BMJ Open

Machine learning models for prediction of lymph node metastasis in patients with gastric cancer: a Chinese single-centre study with external validation in an Asian American population

Qian Li,[1] Shangcheng Yan,[2] Weiran Yang,[3] Zhuan Du,[1] Ming Cheng,[1] Renwei Chen,[1] Qiankun Shao,[1] Yuan Tian,[1] Mengchao Sheng,[1] Wei Peng,[1] Yongyou Wu [1]

To cite: Li Q, Yan S, Yang W, *et al*. Machine learning models for prediction of lymph node metastasis in patients with gastric cancer: a Chinese single-centre study with external validation in an Asian American population. *BMJ Open* 2025;**15**:e098476. doi:10.1136/bmjopen-2024-098476

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2024-098476).

QL, SY and WY contributed equally.

Received 24 December 2024
Accepted 28 February 2025

Check for updates

For numbered affiliations see end of article.

Correspondence to
Dr Yongyou Wu;
wyoyo111@163.com

ABSTRACT

Objective To develop and validate machine learning (ML)-based models to predict lymph node metastasis (LNM) in patients with gastric cancer (GC).

Design Retrospective cohort study.

Setting Second Affiliated Hospital of Soochow University.

Participants A total of 500 inpatients from the Second Affiliated Hospital of Soochow University, collected retrospectively between 1 April 2018 and 31 March 2023, were used as the training set, while 824 Asian patients from the Surveillance, Epidemiology and End Results database comprised the external validation set.

Main outcome measures Prediction models were developed using multiple ML algorithms, including logistic regression, support vector machine, k-nearest neighbours, naive Bayes, decision tree (DT), gradient boosting DT, random forest and artificial neural network (ANN). The predictive value of these models was validated and evaluated through receiver operating characteristic curves, precision-recall (PR) curves, calibration curves, decision curve analysis and accuracy metrics.

Results Among the ML algorithms, the ANN outperformed others, achieving the highest accuracy (0.722; 95% CI: 0.692 to 0.751), precision (0.732; 95% CI: 0.694 to 0.776), F1 score (0.733; 95% CI: 0.695 to 0.773), specificity (0.728; 95% CI: 0.684 to 0.770) and area under the PR curve (0.781; 95% CI: 0.740 to 0.821) in the external validation results. Moreover, it demonstrated superior calibration and clinical utility. Shapley Additive Explanations analysis identified the depth of invasion, tumour size and Lauren classification as the most influential predictors of LNM in patients with GC. Furthermore, a user-friendly web application was developed to provide individual prediction results.

Conclusions This study introduces an accurate, reliable and clinically applicable approach for predicting the risk of LNM in patients with GC. The model demonstrates its potential to enhance the personalised management of GC in diverse populations, supported by external validation and an accessible web application for practical use.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ Employed one-hot encoding for non-ordinal variables and ordinal encoding for relevant ordinal variables to ensure proper data representation.
- ⇒ Used a univariate strategy for feature selection based on 100 iterations of fivefold cross-validation to optimise model features.
- ⇒ Performed grid search with fivefold cross-validation to identify optimal hyperparameters for each of the eight machine learning algorithms.
- ⇒ The study was retrospective, which may introduce biases such as selection and information biases.
- ⇒ Feature selection was based on accuracy in cross-validation, which might not fully capture the most relevant features for all scenarios.

INTRODUCTION

The incidence and mortality rates of gastric cancer (GC) rank fifth and fourth, respectively, among all malignant tumours globally, representing a significant threat to human health and survival.[1 2] Each year, more than one million individuals are diagnosed with GC, and the global 5-year survival rate remains alarmingly low at less than 40%.[3] In China, the situation is particularly dire, with GC ranking third in both incidence and mortality. Patients with early-stage GC in China are relatively rare, comprising less than 20% of all diagnoses, as most patients present with advanced disease.[4 5] This is largely attributed to the absence of distinct clinical symptoms in early GC, making timely diagnosis and treatment extremely challenging.[6] At the advanced stage, approximately 80% of patients develop local lymph node metastasis (LNM), a critical factor influencing prognosis and determining treatment strategies.[7] Given the pivotal role of LNM, establishing an effective and reliable risk prediction model is essential to assess the likelihood of LNM in patients with GC before therapeutic intervention.

The unique physiological structure of the stomach makes LNM a critical pathway in the progression and dissemination of GC, playing a pivotal role in the tumour-node-metastasis staging system.[8 9] Furthermore, advancements in radical gastrectomy have refined the extent of lymph node dissection, underscoring the necessity of accurately identifying LNM.[10] Previous research has established that factors such as tumour invasiveness, size and pathological type are risk factors for LNM, and prediction models based on these parameters have demonstrated moderate predictive efficacy.[11–13] Among these models, the eCura system is the most widely used for risk assessment.[14] However, critics have highlighted the stringency of the eCura criteria, emphasising the urgent need for more precise and adaptable methods to enhance predictive accuracy.[15]

The rapid advancement of various machine learning (ML) algorithms has led to their extensive integration into the medical field. Developing ML-based prediction models using clinical and pathological data offers several advantages, including ease of data acquisition, low computational complexity and strong model interpretability.[16 17] Previous research has proposed prediction models for LNM in patients with GC using radiomics and clinical features. However, many of these studies are often limited by their focus on specific GC populations or the absence of external validation, reducing the representativeness and generalisability of the models.[18–20] Therefore, this study aimed to use ML methods to construct robust LNM prediction models for GC and externally validate their predictive performance using data from Asian American patients with GC in the Surveillance, Epidemiology and End Results (SEER) database. Furthermore, based on the optimal model, a user-friendly visualisation tool was developed to enhance the intuitive presentation of prediction results.

METHODS

This study adheres to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis Artificial Intelligence (TRIPOD AI) guidelines for reporting ML-based prediction models.[21] We have carefully considered each item in the TRIPOD AI checklist during the preparation of this manuscript to ensure comprehensive and transparent reporting, enhancing the reproducibility and reliability of our study.

Study setting

Patients who underwent GC surgery at the Second Affiliated Hospital of Soochow University in Suzhou, China, between 1 April 2018 and 31 March 2023, were retrospectively included as the training dataset. Using the same criteria, Asian Americans (classified as 'Non-Hispanic Asian or Pacific Islander' in the Surveillance, Epidemiology and End Results [SEER] database) diagnosed with GC between 1 January 2018 and 31 December 2021, were extracted to form the external validation dataset.

