV (fL)	89.4±7.4	90.8±5.0	1.04 (1.02–1.07, *P*<0.001)	1.02 (0.86–1.21, *P*=0.807)
MPV (fL)	11.3±11.3	10.9±1.3	0.99 (0.97–1.01, *P*=0.542)	
PLT ($10^{\wedge}9$/L)	213.6±66.1	181.7±64.7	0.99 (0.99–0.99, *P*<0.001)	1.00 (0.99–1.00, *P*=0.382)
TP (g/L)	70.8±7.4	68.0±14.6	0.96 (0.94–0.97, *P*<0.001)	0.96 (0.91–1.02, *P*=0.225)
ALB (g/L)	43.1±3.1	40.9±4.1	0.83 (0.80–0.86, *P* <0.001)	1.08 (0.99–1.18, *P*=0.085)
GLB (g/L)	28.5±14.9	27.9±24.5	1.00 (0.99–1.00, *P*=0.611)	
A/G	1.6±0.2	1.6±1.1	1.34 (0.83–2.18, *P*=0.231)	
PAB (mg/L)	286.3±64.4	265.7±67.9	1.00 (0.99–1.00, *P*<0.001)	1.00 (1.00–1.00, *P*=0.973)
ALT (U/L)	27.9±25.5	24.7±17.2	0.99 (0.99–1.00, *P* =0.020)	1.02 (1.01–1.03, *P*=0.056)
AST (U/L)	25.0±33.0	23.2±10.7	1.00 (0.99–1.00, *P*=0.288)	
ALT/AST	1.3±1.9	1.3±4.1	1.01 (0.97–1.05, *P* =0.722)	
TBIL (μmol/L)	14.0±5.4	14.1±5.5	1.00 (0.98–1.02, *P*=0.854)	
DBIL (μmol/L)	4.2±2.0	4.5±1.9	1.10 (1.03–1.17, *P*=0.003)	0.90 (0.80–1.02, *P*=0.107)
IBIL (μmol/L)	10.1±4.9	9.7±3.9	0.98 (0.95–1.01, *P*=0.134)	
DBIL/TBIL	0.3±0.4	0.3±0.1	0.91 (0.60–1.38, *P*=0.643)	
ADA (U/L)	11.1±3.5	12.3±4.7	1.08 (1.04–1.11, *P* <0.001)	1.01 (0.95–1.08, *P*=0.655)
AFU	21.1±5.7	19.0±3.5	0.91 (0.88–0.94, *P* <0.001)	0.91 (0.84–0.98, *P*=0.017)*
ALP (U/L)	68.0±24.1	76.2±26.2	1.01 (1.01–1.02, *P* <0.001)	1.02 (1.01–1.03, *P*=0.001)**
GGT (U/L)	30.9±38.1	28.6±32.1	1.00 (0.99–1.00, *P*=0.282)	

(*Continued*)

https://doi.org/10.2147/IJGM.S559103

Table 1 (Continued).

Variables	0 (N=616)	1 (N=509)	OR (Univariable)	OR (Multivariable)
CG (mg/L)	1.5±1.2	1.7±0.6	1.44 (1.16–1.78, *P*<0.001)	1.02 (0.78–1.33, *P*=0.828)
GR (U/L)	34.4±2.5	36.8±2.1	0.90 (0.86–0.95, *P*<0.001)	1.05 (0.88–1.14, *P*=0.029)*
TBA (μmol/L)	5.6±7.5	7.0±13.8	1.01 (1.00–1.03, *P*=0.047)	1.00 (0.98–1.02, *P*=0.733)
BUN (mmol/L)	7.9±36.0	5.7±3.6	0.99 (0.98–1.00, *P*=0.112)	
CO2CP (mmol/L)	27.6±4.8	27.4±7.9	1.00 (0.98–1.02, *P*=0.747)	
CREA (μmol/L)	61.6±14.4	68.0±19.6	1.03 (1.02–1.03, *P*<0.001)	0.95 (0.93–0.98, *P*<0.001)***
CYS.C (mg/L)	0.9±0.2	0.9±0.1	0.84 (0.42–1.67, *P*=0.611)	
eGFR	103.2±8.7	91.9±9.7	0.84 (0.82–0.86, *P*<0.001)	0.89 (0.84–0.94, *P*<0.001)***
GLU (mmol/L)	5.8±6.9	6.4±15.4	1.00 (0.99–1.02, *P*=0.436)	
RBP (mg/L)	41.9±5.6	40.4±8.0	0.96 (0.94–0.98, *P*<0.001)	0.98 (0.93–1.03, *P*=0.430)
UA (μmol/L)	319.5±89.5	324.8±93.2	1.00 (1.00–1.00, *P*=0.326)	
CK (U/L)	110.2±60.4	108.4±76.2	1.00 (1.00–1.00, *P*=0.247)	
CKMB (U/L)	7.5±0.5	7.6±0.7	0.95 (0.04–0.07, *P*=0.101)	
HBDH (U/L)	157.2±10.9	154.7±16.4	0.99 (0.98–1.00, *P*=0.002)	0.99 (0.97–1.02, *P*=0.621)
LDH.L (U/L)	189.6±45.6	186.8±131.5	1.00 (1.00–1.00, *P*=0.621)	
m.AST (U/L)	11.7±8.5	11.6±16.0	1.00 (0.99–1.01, *P*=0.949)	
TG (mmol/L)	1.6±1.4	1.9±5.8	1.02 (0.98–1.07, *P*=0.264)	
CHOL (mmol/L)	4.8±0.9	5.8±12.8	1.02 (0.99–1.05, *P*=0.173)	
APO.A1 (g/L)	1.5±0.4	1.4±0.2	0.15 (0.09–0.26, *P*<0.001)	0.11 (0.05–0.23, *P*<0.001)***
APO.B (g/L)	2.0±14.5	1.1±1.4	0.97 (0.92–1.02, *P*=0.276)	
HDL.C (mmol/L)	1.7±2.6	1.3±0.6	0.51 (0.38–0.70, *P*<0.001)	0.96 (0.76–1.22, *P*=0.751)
LDL.C (mmol/L)	3.3±7.1	2.8±2.0	0.97 (0.94–1.01, *P*=0.201)	
Na (mmol/L)	140.1±5.3	139.4±16.4	0.99 (0.97–1.00, *P*=0.132)	
P (mmol/L)	5.2±21.1	5.5±26.9	0.98 (0.98–0.99, *P*=0.052)	
K (mmol/L)	4.4±1.3	4.3±0.4	0.37 (0.27–0.50, *P*=0.083)	
Cl (mmol/L)	102.8±2.5	103.0±6.8	1.01 (0.98–1.03, *P*=0.601)	
Ca (mmol/L)	3.3±3.2	3.3±9.2	1.00 (0.98–1.02, *P*=0.915)	
AMY (U/L)	63.5±21.0	61.7±22.0	1.00 (0.99–1.00, *P*=0.172)	

