

<https://doi.org/10.1038/s41698-025-01055-9>

Development of a deep learning model for T1N0 gastric cancer diagnosis using 2.5D radiomic data in preoperative CT images

Check for updates

Jingyang He^{1,8}, Jingli Xu^{1,2,8}, Wujie Chen^{3,4,5,8}, Mengxuan Cao^{3,4,5}, Jiaqing Zhang¹, Qing Yang¹, Enze Li¹, Ruolan Zhang¹, Yahang Tong^{3,4,5}, Yanqiang Zhang^{1,4,5}, Chen Gao⁶, Qianyu Zhao¹, Zhiyuan Xu^{1,4,5}, Lijing Wang¹, Xiangdong Cheng^{1,4,5}, Guoliang Zheng⁷✉, Siwei Pan^{1,4,5}✉ & Can Hu^{1,4,5}✉

Early detection and precise preoperative staging of early gastric cancer (EGC) are critical. Therefore, this study aims to develop a deep learning model using portal venous phase CT images to accurately distinguish EGC without lymph node metastasis. This study included 3164 patients with gastric cancer (GC) who underwent radical surgery at two medical centers in China from 2006 to 2019. Moreover, 2.5D radiomic data and multi-instance learning (MIL) were novel approaches applied in this study. By basing the selection of features on 2.5D radiomic data and MIL, the ResNet101 model combined with the XGBoost model represented a satisfactory performance for diagnosing pT1N0 GC. Furthermore, the 2.5D MIL-based model demonstrated a markedly superior predictive performance compared to traditional radiomics models and clinical models. We first constructed a deep learning prediction model based on 2.5D radiomics and MIL for effectively diagnosing pT1N0 GC patients, which provides valuable information for the individualized treatment selection.

Gastric cancer (GC) is a prevalent malignancy and ranks among the top three causes of cancer-related mortality¹. The five-year survival rate for advanced gastric cancer (AGC) is less than 30%, while early gastric cancer (EGC) shows a significantly higher survival rate of 90%². EGC refers to invasive gastric cancer that is confined to the submucosa, regardless of lymph node involvement³. The primary treatment for EGC without lymph node metastasis is endoscopic resection⁴, including endoscopic submucosal dissection (ESD) and endoscopic mucosal resection (EMR). For EGC patients with cN+, gastrectomy plus D2 lymph node dissection might be the first choice for treatment⁵. Therefore, precise diagnosis of EGC, along with a thorough evaluation of the presence and risk of lymph node metastasis (LNM), is crucial for determining the appropriate treatment strategy.

For the preoperative diagnosis of EGC, computed tomography (CT)⁶ and endoscopic ultrasonography (EUS)⁷ are the most widely used techniques. Reports indicate that CT's overall diagnostic accuracy for

T-staging ranges from 73.8% to 88.9%^{8–10}. However, radiologists' evaluation of its diagnostic accuracy for the T1 stage was moderate, ranging from 63.0% to 82.7%^{11,12}. Although EUS can differentiate between superficial (T1–T2) and advanced (T3–T4) primary gastric tumors with a sensitivity and specificity exceeding 85%, certain studies have reported reduced diagnostic precision of EUS in distinguishing between subtypes of early-stage tumors (T1a vs T1b) and in identifying tumors with or without LNM¹³. The accuracy of current auxiliary examinations, whether EUS or CT with manual evaluation, in diagnosing lymph node status in GC patients is very limited, and there are inevitable limitations in evaluating TNM staging. This impetus fosters the advancement of novel methods aimed at enhancing the diagnostic accuracy of EGC, with particular emphasis on cases devoid of LNM.

Recently, deep learning has been widely discussed in the context of medical image analysis, including disease diagnosis, prognosis, and therapy^{14–16}. Several prior studies have employed deep learning methods

¹Department of Gastric Surgery, Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, Zhejiang, China.

²Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China. ³Department of Radiology, Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences, Hangzhou, Zhejiang, China. ⁴Key Laboratory of Prevention, Diagnosis and Therapy of Upper Gastrointestinal Cancer of Zhejiang Province, Hangzhou, Zhejiang, China. ⁵Zhejiang Provincial Research Center for Upper Gastrointestinal Tract Cancer, Zhejiang Cancer Hospital, Hangzhou, Zhejiang, China. ⁶Zhejiang Hospital of Traditional Chinese Medicine, Hangzhou, Zhejiang, China. ⁷Department of Gastric Surgery, Cancer Hospital of China Medical University, Cancer Hospital of Dalian University of Technology, Liaoning Cancer Hospital & Institute, Shenyang, Liaoning, China. ⁸These authors contributed equally: Jingyang He, Jingli Xu, Wujie Chen.

✉ e-mail: zhengboren1@126.com; pansw@zjcc.org.cn; hucan@zjcc.org.cn

to analyze CT data for predicting the pT1 stage in GC patients, and demonstrating high predictive accuracies^{17,18}. However, despite the fact that the CT models developed in the aforementioned studies demonstrated superior predictive ability compared to manual work, there are no studies that have utilized CT data to predict pT1N0 GC by concurrently considering the depth of invasion and the presence of LNM as a combined factor. Thus, this study aims to develop a deep learning model using portal venous phase CT images to accurately distinguish EGC without LNM via 2.5D radiomic data and a novel deep learning method.

Results

Clinical characteristics of cohorts

This study finally included a total of 3164 GC patients (Supplementary Fig. 1) from Zhejiang Cancer Hospital (the training cohort, $n = 1953$; the internal validation cohort, $n = 837$) and Zhejiang Hospital of Traditional Chinese Medicine (the external validation cohort, $n = 374$). Table 1 lists the detailed pathological features of all patients. The median age of all patients was 62 years (range: 21–92 years), including 2253 males (71.2%),

911 females (28.8%), 923 pT1N0 patients (29.2%), and 2241 pT1N + / pT2–4Nx patients (70.8%).

Models pre-trained at the slice level for diagnosing pT1N0 GC

In this study, we first evaluated the deep learning models constructed at the slice level for two distinct stages of the disease: pT1N0 GC and pT1N + / pT2–4Nx GC. A comprehensive workflow is employed as illustrated in Fig. 1. We analyzed the performance of Densenet121, inception_v3, ResNet101, ResNet50, and VGG19 to determine the best model for diagnosing T1N0 GC as the pre-training step. The results showed that the ResNet101 model represented the best performance for diagnosing pT1N0 GC with Area Under the Curve (AUC) 0.771 (95% CI: 0.763–0.779) in the training cohort, 0.784 (95% CI: 0.772–0.796) and 0.754 (95% CI: 0.735–0.774) in the internal and external validation cohorts, respectively. Furthermore, the training cohort had an accuracy of 74.3%, a sensitivity of 90.6%, and a specificity of 37.5%, the internal validation cohort had an accuracy of 74.9%, a sensitivity of 86.7%, and a specificity of 44.6%, meanwhile, the external validation cohort had an accuracy of 73.8%, a sensitivity of 85.0%, and a specificity of 39.6% (Table 2).