Inclusion and exclusion criteria

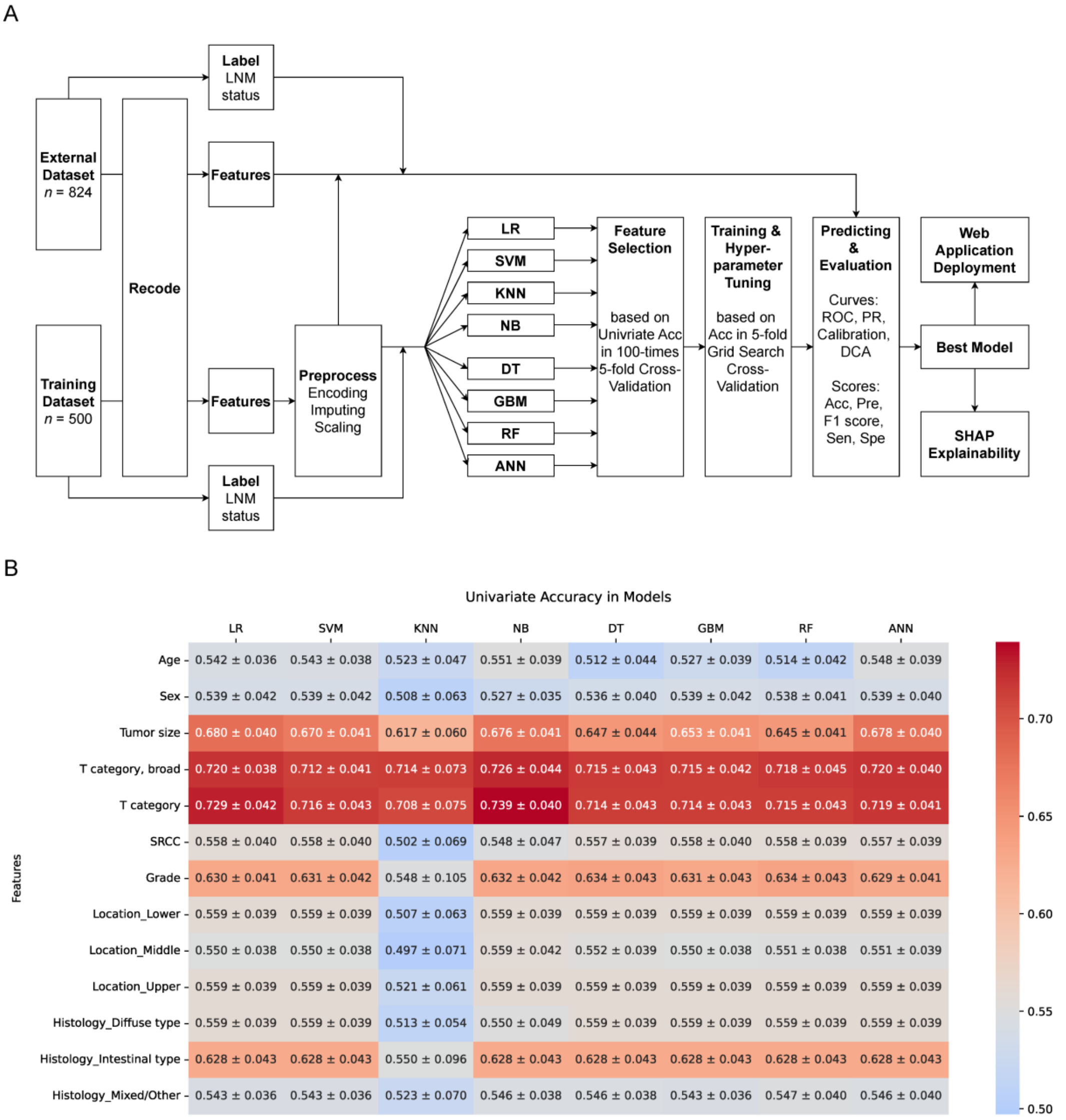
The specific inclusion criteria for patients were (1) histologically confirmed malignant epithelial tumour of the stomach and (2) availability of complete clinical and pathological data. Patients were excluded if they met any of the following criteria: (1) fewer than 16 lymph nodes were examined; (2) diagnosed with in situ, distant metastasis or unknown stage; (3) diagnosed with squamous cell carcinoma of the esophagogastric junction or (4) age ≥90 years.

Data collection and outcome definition

Demographic and clinicopathological characteristics were collected from electronic medical records and the case listing session of SEER*Stat (V.8.4.3, Surveillance Research Program, National Cancer Institute, Calverton, Maryland, USA), respectively. The predictor variables (features) collected in this study included age, sex, tumour location, tumour size, depth of invasion (T category), histology (Lauren classification), signet ring cell carcinoma (SRCC) and grade. The T category and grade were recoded according to the eighth edition of the American Joint Committee on Cancer staging system.[22] Since detailed T category information was unavailable for a few patients, we created an additional feature called 'broad T category' (T1, T2, T3 and T4) to preserve this information and its ordinal relationship. The outcome variable in this study was the status of LNM, defined as the presence of histologically confirmed metastatic foci in the lymph nodes or the diagnosis of metastatic recurrence in the lymph nodes, as detected using CT.

Data preprocessing

The flow chart of this study is shown in figure 1A. One-hot encoding was applied to sex, SRCC, tumour location and histology, while ordinal encoding was retained for 'broad T category', T category, and grade. The T category, which represents the depth of tumour invasion, was ordinal-encoded to preserve the inherent order of the categories. This encoding method is crucial as it allows ML algorithms to recognise the increasing severity of tumour invasion levels. After encoding, multiple imputation was performed using an iterative imputer with Bayesian ridge as the estimator.[23] To prevent data leakage and ensure the reliability of our model evaluation, the training and test datasets were imputed separately (online supplemental figures S1 and S2). Multiple imputation was first carried out on the training dataset using an iterative imputer with Bayesian ridge as the estimator. This process involved leveraging the relationships between variables within the training dataset to generate imputed values for missing data. Once the imputation parameters were determined from the training dataset, they were applied to the test



