scientific reports

Check for updates

OPEN

An ensemble learning model to predict lymph node metastasis in early gastric cancer

Kaiqing Song[1], Jiaming Wu[1], Muchen Xu[2,3], Mengying Li[1], Yuqi Chen[1], Yi Zhang[4,5], Hong Chen[4]✉ & Caifeng Jiang[1]✉

Lymph node metastasis is a critical factor for determining therapeutic strategies and assessing the prognosis of early gastric cancer. This work aimed to establish a more dependable predictive model for identify lymph node metastasis in early gastric cancer. The study utilized both univariate and multivariate logistic regression analyses to identify independent risk factors for lymph node metastasis of early gastric cancer, while employing five distinct algorithms to calculate feature weights. The optimal feature combination for each algorithm model was determined by combining the six highest weight features from all five models along with the independent risk factors. An ensemble learning model was subsequently constructed by integrating these five models. The model's performance was evaluated by the AUC, accuracy, and F1 score. Following this, a threshold was determined based on the F1 score, and the model's performance was assessed using an external validation set. The lymph node metastasis rate of early gastric cancer in our study was 16.4%. The ensemble learning model achieved an AUC value of 0.860 in the test set, with an accuracy of 82.35% and an F1 score of 0.611. Based on the F1 score, the model's threshold was set at 0.18. Additionally, the model demonstrated an AUC of 0.892 in the external validation set, along with an accuracy of 78.30% and an F1 score of 0.60.We constructed an ensemble learning model for predicting lymph node metastasis of early gastric cancer. Gastric surgery should be considered as the preferred treatment when the risk of lymph node metastasis exceeds 18%.

Keywords Ensemble learning model, Lymph node metastasis, Early gastric cancer

Abbreviations

EGC	Early gastric cancer
LNM	Lymph node metastasis
RF	Random forest
SVM	Support vector machine
CART	Classification and regression tree
LR	Logistic regression
XGB	Extreme gradient boosting
ER	Endoscopic resection

The latest statistical data underscores the worldwide incidence of gastric cancer, placing it fifth in terms of new occurrences and fourth in terms of cancer-related deaths[1]. It is worth noting that China constitutes a significant proportion, with 43.9% of newly diagnosed cases and 48.6% of associated deaths. Unfortunately, the early detection rate of gastric cancer in China is low, with over 80% of patients being diagnosed at an advanced stage[2].

Early gastric cancer (EGC), which stands for gastric cancer with lesions limited to the mucosa and submucosa, regardless of the presence of lymph node metastasis (LNM)[3]. The incidence of lymph node metastasis (LNM) in early gastric cancer (EGC) is estimated to be between 12.3% and 20.7% according to statistics[4–7]. Patients without lymph node metastasis (LNM) can achieve an impressive 5-year survival rate of over 90%, while once

[1]Department of Gastroenterology, The Fourth Affiliated Hospital of Soochow University, Suzhou, China. [2]Department of Radiotherapy, The Fourth Affiliated Hospital of Soochow University, Suzhou, China. [3]Department of Radiotherapy, The Affiliated Hospital of Xuzhou Medical University, Xuzhou, China. [4]Department of General Surgery, The Fourth Affiliated Hospital of Soochow University, Suzhou, China. [5]Department of General Surgery, The First Affiliated Hospital of Soochow University, Suzhou, China. ✉email: Hongchen_506@126.com; cfjiang1999@163.com

LNM occurs, the overall survival rate of node-positive early gastric cancer (EGC) will decrease to 70–80%, and the recurrence rate will increase to 8%[4,8–10]. The primary treatment modalities for EGC are endoscopic resection (ER) and gastric surgery, as indicated by various studies[11,12]. Endoscopic resection (ER) is preferred due to its minimally invasive nature, high safety profile, and superior postoperative quality of life[13,14]. For tumors with a very low likelihood of LNM which are amenable to en bloc resection, the 2018 Japanese gastric cancer treatment guidelines recommend endoscopic resection (ER), while gastric surgery is much preferred for EGC with LNM[15]. Therefore, accurate evaluation of LNM status in EGC is crucial to avoid unnecessary surgical procedures and predict patient prognosis.

In recent years, numerous studies have constructed prediction models for LNM of EGC. These models are primarily based on various algorithms, such as logistic models, nomograms, decision trees, and naive Bayes methods, among others. The AUC values of these models range from 0.690 to 0.865[5,16–20]. However, the incidence of LNM of EGC is relatively low, resulting in an imbalanced distribution of positive and negative samples in the dataset. This imbalanced data distribution can reduce the performance of traditional classification algorithms[21]. Relevant studies have demonstrated that utilizing ensemble learning techniques to process this type of dataset can yield positive outcomes[22]. In this work, we employed ensemble learning technology to construct a more dependable prediction model for LNM of EGC. This model provides a crucial foundation for guiding clinical rational selection of therapeutic strategies and assessment of prognosis.