Table 1 | Clinicopathologic characteristics of patients in the three cohorts

Characteristics	Training cohort ($n = 1953$) Mean $\pm$ SD /N (%)	Internal validation cohort ($n = 837$) Mean $\pm$ SD /N (%)	External validation cohort ($n = 374$) Mean $\pm$ SD /N (%)	P-value*
Age	60.78 $\pm$ 10.55	61.17 $\pm$ 10.42	63.05 $\pm$ 10.46	<0.001
Height (cm)	164.02 $\pm$ 7.67	164.15 $\pm$ 6.99	162.27 $\pm$ 7.90	0.001
Weight (kg)	59.82 $\pm$ 10.06	59.76 $\pm$ 9.88	60.74 $\pm$ 10.59	0.169
BMI	22.17 $\pm$ 3.03	22.11 $\pm$ 2.94	23.05 $\pm$ 3.09	<0.001
CEA (ng/ml)	9.11 $\pm$ 69.00	8.36 $\pm$ 38.23	6.12 $\pm$ 20.78	0.766
CA199 (U/ml)	138.13 $\pm$ 1153.22	102.63 $\pm$ 632.49	70.69 $\pm$ 319.11	0.567
CA125 (U/ml)	14.75 $\pm$ 20.41	15.50 $\pm$ 39.17	15.03 $\pm$ 52.67	0.973
AFP (ng/ml)	8.17 $\pm$ 82.07	14.48 $\pm$ 183.18	7.81 $\pm$ 56.45	0.992
CA242 (IU/ml)	17.99 $\pm$ 74.23	17.63 $\pm$ 78.82	17.88 $\pm$ 75.62	0.629
CA724 (U/ml)	6.89 $\pm$ 21.60	6.68 $\pm$ 21.24	5.65 $\pm$ 18.11	0.933
CA50 (IU/ml)	14.51 $\pm$ 63.98	14.80 $\pm$ 55.73	8.12 $\pm$ 22.43	0.262
Gender				0.442
Female	572 (29.3)	225 (26.9)	114 (30.5)	
Male	1381 (70.7)	612 (73.1)	260 (69.5)	
Tumor location				<0.001
Proximal	413 (21.2)	190 (22.7)	78 (20.9)	
Middle	410 (21.0)	185 (22.10)	122 (32.6)	
Distal	1094 (56.0)	447 (53.41)	173 (46.2)	
Whole	36 (1.8)	15 (1.8)	1 (0.3)	
Pathological type				<0.001
Adenocarcinoma	1513 (77.5)	625 (74.7)	80 (21.4)	
Signet ring cell carcinoma	440 (22.5)	212 (25.3)	294(78.6)	
pT stage				<0.001
pT1	646 (33.1)	249 (29.7)	124 (33.1)	
pT2	111 (5.7)	59 (7.0)	24 (6.4)	
pT3	143 (7.3)	54 (6.5)	99 (26.5)	
pT4	1053 (53.9)	475 (56.8)	127 (34.0)	
pN stage				0.214
pN0	803 (41.1)	336 (40.1)	162 (43.3)	
pN1	244 (12.5)	122 (14.6)	54 (14.4)	
pN2	370 (19.0)	156 (18.6)	54 (14.4)	
pN3	536 (27.4)	223 (26.6)	104 (27.8)	

*P-values represent the results of the consistency analysis comparing the two medical centers.