Features	LR	SVM	KNN	NB	DT	GBM	RF	ANN
Age	0.542 ± 0.036	0.543 ± 0.038	0.523 ± 0.047	0.551 ± 0.039	0.512 ± 0.044	0.527 ± 0.039	0.514 ± 0.042	0.548 ± 0.039
Sex	0.539 ± 0.042	0.539 ± 0.042	0.508 ± 0.063	0.527 ± 0.035	0.536 ± 0.040	0.539 ± 0.042	0.538 ± 0.041	0.539 ± 0.040
Tumor size	0.680 ± 0.040	0.670 ± 0.041	0.617 ± 0.060	0.676 ± 0.041	0.647 ± 0.044	0.653 ± 0.041	0.645 ± 0.041	0.678 ± 0.040
T category, broad	0.720 ± 0.038	0.712 ± 0.041	0.714 ± 0.073	0.726 ± 0.044	0.715 ± 0.043	0.715 ± 0.042	0.718 ± 0.045	0.720 ± 0.040
T category	0.729 ± 0.042	0.716 ± 0.043	0.708 ± 0.075	0.739 ± 0.040	0.714 ± 0.043	0.714 ± 0.043	0.715 ± 0.043	0.719 ± 0.041
SRCC	0.558 ± 0.040	0.558 ± 0.040	0.502 ± 0.069	0.548 ± 0.047	0.557 ± 0.039	0.558 ± 0.040	0.558 ± 0.039	0.557 ± 0.039
Grade	0.630 ± 0.041	0.631 ± 0.042	0.548 ± 0.105	0.632 ± 0.042	0.634 ± 0.043	0.631 ± 0.043	0.634 ± 0.043	0.629 ± 0.041
Location_Lower	0.559 ± 0.039	0.559 ± 0.039	0.507 ± 0.063	0.559 ± 0.039	0.559 ± 0.039	0.559 ± 0.039	0.559 ± 0.039	0.559 ± 0.039
Location_Middle	0.550 ± 0.038	0.550 ± 0.038	0.497 ± 0.071	0.559 ± 0.042	0.552 ± 0.039	0.550 ± 0.038	0.551 ± 0.038	0.551 ± 0.039
Location_Upper	0.559 ± 0.039	0.559 ± 0.039	0.521 ± 0.061	0.559 ± 0.039	0.559 ± 0.039	0.559 ± 0.039	0.559 ± 0.039	0.559 ± 0.039
Histology_Diffuse type	0.559 ± 0.039	0.559 ± 0.039	0.513 ± 0.054	0.550 ± 0.049	0.559 ± 0.039	0.559 ± 0.039	0.559 ± 0.039	0.559 ± 0.039
Histology_Intestinal type	0.628 ± 0.043	0.628 ± 0.043	0.550 ± 0.096	0.628 ± 0.043	0.628 ± 0.043	0.628 ± 0.043	0.628 ± 0.043	0.628 ± 0.043
Histology_Mixed/Other	0.543 ± 0.036	0.543 ± 0.036	0.523 ± 0.070	0.546 ± 0.038	0.546 ± 0.038	0.543 ± 0.036	0.547 ± 0.040	0.546 ± 0.040

Figure 1 Study design and feature selection. (A) Flow chart of the study; (B) Univariate accuracy (mean±SD) of each feature in different models, calculated using fivefold cross-validation. Features with a mean accuracy greater than 0.55 were selected for model training. Acc, accuracy; ANN, artificial neural network; DCA, decision curve analysis; DT, decision tree; GBM, gradient boosting machine; KNN, k-nearest neighbours; LNM, lymph node metastasis; LR, logistic regression; NB, naive Bayes; PR, precision-recall; Pre, precision; RF, random forest; ROC, receiver operating characteristic; Sen, sensitivity; SHAP, Shapley Additive Explanations; Spe, specificity; SRCC, signet ring cell carcinoma; SVM, support vector machine.

dataset for a single imputation. This approach maintained the integrity of both datasets and ensured that the performance metrics were not inflated due to improper data handling. Features were then scaled using the power transformer to transform the data to a more normal distribution, which can enhance the performance of ML algorithms and alleviate the potential negative impacts of unevenly distributed variables like the T category.[24] A univariate strategy was used for feature selection. For each of the eight ML algorithms (discussed later), the accuracy of each feature was calculated using 100 iterations of fivefold cross-validation. Features with an accuracy below 0.55 were eliminated from the model training. This univariate feature-selection strategy based on cross-validation was effective in reducing the influence of variables with potentially problematic distributions, such as the T category variable with its multiple and unevenly distributed categories. The high accuracy of the T category variable in most algorithms indicated its significance and stability in the prediction models. This comprehensive data preprocessing approach ensured that the data used for model training were of high quality and minimised the risk of bias introduced by data-related issues.

Model training

Eight algorithms were selected for model training: logistic regression (LR),[25] support vector machine (SVM),[26] k-nearest neighbours,[27] naive Bayes (NB),[28] decision tree (DT),[29] gradient boosting DT,[30] random forest (RF)[31] and artificial neural network (ANN).[32] Grid search with fivefold cross-validation was conducted to identify the optimal hyperparameters for each model, using accuracy as the evaluation metric. Once the optimal hyperparameters were determined, the best model for each ML algorithm was trained using the entire training dataset. It is important to note that the external dataset was used for both comparing the performance of different ML algorithms to select the optimal model and for external validation. This approach is within a validation framework rather than a strict test framework, where a test set is typically used only for the final evaluation of model performance.

Statistical analysis

The continuous variables were described as mean (SD) and median (IQR) according to data distribution. Differences between the training set and validation set were then compared using an independent t-test or Mann-Whitney U test, as appropriate. The categorical variables were expressed as n (%) and compared using the χ^2 test. A two-sided $p<0.05$ was considered statistically significant. The performance of models trained using various algorithms was evaluated and compared on an external validation set to determine the best model. All metrics and plots were calculated using the mean and 95% CI obtained from 1000 iterations of Bootstrap resampling to assess the robustness of the models. Accuracy, precision, F1 score, sensitivity (recall) and specificity were calculated to evaluate the model's discrimination. Additionally, the receiver operating characteristic curve and the precision-recall (PR) curve were plotted, and their areas under the curves (AUCs) were calculated. The calibration of the models was assessed by plotting the calibration curve. We compared the clinical value of each model through decision curve analysis (DCA).[33] The Shapley Additive Explanations (SHAP) model was used for the visual analysis of the contribution of each feature to the prediction results.[27] SHAP is a game theory-based ML additive explanation model where all features are considered 'contributors'. The model generates a predicted value for each prediction sample, and the SHAP value is the numerical value assigned to each feature in that sample.[34] R (V.4.3.2) was used for data cleansing, patient selection and comparison of baseline characteristics, while all other analyses were conducted using Python (V. 3.11.5). Data preprocessing, model training and evaluation were performed using scikit-learn (V.1.3.2).[35] The model explanation was conducted using SHAP (V.0.43.0). Additionally, we developed a web application for the best model using Streamlit (V.1.28.2). The code and datasets used in this study are available at https://github.com/SCYAN0401/Predict-GC-LNM.

Patient and public involvement

Patients and the public were not involved in the development of the research question.

RESULTS

Patient characteristics

The training set and external validation set in this study included 500 and 824 patients with GC, respectively. The baseline characteristics of the patients are shown in table 1. Significant differences exist between the training set and external validation set for sex ($p<0.001$), location of GC ($p<0.001$), T category ($p<0.001$), histology of GC ($p<0.001$) and SRCC ($p<0.001$). However, no significant differences were observed between the groups for age ($p=0.148$), tumour size ($p=0.246$), tumour grade ($p=0.057$) and LNM ($p=0.334$).