Notes: The control group is defined as 0 (no event occurred), and the EGC group is defined as 1 (event occurred group). Mean ± SD/n (%). Representative with statistical significance. *P<0.05, **P<0.01, ***P<0.001.

were selected to construct diagnostic model in the training set, and were used to test the performance of the model in the validation set (Figure 3A). Finally, 26 variables were used as key variables for subsequent analysis. They were age, gender, CK-MB, CRP, Ca, Cl, LYM_Count, Na, eGFR, BASO_Count, LYM_Percent, P, RBP, ALP, AMY, CA724, GR, HDL-C, K, ApoA1, ALB, CA50, AFU, RBC, TP, and CA242. In addition, the importance ranking of the 26 variables was more consistent by using RF, XGBoost, and SVM-RFE training methods (Figure 3B). To balance the model and avoid overfitting, we further adjust the parameters of the training set by using 10-fold cross validation to determine the optimal model parameters. In the test set, XGBoost (AUC: 0.954, 95% CI: 0.915–0.994) and RF (AUC: 0.954, 95% CI: 0.926–0.996) had the highest area under curve (AUC) values compared to other methods, followed by SVM-RFE (AUC: 0.942, 95% CI: 0.912–0.965), rpart (AUC: 0.877, 95% CI: 0.813–0.885), and LGBM (AUC: 0.859, 95% CI: 0.825–0.884) (Figure 3C). A confusion matrix was used to compare the classification results and the actual measured values. The overall performance of the models was detailed in Table 2.

To further identify EGC from AGC better, we operated XGBoost, RF, and SVM models by using the 26 variables. The XGBoost model showed the best diagnostic performance with the AUC was 0.8725 (Figure 3D). Similarly, confusion matrix was used to evaluate the overall performance of the model. XGBoost performed the best: (accuracy: 0.7754, kappa: 0.5501, sensitivity: 0.8057, specificity: 0.7438, recall: 0.8057, and F1: 0.7852) (Table 3). The SHAP