Materials and methods

Data source and processing

We conducted a thorough search for gastrectomy cases between 2014 and 2023 at the First Affiliated Hospital of Soochow University, the Fourth Affiliated Hospital of Soochow University, and the Affiliated Hospital of Xuzhou Medical University. After screening, a total of 1423 cases with postoperative pathological diagnosis of EGC were selected. Among them, 1189 cases from the First Affiliated Hospital of Soochow University and the Fourth Affiliated Hospital of Soochow University were randomly divided into a training set and a test set in a 7:3 ratio. Additionally, data from 234 cases from the Affiliated Hospital of Xuzhou Medical University were utilized as the external validation set. This research was carried out in accordance with the Declaration of Helsinki and approved by the Medical Ethics Committee of Suzhou Dushu Lake Hospital (no.231004).

The enrollment of patients was based on the following inclusion criteria: firstly, the tumor must be in pT1 stage, and secondly, there must be more than 15 lymph node dissections. Conversely, the exclusion criteria were as follows: firstly, the presence of other primary tumors, secondly, the use of neoadjuvant therapy, and thirdly, the existence of other pathological types of gastric cancer, such as lymphoma, gastric stromal tumor, neuroendocrine tumors, and so on.

The comprehensive clinical and pathological features encompass various factors such as age, gender, tumor size, tumor location, depth of invasion, tumor histologic type, vascular invasion, neural invasion, ulceration, signet ring cell, macroscopic type, and histopathological type. The tumor location is further categorized into upper, middle, and lower regions, with the upper part comprising the esophageal gastric junction and gastric fundus, the middle part being the gastric body, and the lower part including the gastric horn and gastric antrum. As per the Japanese guide[15,23], macroscopic types are classified into elevated, flat, and depressed, while tumor histologic types are differentiated and undifferentiated. The histopathological type types are further categorized into mucinous adenocarcinoma, papillary adenocarcinoma, tubular adenocarcinoma, and signet ring cell carcinoma.

Model construction and statistical analysis

The study employed univariate analysis and multivariate logistic regression analysis to identify independent risk factors for lymph node metastasis (LNM) of early gastric cancer (EGC). Additionally, five algorithm models, namely random forest (RF), support vector machine (SVM), classification and regression tree (CART), logistic regression (LR), and extreme gradient boosting (XGB), were utilized to compute the feature weights in their respective models.

After considering the six features with higher weights in various models and independent risk factors for LNM of EGC, a total of ten features were selected for the subsequent feature screening process. These selected features were then randomly combined and substituted into five models to calculate the AUC value for each combination. The combination that yielded the highest AUC value was deemed the optimal feature combination for this algorithm, and as such, five models were constructed accordingly.

Through the utilization of a soft voting approach[24], we have amalgamated these 5 models to construct an ensemble learning model. The test set was utilized to intercept the threshold, and validated on the external validation set. The model's performance was evaluated by calculating the AUC value, accuracy, and F1 score. The flowchart depicting the process is illustrated in Fig. 1.

The statistical analyses were performed utilizing the Python programming language. The comparison between the two groups was executed by means of t-tests for continuous variables and chi square tests for categorical variables. A statistical significance level of $P < 0.05$ was deemed to be significant.

Results

Demographic and clinicopathological features

This study included a total of 1423 patients diagnosed with EGC, with a recorded LNM rate of 16.4% (233/1190). The prevalence of EGC was found to be higher in men (68.1%) than in women (31.9%), while the LNM rate in female patients (18.1%) was slightly higher than that in men (15.6%). Significant differences were observed between the LNM (-) group and LNM (+) group in terms of age, tumor size, tumor location, depth of invasion,