Feature selection

In the fivefold cross-validation, the variables 'T category (broad)', T category, tumour size, grade and histology exhibited the highest accuracy in almost all algorithms (figure 1B). The 'T category (broad)' and T category achieved the highest accuracies in the NB model, with values of 0.739±0.040 and 0.726±0.044, respectively. Tumour size had the highest accuracy in the LR model, with a value of 0.680±0.040. Features with an accuracy of less than 0.55 were excluded from subsequent model training.

As shown in figure 1B, some features, such as the 'T category (broad)', T category, tumour size, have strong predictive values despite being heterogeneously distributed between the training and validation datasets. By comparing the performance of these features in the training stage, we can gain a better understanding of their transferability. The 'T category (broad)' feature, which had high accuracy in most algorithms during the training stage, also maintained a relatively strong predictive power in the external validation set. This indicates that although its distribution varied between the two datasets, the information it carried was still relevant and useful for the model's prediction. However, for some features with large differences in distribution, further research may be needed to explore how to better use them in different datasets to improve the model's generalisation ability.

Model training and evaluation

The performance of the eight ML models during the training stage was evaluated using several metrics, including accuracy, precision, F1-score, sensitivity and specificity (figure 2A). Within the external validation framework, where the external dataset was also used for model selection, the ANN demonstrated high accuracy (0.722; 95% CI: 0.692 to 0.751), precision (0.732; 95% CI: 0.694 to 0.776), F1 score (0.733; 95% CI: 0.695 to 0.773) and specificity (0.728; 95% CI: 0.684 to 0.770; figure 2A). The SVM had the highest sensitivity (0.740; 95% CI: 0.632 to 0.885), followed by ANN (0.731; 95% CI: 0.693 to 0.772), which outperformed other algorithms (figure 2A). LR had the highest AUC on the receiver operating characteristic curve (0.823; 95% CI: 0.785 to 0.859; figure 2B), while ANN had the highest AUC on the PR curve (0.781;

Table 1 Characteristics of GC patients in training and external validation sets

Variable	Category	Training set (n=500)	External validation set (n=824)	P value
Age (years), median (IQR)		67 (59, 73)	68 (59, 75)	0.148
Sex, n (%)	Female	130 (26.0)	343 (41.6)	<0.001
	Male	370 (74.0)	481 (58.4)	
Tumour location, n (%)	Lower	273 (54.7)	318 (45.1)	<0.001
	Middle	62 (12.4)	272 (38.6)	
	Upper	164 (32.9)	115 (16.3)	
Tumour size (mm), median (IQR)		35 (20, 50)	35 (21, 55)	0.246
T category, broad, n (%)	T1	140 (28.3)	208 (25.5)	<0.001
	T2	112 (22.6)	118 (14.5)	
	T3	219 (44.2)	301 (36.9)	
	T4	24 (4.8)	189 (23.2)	
T category, n (%)	T1a	56 (11.3)	70 (8.9)	<0.001
	T1b	84 (17.0)	108 (13.7)	
	T2	112 (22.6)	118 (15.0)	
	T3	219 (44.2)	301 (38.3)	
	T4a	16 (3.2)	160 (20.4)	
	T4b	8 (1.6)	29 (3.7)	
Histology, n (%)	Diffuse type	125 (25.0)	246 (29.9)	<0.001
	Intestinal type	117 (23.4)	541 (65.7)	
	Mixed/other	258 (51.6)	37 (4.5)	
SRCC, n (%)	No	467 (93.4)	682 (82.8)	<0.001
	Yes	33 (6.6)	142 (17.2)	
Tumour grade, n (%)	G1	8 (1.7)	25 (3.4)	0.057
	G2	103 (21.7)	187 (25.3)	
	G3	364 (76.6)	528 (71.4)	
LNM, n (%)	No	221 (44.2)	388 (47.1)	0.334
	Yes	279 (55.8)	436 (52.9)	

GC, gastric cancer; LNM, lymph node metastasis; SRCC, signet ring cell carcinoma.

95% CI: 0.740 to 0.821; figure 2C). Additionally, ANN exhibited the best calibration curve (figure 2D) and DCA performance (figure 2E). In summary, the ANN model demonstrated superior discrimination and calibration performance, making it the most suitable for clinical application. Therefore, it was selected as the optimal algorithm for further analysis.

Model tuning process for the ANN

The hyperparameters of the ANN were carefully selected through a combination of literature-based insights and experimental optimisation. The choice of three hidden layers was based on previous research in medical predictive modelling using ANNs, which suggested that 2–4 hidden layers can effectively capture complex relationships in the data. Initial experiments comparing models with two and three hidden layers showed that the latter achieved better performance in terms of accuracy, precision and F1-score on both the training and validation datasets. For the number of neurons in each hidden layer, a grid-search approach with fivefold cross-validation was employed. In the first hidden layer, values ranging from 10 to 50 neurons were tested. A value of 20 neurons was chosen as it provided an optimal balance between model complexity and performance, effectively extracting initial features from the input data without overfitting. In the second hidden layer, values from 30 to 100 neurons were explored, and 50 neurons were selected as they enabled the model to further process and combine the features learnt in the first layer. For the third hidden layer, after testing values from 80 to 150 neurons, 110 neurons were determined to be the best choice, leading to the highest overall performance of the model on the validation set. To address the potential issue of overfitting, especially considering the limited data, we implemented several strategies. We used fivefold cross-validation during model training and hyperparameter tuning. Additionally,