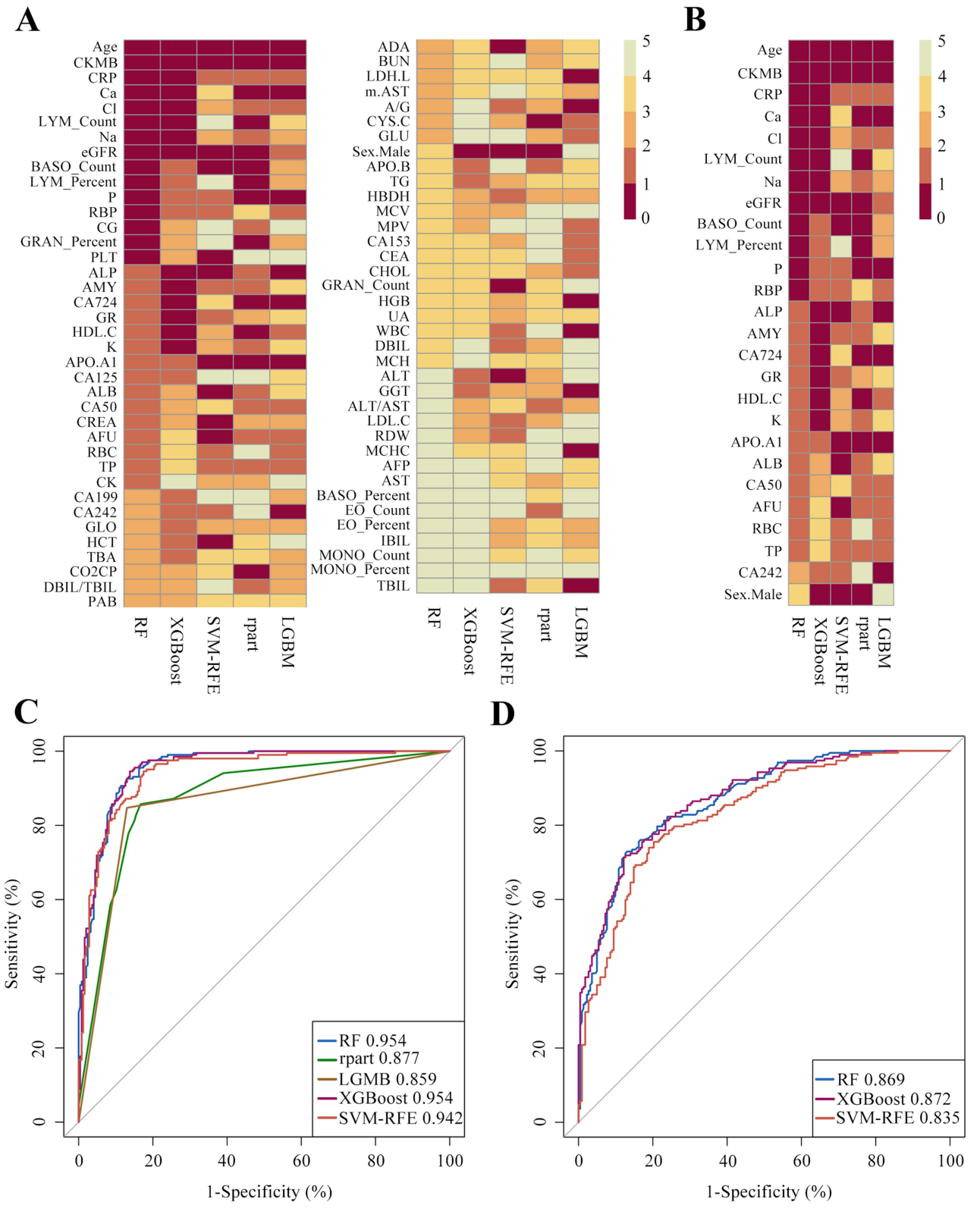

Figure 3 Machine learning models for EGC diagnosis. (**A**) The importance ranking of 75 laboratory indicators by using five machine learning models. (**B**) The importance ranking of 26 key laboratory indicators by using five machine learning models. (**C**) ROC curves of five machine learning models in the test set. The data come from all gastric cancer patients. (**D**) ROC curves of machine learning models. The data come from EGC patients.

https://doi.org/10.2147/IJGM.S559103

Table 2 The Overall Performance of the Machine Learning Models

Methods	Accuracy	Kappa	Sensitivity	Specificity	Recall	F1
XGBoost	0.8864	0.7704	0.8916	0.8800	0.8916	0.8970
RF	0.8931	0.7844	0.8984	0.8867	0.8984	0.9020
SVM-RFE	0.8597	0.7184	0.8455	0.8768	0.8455	0.8685
rpart	0.8330	0.6627	0.8462	0.8168	0.8462	0.8479
LGBM	0.8597	0.7169	0.8735	0.8431	0.8735	0.8717

Abbreviations: XGBoost, extreme gradient boosting; RF, random forest; SVM-RFE, support vector machine-recursive feature elimination; LGBM, light gradient boosting machine; rpart, recursive partitioning.

Table 3 The Overall Performance Evaluation of Three Machine Learning Models in Early Gastric Cancer Diagnosis Application

Methods	Accuracy	Kappa	Sensitivity	Specificity	Recall	F1
XGBoost	0.7754	0.5501	0.8057	0.7438	0.8057	0.7852
RF	0.7899	0.5362	0.8063	0.7708	0.8063	0.8045
SVM-RFE	0.7705	0.5401	0.7658	0.7760	0.7658	0.7816

Abbreviations: XGBoost, extreme gradient boosting; RF, random forest; SVM-RFE, support vector machine-recursive feature elimination.

analysis was also used to evaluate the importance and contribution of the 26 variables by XGBoost model (Figure 4A). The top 5 variables were GR, CA724 RBC, CA242, and ALB. DCA curve (Figure 4B) and calibration curve (Figure 4C) were used to assess model. The curve of model covered a wide range of threshold values and was located at the top of the clinically relevant threshold range, significantly higher than the Treat All and Treat Non groups. The net profit value shows a clinically significant increase at the critical threshold (Figure 4B). The calibration curve is smooth and closely follows the diagonal, and the confidence interval includes the diagonal (Figure 4C). Therefore, the performance of XGBoost model was good.