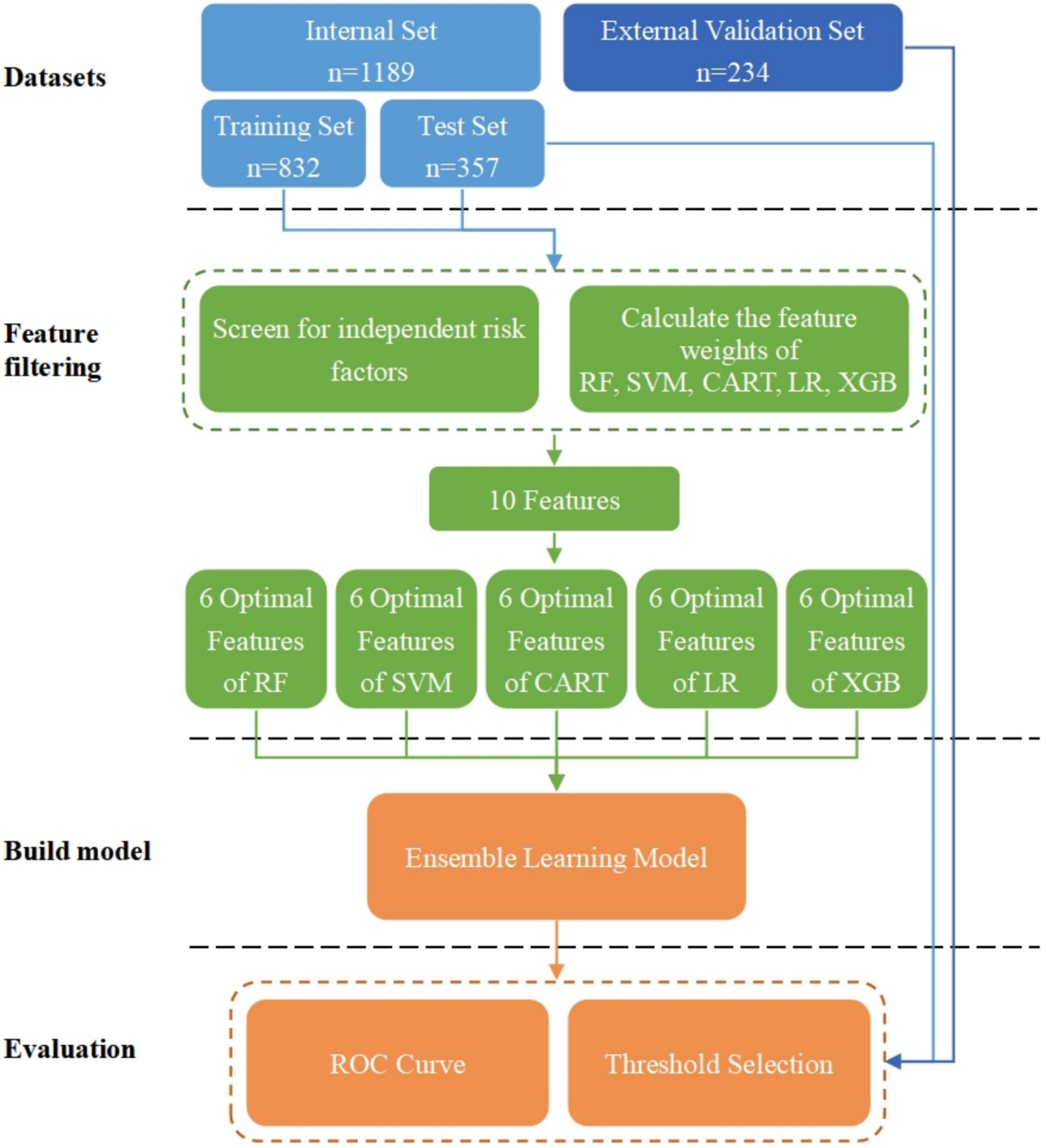


Fig. 1. Flowchart of the whole work.

tumor histologic type, vascular invasion, neural invasion, ulceration, and macroscopic type (Table 1). However, no statistical difference was found between gender, signet ring cell, and histopathological type. The demographic and clinicopathological features were showed in Table 1.

Risk factors of LNM in univariate and multivariate analyses

Table 2 present the risk factors of LNM in univariate and multivariate analyses. The results of the univariate analysis indicate that several factors, including age, tumor size, tumor location, depth of invasion, tumor histologic type, vascular invasion, neural invasion, ulceration, and macroscopic type, are significantly associated with LNM ($P<0.05$). However, gender, signet ring cell, and histopathological type do not show any significant differences between the LNM(-) group and LNM(+) group ($P>0.05$). The multivariate analysis further reveals

Features	Classification	All patients	LNM (−)	LNM (+)	*P*
N		1423	1190 (83.6)	233 (16.4)	
Gender	Male	969 (68.1)	818 (68.7)	151 (64.8)	0.271
	Female	454 (31.9)	372 (31.3)	82 (35.2)	
Age, median (range)		63 (25,91)	63 (26,91)	63 (25,88)	<0.001
Tumor size	>2 cm	566 (40.0)	396 (33.3)	170 (73.0)	<0.001
	≤2 cm	857 (60.0)	794 (66.7)	63 (27.0)	
Tumor location	Upper	265 (18.6)	234 (19.7)	31 (13.3)	<0.001
	Middle	423 (29.7)	369 (31.0)	54 (23.2)	
	Lower	735 (51.7)	587 (49.3)	148 (63.5)	
Depth of invasion	Mucosa	658 (46.2)	619 (52.0)	39 (16.7)	<0.001
	Submucosa	765 (53.8)	571 (48)	194 (83.3)	
Tumor histologic type	Differentiated	703 (49.4)	634 (53.3)	69 (29.6)	<0.001
	Undifferentiated	720 (50.6)	556 (46.7)	164 (70.4)	
Vascular invasion	Absent	1337 (94.0)	1163 (97.7)	174 (74.7)	<0.001
	Present	86 (6.0)	27 (3.3)	59 (25.3)	
Neural invasion	Absent	1417 (99.6)	1188 (99.8)	229 (98.3)	0.005
	Present	6 (0.4)	2 (0.2)	4 (1.7)	
Ulceration	Absent	694 (48.8)	595 (50.0)	99 (42.5)	0.043
	Present	729 (51.2)	595 (50.0)	134 (57.5)	
Signet ring cell	Absent	1154 (81.1)	975 (81.9)	179 (76.8)	0.084
	Present	269 (18.9)	215 (18.1)	54 (23.3)	
Macroscopic type	Elevated	152 (10.7)	123 (10.3)	29 (12.4)	<0.001
	Flat	771 (54.2)	690 (58.0)	81 (34.8)	
	Depressed	500 (35.1)	377 (31.7)	123 (52.8)	
Histopathological type	Signet ring cell carcinoma	104 (7.3)	84 (7.1)	20 (8.6)	0.740
	Mucinous adenocarcinoma	30 (2.1)	24 (2.0)	6 (2.6)	
	Papillary adenocarcinoma	9 (0.6)	7 (0.6)	2 (0.8)	
	Tubular adenocarcinoma	1280 (90.0)	1075 (90.3)	205 (88.0)	