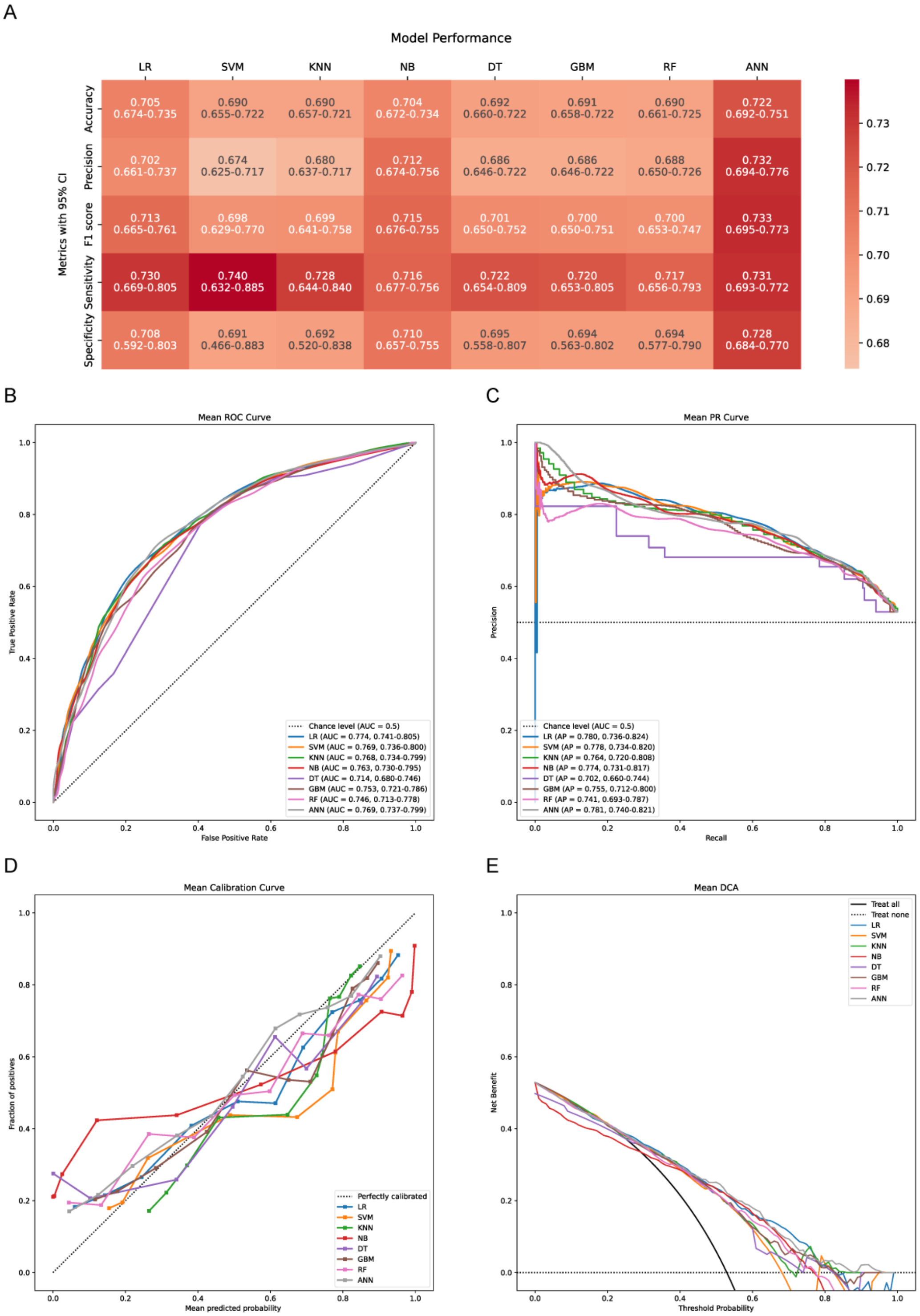
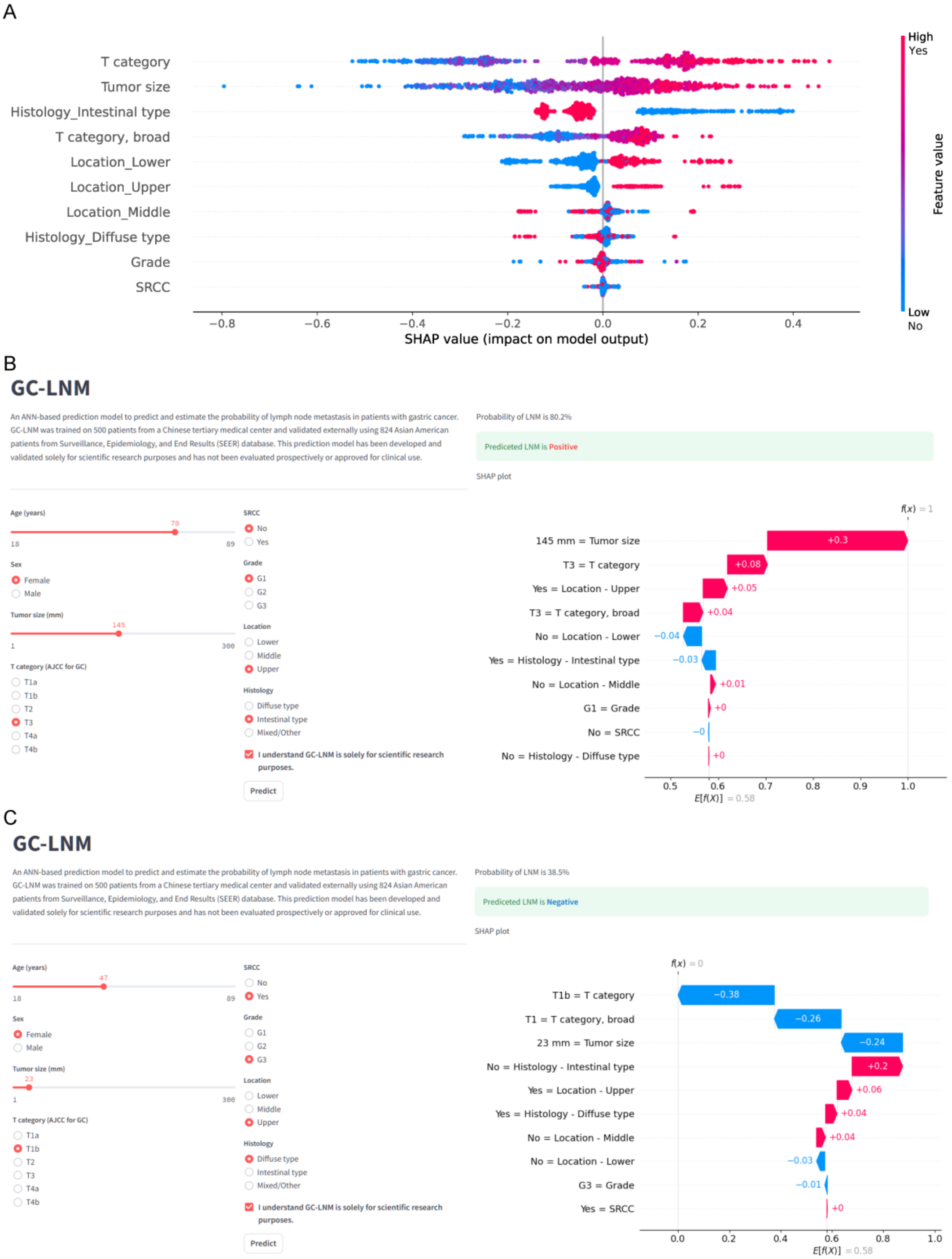


Figure 2 Performance of models in the external testing dataset. All metrics and curves are expressed as means with 95% CIs, computed via 1000 bootstrap resamples. (A) Accuracy, precision, F1 score, sensitivity and specificity; (B) Receiver operating characteristic (ROC) curves; (C) Precision-recall (PR) curves; (D) Calibration curves; (E) Decision curve analysis (DCA). (B, C) The dashed line represents the chance level. (D) The dashed line represents perfect calibration. ANN, artificial neural network; AUC, area under the curve; DT, decision tree; GBM, gradient boosting machine; KNN, k-nearest neighbours; LNM, lymph node metastasis; LR, logistic regression; NB, naive Bayes; RF, random forest; SVM, support vector machine.

we closely monitored the performance of the model on both the training and validation sets. The consistent performance of the ANN on these two sets, along with the acceptable variance calculated through 1000 iterations of Bootstrap resampling, indicated that the chosen number of neurons and layers did not result in overfitting.