In summary, we confirmed that XGBoost, RF, and SVM-RFE models had high potential in the diagnosis of EGC through detailed multidimensional evaluation. These models not only operated effectively in disease screening scenarios with limited data, but also met the clinical demand for high accuracy and recall rate. In addition, the high AUC values of the ROC curve further validate the advantages of these models in ensuring diagnostic reliability, providing a solid scientific foundation for future clinical applications.

Construction of Model Combined with Blood Indicators and Pathological Parameters

We also collected pathological parameters in addition to blood indicators, including tumor size, tumor distribution area, degree of differentiation, histological type, lymph node metastasis status, depth of infiltration and TNM stage. The results of univariate, multivariate logistic regression analysis (Table 4) and multiple comparison corrections (Supplementary Table 3) showed that the diameter of tumor tissue less than 2 mm occurred more frequently in EGC patients, was a risk factor for the occurrence of the event ($P<0.001$). The diameter of tumor tissue in the AGC group was usually greater than 5 mm. In addition, EGC was usually in T2 phase (Table 4, and Supplementary Table 3). Combining blood indicators and pathological parameters, the AUC of XGBoost model was 0.9909 (Figure 5A). SHAP analysis showed that in XGBoost model, the top 5 variables were depth of invasion, tumor size, lymph node metastasis, GR and tumor stage (Figure 5B and C).

Discussion

The majority of gastric cancer patients in China are diagnosed in the mid to late stages. This situation called over 400,000 deaths of gastric cancer cases annually, which underscores the importance of early diagnosis in improving patient prognoses.[14] Limitations of standard diagnostic methods for gastric cancer, such as endoscopic examination, have spurred the development of more effective and precise early diagnostic technologies to enable timely therapeutic

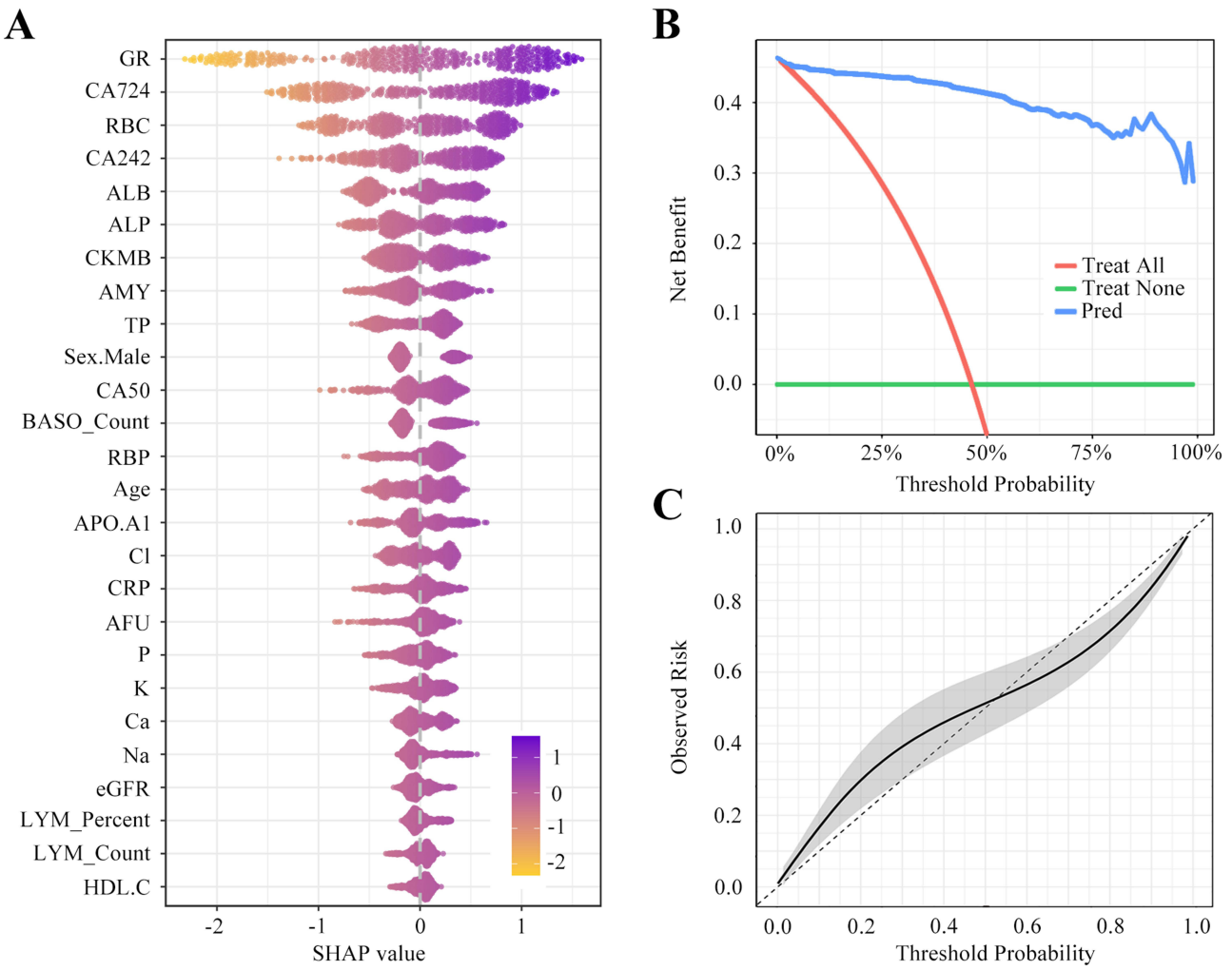

Figure 4 SHAP analysis and calibration assessed for XGBoost model. (**A**) SHAP analysis for XGBoost model. (**B**) DCA curve analysis. (**C**) Correction curve analysis.