Table 1. Clinicopathologic characteristics of cohort in patients with EGC by LNM. LNM(−): lymph node metastasis (negative), LNM(+): lymph node metastasis (positive), significant at $P = 0.05$.

that tumor size, depth of invasion, tumor histologic type, and vascular invasion are independent risk factors for LNM of EGC. Notably, submucosal tumors larger than 2 cm, undifferentiated, and accompanied by vascular invasion are more likely to undergo LNM.

Feature importance of different algorithm model

Figure 2 illustrates the six most prominent attributes among the algorithm weights of five distinct types. In conjunction with the independent risk factors of LNM, a total of ten characteristics were initially selected for the subsequent phase of feature screening. These characteristics encompass age, gender, tumor size, depth of invasion, tumor histologic type, vascular invasion, macroscopic type_elevated, macroscopic type_depressed, neural invasion, and tumor location_lower.

Select the optimal feature combination

By randomly combining 10 features, the total number of combinations that can be obtained is 1023. The calculation formula is as follows.

$$\sum_{m=1}^{10} \frac{10!}{m! \times (10-m)!} = 1023$$

A total of five algorithms were utilized in constructing models for 1023 feature combinations, resulting in the calculation of 5115 AUC values. The ROC curves for each of the five algorithm models, featuring the optimal feature combination, are depicted in Fig. 3. Notably, the optimal feature combination for SVM (AUC = 0.838) and LR (AUC = 0.856) includes tumor size, tumor histologic type, depth of invasion, vascular invasion, age, and tumor location_lower. Similarly, the optimal feature combination for RF (AUC = 0.867), CART (AUC = 0.858), and XGB (AUC = 0.862) includes tumor size, tumor histologic type, depth of invasion, vascular invasion, gender, and tumor location lower.

Variables	Classification	Univariate analysis	Multivariate analysis		
		P Value	OR	95%CI	P值
Gender	Male	0.079			
	Female				
Age, median (range)		<0.001	0.41	0.13–1.29	0.126
Tumor size	>2 cm	<0.001	4.57	3.14–6.67	<0.001
	≤2 cm				
Tumor location	Upper				
	Middle	0.053	0.85	0.48–1.53	0.595
	Lower	<0.001	1.57	0.92–2.67	0.095
Depth of invasion	Mucosa	<0.001	0.29	0.18–0.45	<0.001
	Submucosa				
Tumor histologic type	Differentiated	<0.001	0.50	0.34–0.74	0.001
	Undifferentiated				
Vascular invasion	Absent	<0.001	5.80	3.18–10.58	<0.001
	Present				
Neural invasion	Absent	0.035	4.33	0.53–35.44	0.172
	Present				
Ulceration	Absent	0.024	0.93	0.54–1.59	0.780
	Present				
Signet ring cell	Absent	0.057			
	Present				
Macroscopic type	Elevated				
	Flat	<0.001	0.66	0.36–1.22	0.186
	Depressed	<0.001	1.73	0.80–3.71	0.162
Histopathological type	Signet ring cell carcinoma	1.000			
	Mucinous adenocarcinoma	0.755			
	Papillary adenocarcinoma	0.679			
	Tubular adenocarcinoma	0.642			

Table 2. Univariate and multivariate logistic regression analysis of risk factors for LNM in patients with EGC. OR, Odds ratio, CI, Confidence interval, significant at $P = 0.05$.

Build an ensemble learning model

Aforementionedly, we conducted a screening process to identify the optimal feature combinations for five algorithms. Then we developed five models based on these combinations and integrated them to create an ensemble learning model. The ensemble model uses a set probability threshold (0.3) to convert the predicted probabilities into binary classification prediction values. Finally, it is integrated through a soft - voting strategy. Specifically, if at least three base models predict the positive class, the prediction result of the ensemble model is also the positive class. The AUC value of the test set was determined to be 0.860. The model demonstrated strong performance when the threshold was set at 0.18, achieving an accuracy of 82.35%, precision of 50.77%, recall of 76.74%, specificity of 83.59%, and an F1 score of 0.611, as depicted in Fig. 4.

External validation set

Ultimately, the ensemble learning model was implemented on an external validation set, yielding an AUC value of 0.892. Upon setting the threshold to 0.18, the accuracy of the external validation set was determined to be 78.30%, with precision at 46.34%, recall at 84.44%, specificity at 76.84%, and an F1 score of 0.60, as depicted in Fig. 5. It is evident that the ensemble learning model holds significant predictive value for assessing the risk of LNM in the external validation set.