SHAP explainability and web application deployment

The final ANN model consisted of three hidden layers with 20, 50 and 110 neurons, respectively, using the rectified linear unit as the activation function and stochastic gradient descent as the solver. Considering the black-box nature of ANNs, SHAP visualisation was introduced to enhance interpretability. The T category had the highest mean absolute SHAP value, indicating the strongest predictive power, followed by tumour size and Lauren intestinal type (figure 3A). Increased depth of invasion and tumour size made a positive prediction more likely. Intestinal-type tumours were less likely to predict LNM positivity. Compared with upper and lower GCs, mid-gastric tumours were more likely to develop LNM. Grade



GC-LNM

An ANN-based prediction model to predict and estimate the probability of lymph node metastasis in patients with gastric cancer. GC-LNM was trained on 500 patients from a Chinese tertiary medical center and validated externally using 824 Asian American patients from Surveillance, Epidemiology, and End Results (SEER) database. This prediction model has been developed and validated solely for scientific research purposes and has not been evaluated prospectively or approved for clinical use.

Age (years): 70 (18–89)

Sex: ● Female ○ Male

Tumor size (mm): 145 (1–300)

T category (AJCC for GC): ○ T1a ○ T1b ○ T2 ● T3 ○ T4a ○ T4b

SRCC: ● No ○ Yes

Grade: ● G1 ○ G2 ○ G3

Location: ○ Lower ○ Middle ● Upper

Histology: ○ Diffuse type ● Intestinal type ○ Mixed/Other

☑ I understand GC-LNM is solely for scientific research purposes.

Predict

Probability of LNM is 80.2%

Prediceted LNM is Positive

SHAP plot

f(x) = 1
145 mm = Tumor size +0.3
T3 = T category +0.08
Yes = Location - Upper +0.05
T3 = T category, broad +0.04
No = Location - Lower −0.04
Yes = Histology - Intestinal type −0.03
No = Location - Middle +0.01
G1 = Grade +0
No = SRCC −0
No = Histology - Diffuse type +0
0.5
0.6
0.7
0.8
0.9
1.0
E[f(X)] = 0.58
C

GC-LNM

An ANN-based prediction model to predict and estimate the probability of lymph node metastasis in patients with gastric cancer. GC-LNM was trained on 500 patients from a Chinese tertiary medical center and validated externally using 824 Asian American patients from Surveillance, Epidemiology, and End Results (SEER) database. This prediction model has been developed and validated solely for scientific research purposes and has not been evaluated prospectively or approved for clinical use.

Age (years): 47 (18–89)

Sex: ● Female ○ Male

Tumor size (mm): 23 (1–300)

T category (AJCC for GC): ○ T1a ● T1b ○ T2 ○ T3 ○ T4a ○ T4b

SRCC: ○ No ● Yes

Grade: ○ G1 ○ G2 ● G3

Location: ○ Lower ○ Middle ● Upper

Histology: ● Diffuse type ○ Intestinal type ○ Mixed/Other

☑ I understand GC-LNM is solely for scientific research purposes.

Predict

Probability of LNM is 38.5%

Prediceted LNM is Negative

SHAP plot

f(x) = 0
T1b = T category −0.38
T1 = T category, broad −0.26
23 mm = Tumor size −0.24
No = Histology - Intestinal type +0.2
Yes = Location - Upper +0.06
Yes = Histology - Diffuse type +0.04
No = Location - Middle +0.04
No = Location - Lower −0.03
G3 = Grade −0.01
Yes = SRCC +0
0.0
0.2
0.4
0.6
0.8
1.0
E[f(X)] = 0.58

Figure 3 SHAP visualisation and examples of web application. (A) SHAP bee swarm plot of features. The width of the horizontal bars represents the feature's impact on the prediction, while the colour of the horizontal bars indicates the feature's magnitude. The x-axis represents the likelihood of having lymph node metastasis; (B) An example of a patient with an 80.2% probability of lymph node metastasis and a positive lymph node metastasis prediction by our application; (C) An example of a patient with a 38.5% probability of lymph node metastasis and a negative lymph node metastasis prediction by our application. GC, gastric cancer; LNM, lymph node metastasis; SHAP, Shapley Additive Explanations; SRCC, signet ring cell carcinoma.

and SRCC had the least impact on the prediction, with uncertain predictive roles.

The ANN model was deployed into a web application (https://gc-lnm.streamlit.app/), where inputting patient features can display the GC LNM prediction and probability, along with SHAP waterfall plots (figure 3B,C). For example, a 145 mm upper gastric T3 tumour with intestinal histology and G1 grade was predicted to be LNM positive (figure 3B), while a 23 mm upper gastric T1b diffuse G3 SRCC tumour was predicted to be LNM negative (figure 3C). The SHAP waterfall plots help explain why the predictions were positive or negative.

DISCUSSION

This study used eight ML algorithms to develop a predictive model based on data from 500 patients with GC at a hospital in China. The model was externally validated using data from 824 Asian patients with GC in the SEER database. By combining hospital-based and database-driven datasets, our findings offer broader applicability to patients with GC, particularly within Asian populations, compared with studies restricted to single-centre data or database-only analyses. In external validation, the ANN model demonstrated superior performance among all ML algorithms, achieving an accuracy of 0.722. SHAP visualisation analysis identified the depth of invasion, tumour size, and histological type as the most influential predictors of LNM in GC. To enhance clinical adoption, we developed a user-friendly web application that allows practitioners to predict LNM and visualise contributing factors using SHAP waterfall plots. Future studies will focus on prospective validation to further evaluate the clinical utility and impact of the model.

Previous studies have used ML to develop predictive models for assessing the risk of LNM in patients with GC.[19 36 37] For instance, Yue and Xue used ML algorithms to identify risk factors for LNM and created a predictive model for stage II–III patients with GC. Their model incorporated variables such as vascular invasion, maximum tumour diameter, monocyte percentage, haematocrit, and lymphocyte-to-monocyte ratio, demonstrating satisfactory predictive performance.[19] Similarly, Seo *et al* developed and validated a predictive model for LNM in patients with T1b early GC using ML algorithms, identifying the XGBoost model as the most effective.[36] Lu *et al* constructed an ML-based predictive model for LNM risk in patients with GC, where six models (DT, RF, GBM, NB, ANN and LR) exhibited high accuracy and reliability, with clear insights into the impact of each risk factor. However, their study noted relatively poor performance with the SVM algorithm.[37] While these studies highlighted the potential of ML in predicting LNM risk, they were often limited by small sample sizes and lacked extensive external validation, reducing the practical utility of their models in diverse clinical settings. To address these limitations, this study aimed to develop a new predictive model for LNM risk in patients with GC using an MI approach.