interventions. Machine learning uses algorithms to identify patterns in data and makes predictions on new data.[15,16] This technology has become a powerful tool for solving complex medical problems in recent years. In this study, the data of 1036 patients with gastric cancer were analyzed by using machine learning algorithms to help identify EGC patients. This study integrated blood routine indicators, biochemical indicators, traditional tumor indicators, and pathological parameters, to analyze the risk factors of EGC, and establish a predictive diagnostic model for early diagnosis of gastric cancer. By deeply analyzing the biological significance of key variables, not only the clinical acceptance of the model could be improved, but clinical treatment strategies could also be optimized, thereby improving the treatment outcomes and quality of life of gastric cancer patients.

After screening out 26 key variables, our research used 5 different machine learning algorithms to classify. Compared with other models, the XGBoost algorithm had better performance and higher stability, and the AUC of the optimal model was 0.9909. In the evaluation of the importance of model variables, the top five indicators were GR, CA724, RBC, CA242, and ALB. GR is an important component of the human redox system. It uses nicotinamide adenine dinucleotide phosphate (NADPH) as a subunit to catalyze the formation of reduced glutathione (GSH) from oxidized glutathione (GSSG), which is beneficial for the body's antioxidant function.[17] At present, the clinical application of GR is mainly reported in hepatitis, cirrhosis, and metastatic liver cancer.[18] Although there was study showed that the expression of GR was significantly higher in gastric cancer tissue than in tissue adjacent to cancer.[19] But to our knowledge, it's the first time to report the relationship between the expression of GR in serum and EGC. The increased activity of GR in tumor

https://doi.org/10.2147/IJGM.S559103

Table 4 Results of Single-Factor Logistic Regression and Multivariable Logistic Regression Analysis

Name	Desc	0 (N=527)	1 (N=509)	OR (Univariable)	OR (Multivariable)
Size	Diameter 2–5 mm	49 (9.3%)	44 (8.7%)		
	Diameter <2 mm	38 (7.2%)	435 (85.5%)	12.65 (6.23–25.69, *P*<0.001)***	3.93 (1.47–10.49, *P*=0.006)**
	Diameter >5 mm	440 (83.4%)	30 (5.8%)	0.07 (0.03–0.18, *P*<0.001)***	0.55 (0.16–1.93, *P*=0.354)
Area	A: gastric antrum	33 (6.3%)	56 (11%)		
	B: gastric body	196 (37.1%)	201 (39.5%)	0.61 (0.32–1.17, *P*=0.134)	0.56 (0.13–2.29, *P*=0.416)
	BA: gastric body and gastric antrum	2 (0.4%)	3 (0.6%)	0.00 (0.00-Inf, *P*=0.992)	2697.00 (0.00-Inf, *P*=1.000)
	GA: gastric angle	1 (0.2%)	41 (8.1%)	19.89 (2.41–164.42, *P*=0.005)**	5.48 (0.15–198.23, *P*=0.353)
	GAA: gastric angle and gastric antrum	5 (0.9%)	3(0.6%)	0.00 (0.00-Inf, *P*=0.992)	1,620,225,348,001.46 (0.00-Inf, *P*=1.000)
	GEJ: gastroesophageal junction	168 (31.9%)	199 (39%)	0.69 (0.36–1.34, *P*=0.277)	0.61 (0.14–2.62, *P*=0.506)
	GEJ_B: junction of cardia and gastric body	122 (23.1%)	3 (0.6%)	0.01 (0.00–0.11, *P*<0.001)***	0.04 (0.00–0.79, *P*=0.034)*
	P: pylorus	0 (0%)	3 (0.6%)	8,182,155.12 (0.00-Inf, *P*=0.991)	182,152,711.31 (0.00-Inf, *P*=1.000)
Degree of differentiation	G1: well differentiated	231 (43.8%)	154 (30.2%)		
	G1G2: well and moderately differentiated	140 (26.6%)	80 (15.7%)	0.86 (0.51–1.44, *P*=0.559)	0.57 (0.18–1.81, *P*=0.342)
	G2: moderately differentiated	156 (29.6%)	275 (54.1%)	2.65 (1.76–3.98, *P*<0.001)***	1.22 (0.44–3.37, *P*=0.704)
Physiology	AC: adenocarcinoma	526 (99.8%)	509 (100%)		
	SCC: squamous cell carcinoma	1 (0.2%)	0 (0%)	0.00 (0.00-Inf, *P*=0.981)	0.00 (0.00-Inf, *P*=1.000)
Lymph node metastasis	No	162 (30.8%)	494 (97.1%)		
	Yes	365 (69.2%)	15 (2.9%)	0.01 (0.01–0.03, *P*<0.001)***	0.31 (0.07–1.31, *P*=0.111)
Depth of invasion	T1	14 (2.6%)	341 (66.9%)		
	T2	83 (15.9%)	168 (33.1%)	0.08 (0.04–0.16, *P*<0.001)***	0.18 (0.07–0.43, *P*<0.001)***
	T3	418 (79.3%)	0 (0%)	0.00 (0.00-Inf, *P*=0.981)	0.00 (0.00-Inf, *P*=0.990)
	T4	12 (2.3%)	0 (0%)	0.00 (0.00-Inf, *P*=0.997)	0.00 (0.00-Inf, *P*=0.999)
N_Stage	N0	152 (28.9%)	491 (96.5%)		
	N1	174 (32.9%)	18 (3.5%)	0.03 (0.01–0.07, *P*<0.001)***	0.37 (0.10–1.31, *P*=0.122)
	N2	164 (31.2%)	0 (0%)	0.00 (0.00-Inf, *P*=0.983)	0.00 (0.00-Inf, *P*=0.996)
	N3	37 (7%)	0 (0%)	0.00 (0.00-Inf, *P*=0.992)	0.00 (0.00-Inf, *P*=0.999)
M_Stage	M0	508 (96.5%)	473 (93%)		
	M1	15 (2.8%)	0 (0%)	0.00 (0.00-Inf, *P*=0.982)	0.36 (0.00-Inf, *P*=1.000)
	Mx	4 (0.7%)	36 (7%)	10.35 (2.88–37.16, *P*<0.001)***	7,298,311.75 (0.00-Inf, *P*=0.993)