Discussion

We have successfully developed a highly dependable model for predicting lymph node metastasis (LNM) of early gastric cancer (EGC). The LNM rate of EGC in this study was found to be 16.4%, which is consistent with the existing reported lymph node metastasis rate of 12.3–20.7%[4–7]. Gotoda and Sekiguchi et al.[25,26] have discovered that tumor size, invasion depth, ulceration, histological type, and vascular invasion are the primary risk factors for LNM of EGC. Among these factors, tumor size, depth of invasion, ulceration, and histological type are considered important reference indicators in the Japanese gastric cancer treatment guidelines[11]. In our study, the independent risk factors for LNM of EGC were found to be tumor size, depth of invasion, tumor histologic type, and vascular invasion, which is consistent with previous studies. It is worth noting that submucosal tumors larger than 2 cm, undifferentiated, and accompanied by vascular invasion are more likely to undergo LNM.

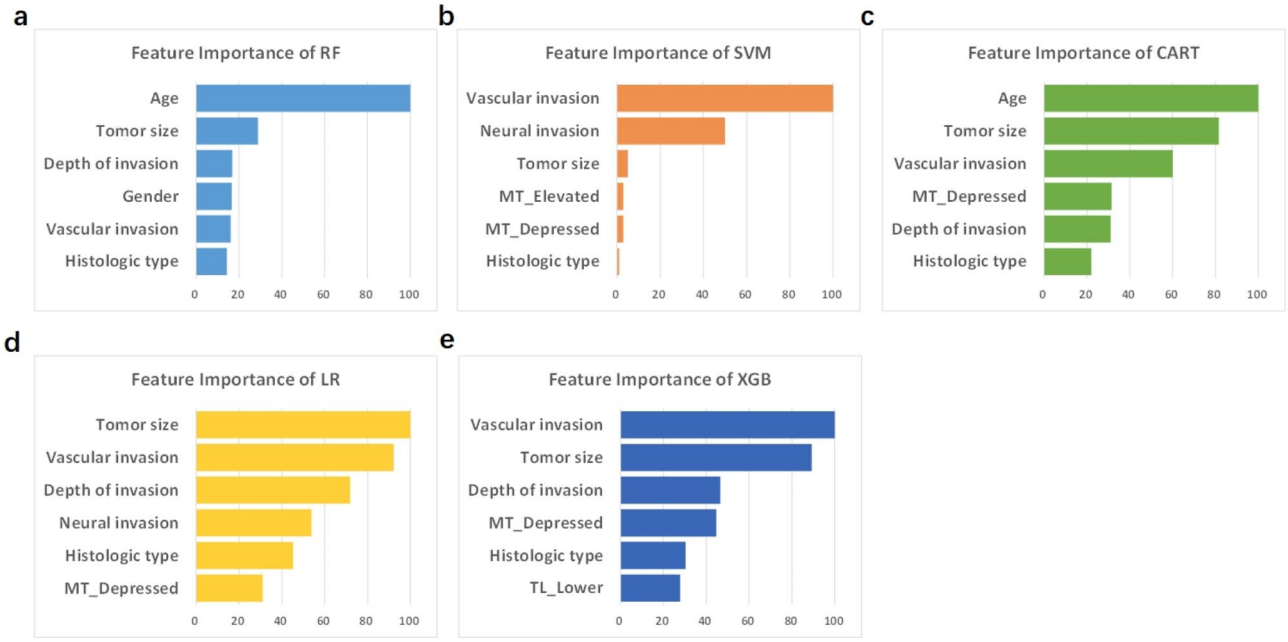


Fig. 2. The feature importance of different algorithm model. (**A**) Feature importance of RF. (**B**) Feature importance of SVM. (**C**) Feature importance of CART. (**D**) Feature importance of LR. (**E**) Feature importance of XGB. MT, Macroscopic type, TL, Tumor location.