Our study used cross-validation to select features for each model using a data-driven approach. To address the 'stage migration' phenomenon,[38] patients with fewer than 16 lymph nodes were excluded, ensuring the development of a robust and reliable ML model for predicting the risk of LNM in patients with GC. Consistent with previous studies,[39] T classification and tumour size emerged as the most predictive factors. However, we observed that age and sex contributed minimally to LNM prediction, which contrasts with earlier studies.[40 41]

Additionally, the impact of SRCC on predictive outcomes was minimal and ambiguous, suggesting that its LNM behaviour may align closely with that of non-SRCC GC. The roles of tumour location and Lauren classification in LNM risk remain contentious. Our findings indicate that the Lauren intestinal type is associated with a lower likelihood of LNM compared with other histological subtypes. Our study observed that the middle stomach is more prone to LNM compared with the upper and lower stomach. This finding may further support the use of pylorus-preserving gastrectomy as an effective treatment for mid-third GC.[42]

To enhance clinical prediction, Steyerberg and Vergouwe proposed the 'ABCD' model evaluation criteria, encompassing A (alpha, wide-range calibration), B (beta, calibration slope), C (C-statistic, discrimination) and D (DCA, clinical significance).[43] Using these criteria, we identified the ANN as the optimal model for predicting LMN in patients with GC. This study is the first to develop a GC-LNM prediction model that meets the 'ABCD' standards. Compared with existing ANN-based prediction models, our model demonstrated significantly improved predictive performance compared with previous studies.[44 45] Additionally, our study incorporated comprehensive comparisons with other ML algorithms to further validate its superiority. To bridge the gap between model development and clinical application, we designed a web-based application that delivers personalised predictions in an intuitive and accessible format.

This study had the following strengths: (1) Non-ordinal variables were encoded using one-hot encoding, avoiding arbitrary categorisation. Moreover, power transformations were applied to achieve a pseudo-normal distribution, and missing values were addressed through multiple imputations, ensuring data integrity instead of excluding samples with missing data; (2) cross-validation was used to select features for each model, enhancing both the reliability and accuracy of the predictive models; (3) eight ML algorithms were evaluated, offering a robust comparative perspective and validating the ANN as the optimal model and (4) a user-friendly web application was developed, providing personalised predictions in a simple and accessible format for clinical use.

However, this study has certain limitations. First, the models were constructed using retrospective data and require further validation through prospective cohort studies to confirm their clinical utility. Second, data collection was limited to clinical routine records and the SEER database, restricting the inclusion of potentially relevant features associated with.[46] Third, the models need further refinement, including the incorporation of imaging and molecular features, to improve predictive performance and applicability. Fourth, missing data in our dataset could have had a substantial impact on model performance. Ignoring these missing values would have likely biased the model towards the non-missing data. Fifth, the external dataset was used for both model selection and validation, which deviates from the ideal train-validation-test split framework. This may affect the generalisability of the model's performance as reported, and future studies should aim to use a separate, independent test set for a more accurate assessment. These limitations highlight the need for continued validation and enhancement of the models to ensure their reliability and effectiveness in real-world clinical scenarios.

CONCLUSIONS

This study used a retrospective design and data-driven methodologies to create a robust and reliable ML model for predicting LNM in patients with GC based on simple clinical and pathological variables. Developing a user-friendly web application enables personalised predictions and supports clinical decision-making. Despite its strong performance on the current dataset, the model requires further validation through prospective cohort studies and enhancements by integrating additional imaging and molecular features to improve its predictive accuracy and clinical applicability.

Author affiliations

[1]Department of Gastrointestinal Surgery, Second Affiliated Hospital of Soochow University, Suzhou, Jiangsu, China

[2]Department of General Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[3]Institute of Exercise Training and Sport Informatics, German Sport University, Cologne, Germany

Contributors Conception and design: YW, WP and SY. Collection and assembly of data: MC, SY, RC, QS, YT and MS; Data analysis and interpretation: SY, WY and ZD; Manuscript writing: QL, SY, WY and ZD; Final approval of manuscript: All authors. SY, QL and WY contributed equally to this work and share first authorship. WY is the guarantor. We have not used AI.

Funding This study was funded by Clinical Key Disease Diagnosis and Treatment Technique Project of Suzhou (LCZX202310); Soochow University Second Hospital Medical New Technology Special Assistance Plan Project (23ZL002).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The study was approved by the Institutional Review Board (IRB) of Second Affiliated Hospital of Soochow University (no: JD-HG-2024-007). Given its retrospective design, the requirement for obtaining informed consent was waived.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. The code and datasets used in this study are also available at https://github.com/SCYAN0401/Predict-GC-LNM.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Yongyou Wu http://orcid.org/0009-0004-3434-2508