Notes: The AGC group is defined as 0 (no event occurred), and the EGC group is defined as 1 (event occurred group). N (%). Representative with statistical significance. *P<0.05, **P<0.01, ***P<0.001.

cells may be related to the increased metabolic demands of tumor cells. GR helps tumor cells to resist oxidative stress and chemotherapy drug toxicity by maintaining GSH levels.[20] Further research is needed to understand the biological and clinical importance of GR in gastric cancer. CA724 and CA242 are laboratory markers for detecting gastric cancer and various digestive tract cancers, and they are also non-specific tumor markers. Their joint application with other indicators is particularly important. Additionally, RBC is an important indicator for diagnosing anemia. Anemia is a common complication of gastric cancer. A meta-analysis involving 13154 cases of gastric cancer found that approximately 36% of patients had anemia at the time of definitive diagnosis, and preoperative anemia levels were correlated with prognosis.[21] The rapid growth of malignant tumor cells consumes a large amount of nutrients in the body, and the toxic substances produced by tumor necrosis causing metabolic disorders in the body, leading to a decrease in liver cell synthesis of ALB.[22,23] The joint analysis of these testing can maximize diagnostic efficiency of gastric cancer.

As for clinical pathological parameters, the depth of invasion is a key factor in determining whether it is EGC. Although there were also cases of EGC that were limited to the mucosa and submucosa but had a larger area. Most EGC lesions had a tumor diameter of less than 2 mm. Taking into account serological indicators and pathological parameters, the serum expression level of GR also ranks among the top 5, indicating its importance in diagnosis.

This study demonstrates the feasibility of using machine learning models based on laboratory indicators for early diagnosis of gastric cancer. Through machine learning technology, exploring the implicit connections between laboratory