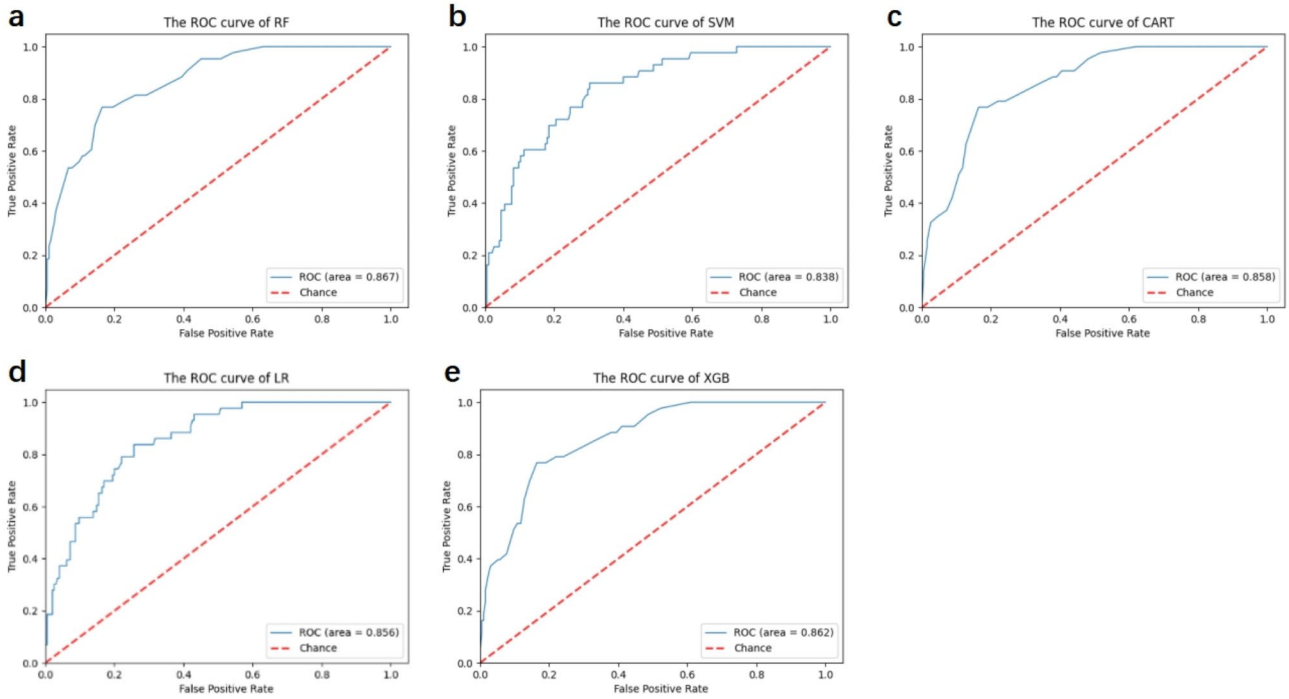


Fig. 3. The ROC curves for five algorithm models with the optimal feature combination. (**A**) The ROC curve of RF. (**B**) the ROC curve of SVM. (**C**) the ROC curve of CART. (**D**) The ROC curve of LR. (**E**) The ROC curve of XGB.

Machine learning has emerged as a potent tool in various domains, including medicine[27]. It is commonly utilized for medical diagnosis, treatment optimization, disease prediction, and patient management in the medical field[28]. Zhao et al.[16] employed nomogram, decision tree, and deep learning models to predict LNM in patients with EGC and determined that the FCNN model (AUC = 0.790) exhibited the most accurate prediction of LNM risk. Zhu et al.[29] conducted a comparison of six machine learning algorithms and discovered that

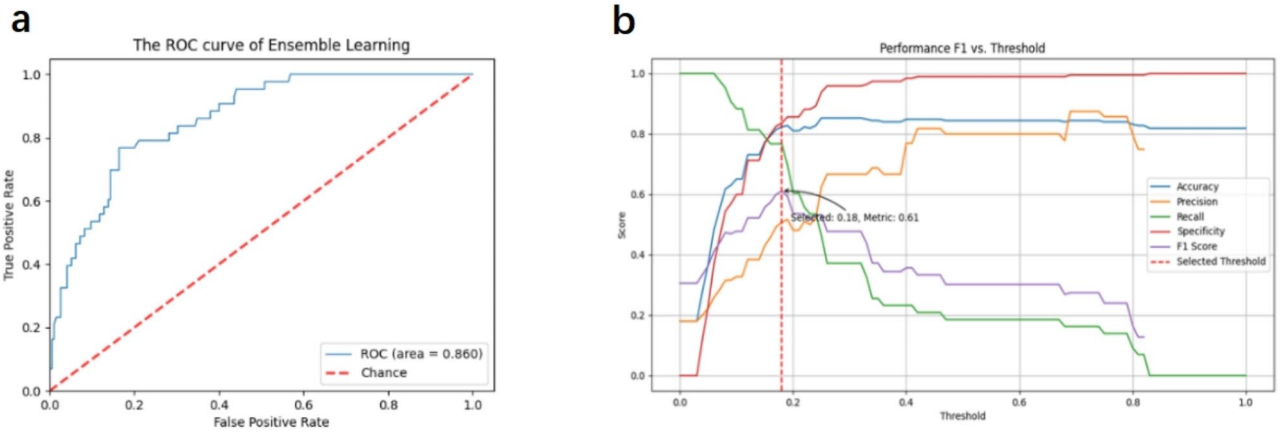
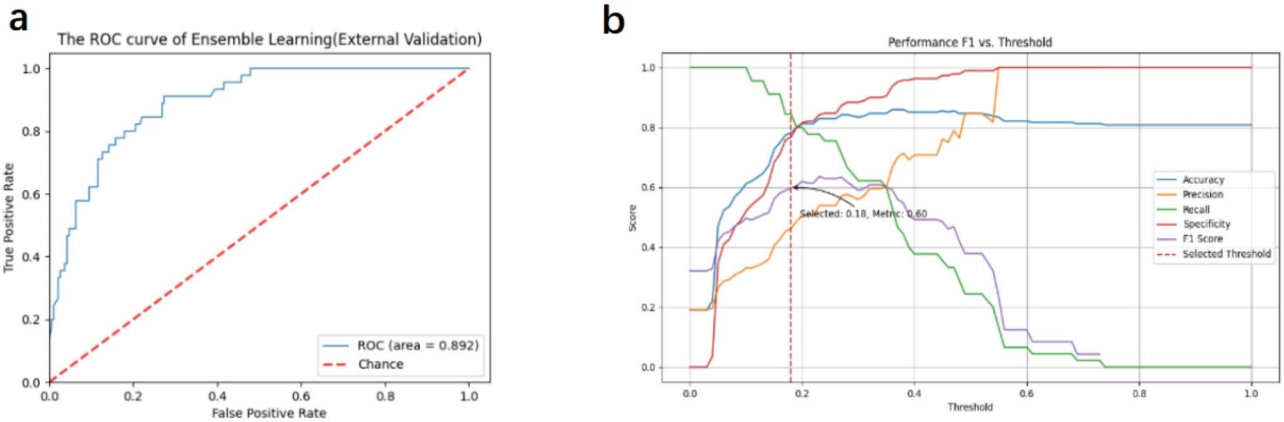