REFERENCES

1 Bray F, Laversanne M, Sung H, *et al*. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024;74:229–63.
2 Smyth EC, Nilsson M, Grabsch HI, *et al*. Gastric cancer. *Lancet* 2020;396:635–48.
3 Siegel RL, Kratzer TB, Giaquinto AN, *et al*. Cancer statistics, 2025. *CA Cancer J Clin* 2025;75:10–45.
4 Wang FH, Zhang XT, Li YF, *et al*. The Chinese Society of Clinical Oncology (CSCO): Clinical guidelines for the diagnosis and treatment of gastric cancer. 2021;41:747–95.
5 Xu H, Li W. Early detection of gastric cancer in China: progress and opportunities. *Cancer Biol Med* 2022;19:1622–8.
6 Maconi G, Manes G, Porro GB. Role of symptoms in diagnosis and outcome of gastric cancer. *World J Gastroenterol* 2008;14:1149–55.
7 Li GZ, Doherty GM, Wang J. Surgical Management of Gastric Cancer: A Review. *JAMA Surg* 2022;157:446–54.
8 Teng F, Fu Y-F, Wu A-L, *et al*. Computed Tomography-Based Predictive Model for the Probability of Lymph Node Metastasis in Gastric Cancer: A Meta-analysis. *J Comput Assist Tomogr* 2024;48:19–25.
9 Wang Z, Liu Q, Zhuang X, *et al*. pT1-2 gastric cancer with lymph node metastasis predicted by tumor morphologic features on contrast-enhanced computed tomography. *Diagn Interv Radiol* 2023;29:228–33.
10 Ajani JA, D'Amico TA, Bentrem DJ, *et al*. Gastric Cancer, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* 2022;20:167–92.
11 Chu Y-N, Yu Y-N, Jing X, *et al*. Feasibility of endoscopic treatment and predictors of lymph node metastasis in early gastric cancer. *World J Gastroenterol* 2019;25:5344–55.
12 Lee JH, Choi IJ, Kook MC, *et al*. Risk factors for lymph node metastasis in patients with early gastric cancer and signet ring cell histology. *Br J Surg* 2010;97:732–6.
13 Folli S, Morgagni P, Roviello F, *et al*. Risk factors for lymph node metastases and their prognostic significance in early gastric cancer (EGC) for the Italian Research Group for Gastric Cancer (IRGGC). *Jpn J Clin Oncol* 2001;31:495–9.
14 Hatta W, Gotoda T, Oyama T, *et al*. A Scoring System to Stratify Curability after Endoscopic Submucosal Dissection for Early Gastric Cancer: "eCura system". *Am J Gastroenterol* 2017;112:874–81.
15 Lee HD, Nam KH, Shin CM, *et al*. Development and Validation of Models to Predict Lymph Node Metastasis in Early Gastric Cancer Using Logistic Regression and Gradient Boosting Machine Methods. *Cancer Res Treat* 2023;55:1240–9.
16 Oliveira AL. Biotechnology, Big Data and Artificial Intelligence. *Biotechnol J* 2019;14:e1800613.
17 Mirza B, Wang W, Wang J, *et al*. Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes (Basel)* 2019;10:87.
18 Ma T, Zhao M, Li X, *et al*. n.d. A machine learning based radiomics approach for predicting No. 14v station lymph node metastasis in gastric cancer. *Front Med*11:1464632.
19 Yue C, Xue H. Construction and validation of a nomogram model for lymph node metastasis of stage II-III gastric cancer based on machine learning algorithms. *Front Oncol* 2024;14:1399970.
20 Lin W-W, Zhong Q, Guo J, *et al*. A Preoperative Prediction Model for Lymph Node Metastasis in Patients with Gastric Cancer Using a Machine Learning-based Ultrasomics Approach. *Curr Med Imaging* 2024;20:e15734056291074.

21 Collins GS, Moons KGM, Dhiman P, *et al*. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378.

22 Amin MB, Greene FL, Edge SB, *et al*. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more “personalized” approach to cancer staging. *CA Cancer J Clin* 2017;67:93–9.

23 Li Y, Ji L, Oravecz Z, *et al*. dynr.mi: An R Program for Multiple Imputation in Dynamic Modeling. *World Acad Sci Eng Technol* 2019;13:302–11.

24 Klawonn F, Jayaram B, Crull K, *et al*. Analysis of contingency tables based on generalised median polish with power transformations and non-additive models. *Health Inf Sci Syst* 2013;1:11.

25 Meurer WJ, Tolles J. Logistic Regression Diagnostics: Understanding How Well a Model Predicts Outcomes. *JAMA* 2017;317:1068–9.

26 Rahman MM, Desai BC, Bhattacharya P. Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion. *Comput Med Imaging Graph* 2008;32:95–108.

27 Yang Y, Xu B, Haverstick J, *et al*. Differentiation and classification of bacterial endotoxins based on surface enhanced Raman scattering and advanced machine learning. *Nanoscale* 2022;14:8806–17.

28 Balfer J, Bajorath J. Introduction of a methodology for visualization and graphical interpretation of Bayesian classification models. *J Chem Inf Model* 2014;54:2451–68.

29 Chien C, Pottie GJ. A universal hybrid decision tree classifier design for human activity classification. *Annu Int Conf IEEE Eng Med Biol Soc* 2012;2012:1065–8.

30 Liang Y, Zhang S, Qiao H, *et al*. iEnhancer-MFGBDT: Identifying enhancers and their strength by fusing multiple features and gradient boosting decision tree. *Math Biosci Eng* 2021;18:8797–814.

31 Zhu J, Zhao Y, Hu Q, *et al*. Coalbed Methane Production Model Based on Random Forests Optimized by a Genetic Algorithm. *ACS Omega* 2022;7:13083–94.

32 Pastur-Romay LA, Cedrón F, Pazos A, *et al*. Deep Artificial Neural Networks and Neuromorphic Chips for Big Data Analysis: Pharmaceutical and Bioinformatics Applications. *Int J Mol Sci* 2016;17:1313.

33 Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;3:18.

34 Ponce-Bobadilla AV, Schmitt V, Maier CS, *et al*. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clin Transl Sci* 2024;17:e70056.

35 Abraham A, Pedregosa F, Eickenberg M, *et al*. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform* 2014;8:14.

36 Seo JW, Park KB, Lim ST, *et al*. Machine learning models for prediction of lymph node metastasis in patients with T1b gastric cancer. *Am J Cancer Res* 2024;14:3842–51.

37 Lu T, Lu M, Wu D, *et al*. Predictive value of machine learning models for lymph node metastasis in gastric cancer: A two-center study. *World J Gastrointest Surg* 2024;16:85–94.

38 Komatsu S, Ichikawa D, Nishimura M, *et al*. Evaluation of prognostic value and stage migration effect using positive lymph node ratio in gastric cancer. *Eur J Surg Oncol* 2017;43:203–9.

39 Li Y, Xie F, Xiong Q, *et al*. Machine learning for lymph node metastasis prediction of in patients with gastric cancer: A systematic review and meta-analysis. *Front Oncol* 2022;12:946038.

40 Takatsu Y, Hiki N, Nunobe S, *et al*. Clinicopathological features of gastric cancer in young patients. *Gastric Cancer* 2016;19:472–8.

41 Zhao X, Cai A, Xi H, *et al*. Predictive Factors for Lymph Node Metastasis in Undifferentiated Early Gastric Cancer: a Systematic Review and Meta-analysis. *J Gastrointest Surg* 2017;21:700–11.

42 Oh SY, Lee HJ, Yang HK. Pylorus-Preserving Gastrectomy for Gastric Cancer. *J Gastric Cancer* 2016;16:63–71.

43 Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35:1925–31.

44 Bollschweiler EH, Mönig SP, Hensler K, *et al*. Artificial neural network for prediction of lymph node metastases in gastric cancer: a phase II diagnostic study. *Ann Surg Oncol* 2004;11:506–11.

45 Xue Z, Lu J, Lin J, *et al*. Establishment of artificial neural network model for predicting lymph node metastasis in patients with stage II-III gastric cancer. *Zhonghua Wei Chang Wai Ke Za Zhi* 2022;25:327–35.

46 Kamarajah SK, Nathan H. Strengths and Limitations of Registries in Surgical Oncology Research. *J Gastrointest Surg* 2021;25:2989–96.