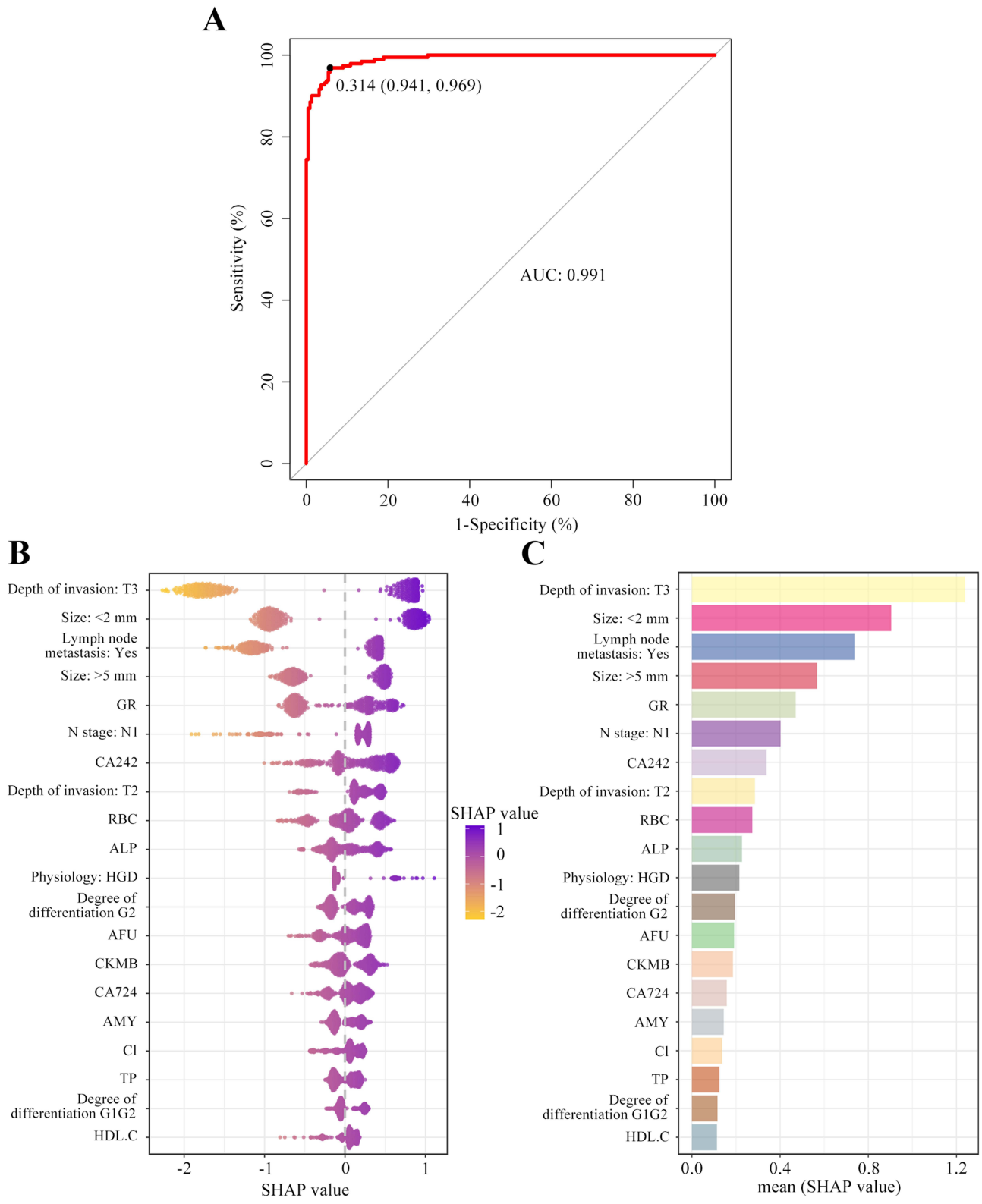

Figure 5 Construction of model combined with blood indicators and pathological parameters. (**A**) ROC curves of XGBoost model combining blood indicators and clinical pathological parameters. (**B**) Scatter plot presenting SHAP analysis for XGBoost model combining blood indicators and pathological parameters. (**C**) Bar chart presenting SHAP analysis of XGBoost model combined with blood indicators and pathological parameters.

https://doi.org/10.2147/IJGM.S559103

indicators and diseases can undoubtedly provide more valuable diagnostic opinions for clinical practice. However, it was a single-center design, absence of external validation, and circularity of using post-surgical pathological parameters were downplayed. The next step of the study needs to conduct multi-center prospective research for external verification to further improve this machine learning model. We will further develop an automatic clinical scoring system based on nomograms or machine learning in order to provide clinicians with more practical and easy-to-understand tools.

Conclusion

In summary, we compared the performance of five machine learning algorithms, the XGBoost presented the best performance. The model included 26 characteristic variables, and the top five indicators were GR, CA724, RBC, CA242, and ALB. Although the clinical practicality needs to be proven, this study provides confirmatory data support for the preclinical implementation of the model.

Institutional Review Board Statement

This study was approved by the Ethics Committee of The Affiliated People's Hospital of Jiangsu University (No. K-20220148-Y). All the study procedures were performed in accordance with the tenets of the Declaration of Helsinki.

Data Sharing Statement

Technical appendix, statistical code, and dataset available upon reasonable request to the corresponding author at runbiji@163.com.

Informed Consent Statement

Before formally entering the study, all participants signed written informed consent forms.

Acknowledgments

Special thanks to Canbiao Wang (RealThinking Biotechnologies, Getang Avenue 11, Nanjing, 210002, China) for making significant contributions to the data statistics in this article.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This work was supported by Zhenjiang Science & Technology Program (JC2024030), Clinical research project funded by The Affiliated People's Hospital of Jiangsu University (Y2022007-Z).

Disclosure

The authors declare that they have no conflicts of interest in this work.