Fig. 4. (**A**) the ROC curve of ensemble learning (test set). (**B**) Performance metrics of the model under different thresholds (test set).



(**A**) The ROC curve of ensemble learning (external validation set). (**B**) Performance metrics of the model under different thresholds (external validation set).

XGBOOST can predict the risk of LNM with the highest accuracy in patients with EGC. In this study, we utilized five algorithms to construct an ensemble learning model, which performed exceptionally better in both the test set (AUC = 0.860) and the external validation set (AUC = 0.892).

The current prediction models for LNM in EGC predominantly depend on a singular algorithmic model or opt for the model with the highest AUC value through comparisons among various models[16,29,30]. However, the dataset we studied exhibits an imbalanced distribution, which negatively impacts the model's predictive ability. To address such datasets, techniques such as ensemble learning, cost-sensitive learning, category weights, and resampling methods, among others, can substantially improve the model's performance[31,32]. In this study, we utilized ensemble learning techniques to augment the model's generalization and stability. In addition, we employed evaluation metrics such as AUC-ROC, precision, recall, and F1 score to evaluate the model's performance, as these metrics are particularly suitable for handling imbalanced data. Based on the F1 score, we established the threshold for the test set at 0.18. The model displayed an AUC of 0.892, an accuracy of 78.30%, a precision of 46.34%, a recall of 84.44%, a specificity of 76.84%, and an F1 score of 0.60 in the external training set. The final prediction model exhibits exceptional generalization and stability.

During the model building process, the selection of appropriate features is of utmost significance. Prior research has primarily focused on computing feature weights or utilizing independent risk factors as features for model construction[29,33]. In contrast, our study takes into account both independent risk factors and feature weights. We conducted a series of 5115 calculations across five models by randomly combining the initially screened features. This methodology allowed us to determine the optimal feature combination for each algorithmic model, thus enabling us to make a more informed decision regarding the modeling features.

This study is subject to several limitations. Firstly, it is a retrospective study, which may have introduced a certain degree of selection bias. Secondly, the dataset used is relatively limited, and the sample size is not sufficiently large. To enhance the stability and generalizability of the model, a larger and more diverse dataset would be beneficial. Additionally, it should be noted that the observed LNM rate in this study may be higher

than the actual rate due to the non-inclusion of the non-surgical group in the analysis. Lastly, the clinical characteristics considered in the study were not comprehensive, and important factors such as molecular markers and imaging data were not included.

Conclusion

To summarize, this work employed an ensemble learning approach to construct a robust prediction model for lymph node metastasis in early gastric cancer. The model demonstrated a high level of accuracy, achieving an AUC value of 0.860 on the test set and an AUC value of 0.892 on the external validation set. Based on the findings, surgical intervention is recommended for patients with a risk of lymph node metastasis exceeding 18%.

Data availability

The data are available from the corresponding author, [cfjiang1999@163.com], upon reasonable request.

Received: 25 February 2024; Accepted: 24 March 2025
Published online: 02 April 2025