References

1. Yasuda T, Wang YA. Gastric cancer immunosuppressive microenvironment heterogeneity: implications for therapy development. *Trends Cancer*. 2024;10(7):627–642. doi:10.1016/j.trecan.2024.03.008
2. Smyth EC, Nilsson M, Grabsch HI, van Grieken NC, Lordick F. Gastric cancer. *Lancet*. 2020;396(10251):635–648. doi:10.1016/S0140-6736(20)31288-5
3. Tian H, Ning Z, Zong Z, et al. Application of machine learning algorithms to predict lymph node metastasis in early gastric cancer. *Front Med Lausanne*. 2022;8:759013. doi:10.3389/fmed.2021.759013
4. Conti CB, Agnesi S, Scaravaglio M, et al. Early Gastric Cancer: update on prevention, diagnosis and treatment. *Int J Environ Res Public Health*. 2023;20(3):2149. doi:10.3390/ijerph20032149

5. Yao K, Uedo N, Kamada T, et al. Guidelines for endoscopic diagnosis of early gastric cancer. *Dig Endosc*. 2020;32(5):663–698. doi:10.1111/den.13684
6. Wang Z, Wu Q. Advancements in non-invasive diagnosis of gastric cancer. *World J Gastroenterol*. 2025;31(6):101886. doi:10.3748/wjg.v31.i6.101886
7. Ma XZ, Zhou N, Luo X, Guo SQ, Mai P. Update understanding on diagnosis and histopathological examination of atrophic gastritis: a review. *World J Gastrointest Oncol*. 2024;16(10):4080–4091. doi:10.4251/wjgo.v16.i10.4080
8. Ma S, Zhou M, Xu Y, et al. Clinical application and detection techniques of liquid biopsy in gastric cancer. *Mol Cancer*. 2023;22(1):7. doi:10.1186/s12943-023-01715-z
9. Li Y, Wang JS, Guo Y, Zhang T, Li LP. Use of the alkaline phosphatase to prealbumin ratio as an independent predictive factor for the prognosis of gastric cancer. *World J Gastroenterol*. 2020;26(44):6963–6978. doi:10.3748/wjg.v26.i44.6963
10. Chen Y, Wang B, Zhao Y, et al. Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer. *Nat Commun*. 2024;15(1):1657. doi:10.1038/s41467-024-46043-y
11. Du H, Yang Q, Ge A, Zhao C, Ma Y, Wang S. Explainable machine learning models for early gastric cancer diagnosis. *Sci Rep*. 2024;14(1):17457. doi:10.1038/s41598-024-67892-z
12. Spanish EURECCA Esophagogastric Cancer Group, Pera M, Gibert J, Gimeno M, et al. Machine learning risk prediction model of 90-day mortality after gastrectomy for cancer. *Ann Surg*. 2022;276(5):776–783. doi:10.1097/SLA.0000000000005616
13. Wang Z, Gu Y, Huang L, et al. Construction of machine learning diagnostic models for cardiovascular pan- disease based on blood routine and biochemical detection data. *Cardiovasc Diabetol*. 2024;23(1):351. doi:10.1186/s12933-024-02439-0
14. Zhu H, Wang G, Zheng J, et al. Preoperative prediction for lymph node metastasis in early gastric cancer by interpretable machine learning models: a multicenter study. *Surgery*. 2022;171(6):1543–1551. doi:10.1016/j.surg.2021.12.015
15. Seo JW, Park KB, Lim ST, Jun KH, Chin HM. Machine learning models for prediction of lymph node metastasis in patients with T1b gastric cancer. *Am J Cancer Res*. 2024;14(8):3842–3851. doi:10.62347/KREL8138
16. Wang Z, Liu Y, Niu X. Application of artificial intelligence for improving early detection and prediction of therapeutic outcomes for gastric cancer in the era of precision oncology. *Semin Cancer Biol*. 2023;93:83–96. doi:10.1016/j.semcancer.2023.04.009
17. Xu FL, Wu XH, Chen C, et al. SLC27A5 promotes sorafenib-induced ferroptosis in hepatocellular carcinoma by downregulating glutathione reductase. *Cell Death Dis*. 2023;14(1):22.
18. Hong KS, Pagan K, Whalen W, et al. The role of glutathione reductase in influenza infection. *Am J Respir Cell Mol Biol*. 2022;67(4):438–445. doi:10.1165/rcmb.2021-0372OC
19. Rzymowska J. The effect of low temperature on the enzyme activities and the level of SH groups in benign gastric ulcer and gastric carcinoma. *J Pharm Pharmacol*. 1994;46(6):517–518. doi:10.1111/j.2042-7158.1994.tb03841.x
20. Brzozowa-Zasada M, Piecuch A, Bajdak-Rusinek K, et al. Glutathione reductase expression and its prognostic significance in colon cancer. *Int J Mol Sci*. 2024;25(2):1097. doi:10.3390/ijms25021097
21. Huang XZ, Yang YC, Chen Y, et al. Preoperative anemia or low hemoglobin predicts poor prognosis in gastric cancer patients: a Meta-Analysis. *Dis Markers*. 2019;2019:7606128. doi:10.1155/2019/7606128
22. Aday U, Tatlı F, Akpulat FV, et al. Prognostic significance of pretreatment serum lactate dehydrogenase-to-albumin ratio in gastric cancer. *Contemp Oncol*. 2020;24(3):145–149. doi:10.5114/wo.2020.100219
23. Gupta D, Lis CG. Pretreatment serum albumin as a predictor of cancer survival: a systematic review of the epidemiological literature. *Nutr J*. 2010;9(1):69. doi:10.1186/1475-2891-9-69