References

1. Sung, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J. Clin.* **71** (3), 209–249 (2021).
2. Zong, L. et al. The challenge of screening for early gastric cancer in China. *Lancet* **388** (10060), 2606 (2016).
3. Japanese Research Society for Gastric Cancer & Kajitani, T. The general rules for the gastric cancer study in surgery and pathology: Part I. Clinical classification. *Jpn. J. Surg.* **11** (2), 127–139 (1981).
4. Li, X. et al. The characteristics, prognosis, and risk factors of lymph node metastasis in early gastric cancer. *Gastroenterol. Res. Pract.* **2018** (2018).
5. Zhang, M. et al. A nomogram to predict risk of lymph node metastasis in early gastric cancer. *Sci. Rep.* **11** (1), 22873 (2021).
6. Wang, Z. K. et al. Higher risk of lymph node metastasis in young patients with early gastric cancer. *J. Cancer* **10** (18), 4389 (2019).
7. Chen, J., Zhao, G. & Wang, Y. Analysis of lymph node metastasis in early gastric cancer: A single institutional experience from China. *World J. Surg. Oncol.* **18** (1), 1–7 (2020).
8. Saka, M., Katai, H., Fukagawa, T., Nijjar, R. & Sano, T. Recurrence in early gastric cancer with lymph node metastasis. *Gastric Cancer*. **11**, 214–218 (2008).
9. Kikuchi, S. et al. Surgical outcome of node-positive early gastric cancer with particular reference to nodal status. *Anticancer Res.* **20**, 3695–3700 (2000).
10. Kim, H. et al. Lymphovascular invasion is an important predictor of lymph node metastasis in endoscopically resected early gastric cancers. *Oncol. Rep.* **25**, 1589–1595 (2011).
11. Japanese Gastric Cancer Association jgca@ koto.kpu-m.ac.jp. JJapanese gastric cancer treatment guidelines 2014 (ver. 4). *Gastric Cancer* **20** (1), 1–19 (2017).
12. Borie, F. et al. Appropriate gastric resection with lymph node dissection for early gastric cancer. *Ann. Surg. Oncol.* **11**, 512–517 (2004).
13. Pimentel-Nunes, P. et al. Endoscopic submucosal dissection: European society of Gastrointestinal endoscopy (ESGE) guideline. *Endoscopy* **47** (09), 829–854 (2015).
14. Tae, C. H. et al. Comparison of subjective quality of life after endoscopic submucosal resection or surgery for early gastric cancer. *Sci. Rep.* **10** (1), 6680 (2020).
15. Japanese Gastric Cancer Association jgca@ koto.kpu-m.ac.jp. Japanese gastric cancer treatment guidelines 2018. *Gastric Cancer* **24** (1), 1–21 (2021).
16. Zhao, L. et al. Using nomogram, decision tree, and deep learning models to predict lymph node metastasis in patients with early gastric cancer: A multi-cohort study. *Am. J. Cancer Res.* **13** (1), 204–215 (2023).
17. Pyo, J. H. et al. A risk prediction model based on lymph-node metastasis in poorly differentiated–type intramucosal gastric cancer. *PLoS One* **11** (5), e0156207 (2016).
18. Li, J. et al. Dual-energy CT–based deep learning radiomics can improve lymph node metastasis risk prediction for gastric cancer. *Eur. Radiol.* **30**, 2324–2333 (2020).
19. Kim, S. M. et al. Nomogram to predict lymph node metastasis in patients with early gastric cancer: A useful clinical tool to reduce gastrectomy after endoscopic resection. *Endoscopy* **52** (06), 435–443 (2020).
20. Eom, B. W. et al. Nomogram incorporating CD44v6 and clinicopathological factors to predict lymph node metastasis for early gastric cancer. *PLoS One* **11** (8), e0159424 (2016).
21. Liu, Y. et al. A classification method based on feature selection for imbalanced data. *IEEE Access* **7**, 81794–81807 (2019).
22. Galar, M. et al. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man. Cybern. Part. C (Appl. Rev.)* **42** (4), 463–484 (2011).
23. Japanese Gastric Cancer Association. Japanese classification of gastric carcinoma: 3rd English edition. *Gastric Cancer* **14**, 101–112 (2011).
24. Bishop, C. M. & Nasrabadi, N. M. *Pattern Recognition and Machine learning* (Springer, 2006).
25. Gotoda, T. et al. Incidence of lymph node metastasis from early gastric cancer: Estimation with a large number of cases at two large centers. *Gastric Cancer* **3**, 219–225 (2000).
26. Sekiguchi, M. et al. Risk stratification and predictive risk-scoring model for lymph node metastasis in early gastric cancer. *J. Gastroenterol.* **51**, 961–970 (2016).
27. Rajkomar, I. Machine learning in medicine. *N. Engl. J. Med.* **380**(14), 1347–1358 (2019).
28. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25** (1), 24–29 (2019).
29. Zhu, H. et al. Preoperative prediction for lymph node metastasis in early gastric cancer by interpretable machine learning models: A multicenter study. *Surgery* **171** (6), 1543–1551 (2022).
30. Tian, H. K. et al. Application of machine learning algorithms to predict lymph node metastasis in early gastric cancer. *Front. Med.* **8**, 759013 (2022).
31. Fernández, A. et al. *Learning from imbalanced data sets* (Springer, 2018).
32. Khushi, M. et al. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access* **9**, 109960–109975 (2021).
33. Feng, X. et al. Development and validation of a machine learning model to predict the risk of lymph node metastasis in renal carcinoma. *Front. Endocrinol.* **13**, 1054358 (2022).

Author contributions

Data collection: JW, ML, MX; Data analysis: KS, HC; Chart drawing: KS, JW; Writing—original draft: KS, JW; Writing—review and editing: CJ, HC, YZ, YC. All the authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethics statement

The study was approved by the Medical Ethics Committee of Suzhou Dushu Lake Hospital (no.231004), with an exemption from informed consent.

Additional information

Correspondence and requests for materials should be addressed to H.C. or C.J.