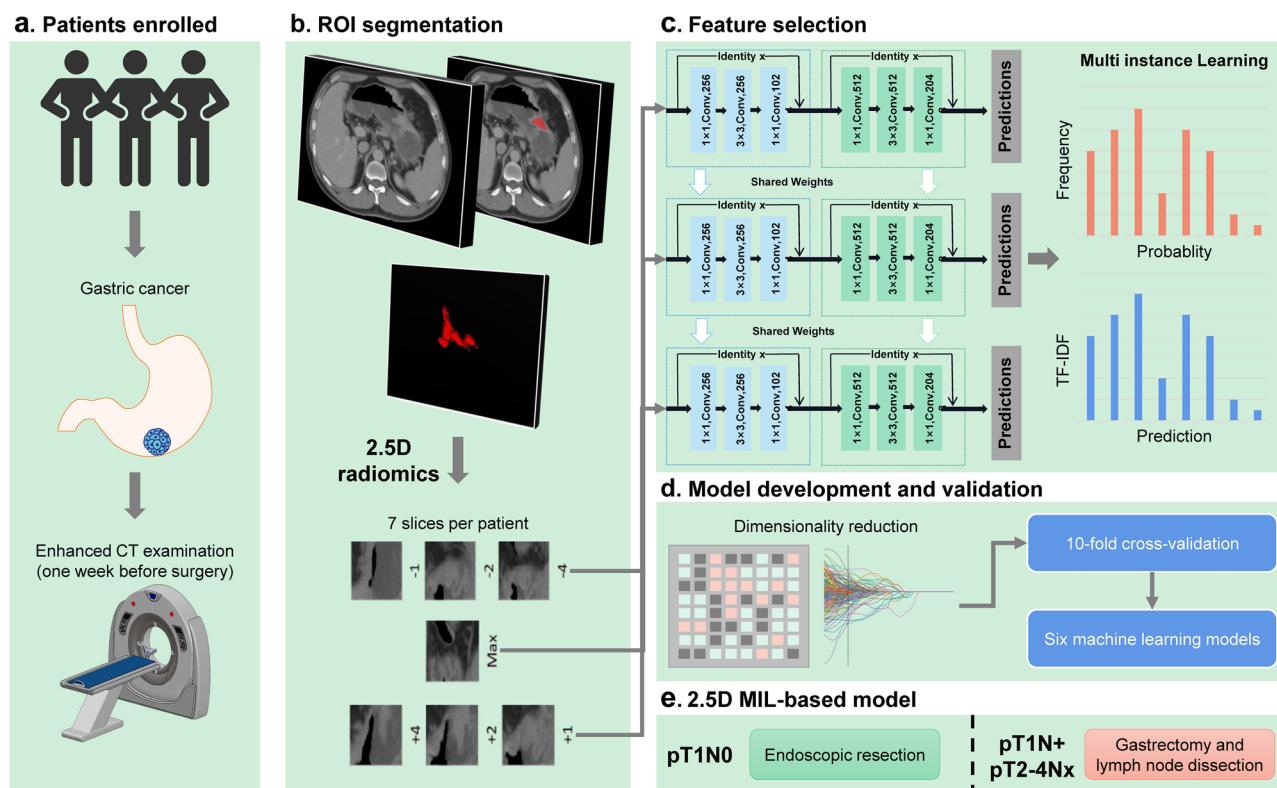


Fig. 1 | Pipeline of the 2.5D MIL-based model for predicting pT1N0 GC. a Patients enrolled; **b** ROI segmentation; **c** Feature selection; **d** Model development and validation; **e** 2.5D MIL-based model.

Table 2 | The performance of several well-known deep learning models

Model	Accuracy	AUC	95% CI	Sensitivity	Specificity	Cohort
Densenet121	0.754	0.786	0.778–0.794	0.909	0.400	Training
Densenet121	0.744	0.778	0.765–0.790	0.846	0.480	Internal validation
Densenet121	0.716	0.732	0.711–0.752	0.800	0.460	External validation
Inception_v3	0.739	0.765	0.757–0.773	0.918	0.333	Training
Inception_v3	0.749	0.777	0.765–0.790	0.873	0.429	Internal validation
Inception_v3	0.729	0.738	0.717–0.758	0.835	0.405	External validation
ResNet101	0.743	0.771	0.763–0.779	0.906	0.375	Training
ResNet101	0.749	0.784	0.772–0.796	0.867	0.446	Internal validation
ResNet101	0.738	0.754	0.735–0.774	0.850	0.396	External validation
ResNet50	0.745	0.778	0.770–0.786	0.911	0.369	Training
ResNet50	0.746	0.777	0.765–0.790	0.842	0.498	Internal validation
ResNet50	0.725	0.736	0.716–0.757	0.820	0.435	External validation
Vgg19	0.738	0.762	0.754–0.771	0.922	0.322	Training
Vgg19	0.742	0.779	0.767–0.792	0.845	0.475	Internal validation
Vgg19	0.724	0.741	0.721–0.761	0.806	0.474	External validation

2.5D MIL-based approach significantly improved the performance of the model

Next, we employed a multi-instance learning (MIL)-based approach for feature fusion to enhance the predictive accuracy of our models. This method integrates multiple data points or instances from a single sample to create a comprehensive feature set, which is essential for effectively analyzing and predicting complex clinical outcomes. In predicting both pT1N0 GC and pT1N+/pT2-4Nx GC, we achieved model effectiveness with scores exceeding 0.7. The results showed that the XGBoost model represented the best performance for diagnosing T1N0 GC with AUC of 0.920 (95% CI: 0.907–0.933), 0.774 (95% CI: 0.738–0.810) and 0.744

(95% CI: 0.686–0.803) in the training, internal and external validation cohorts, respectively (Fig. 2). Furthermore, the training cohort had an accuracy of 86.3%, a sensitivity of 84.9%, and a specificity of 89.6%, the internal validation cohort had an accuracy of 76.7%, a sensitivity of 76.5%, and a specificity of 77.4%, meanwhile, the external validation cohort had an accuracy of 74.9%, a sensitivity of 72.0%, and a specificity of 83.7% (Supplementary Table 1). After applying MIL, the ResNet101 model combined with the XGBoost model significantly improved its recognition results compared to slice-level deep learning models. This improvement is attributed to the robust design of the 2.5D concept and the substantial benefits of multi-instance learning.

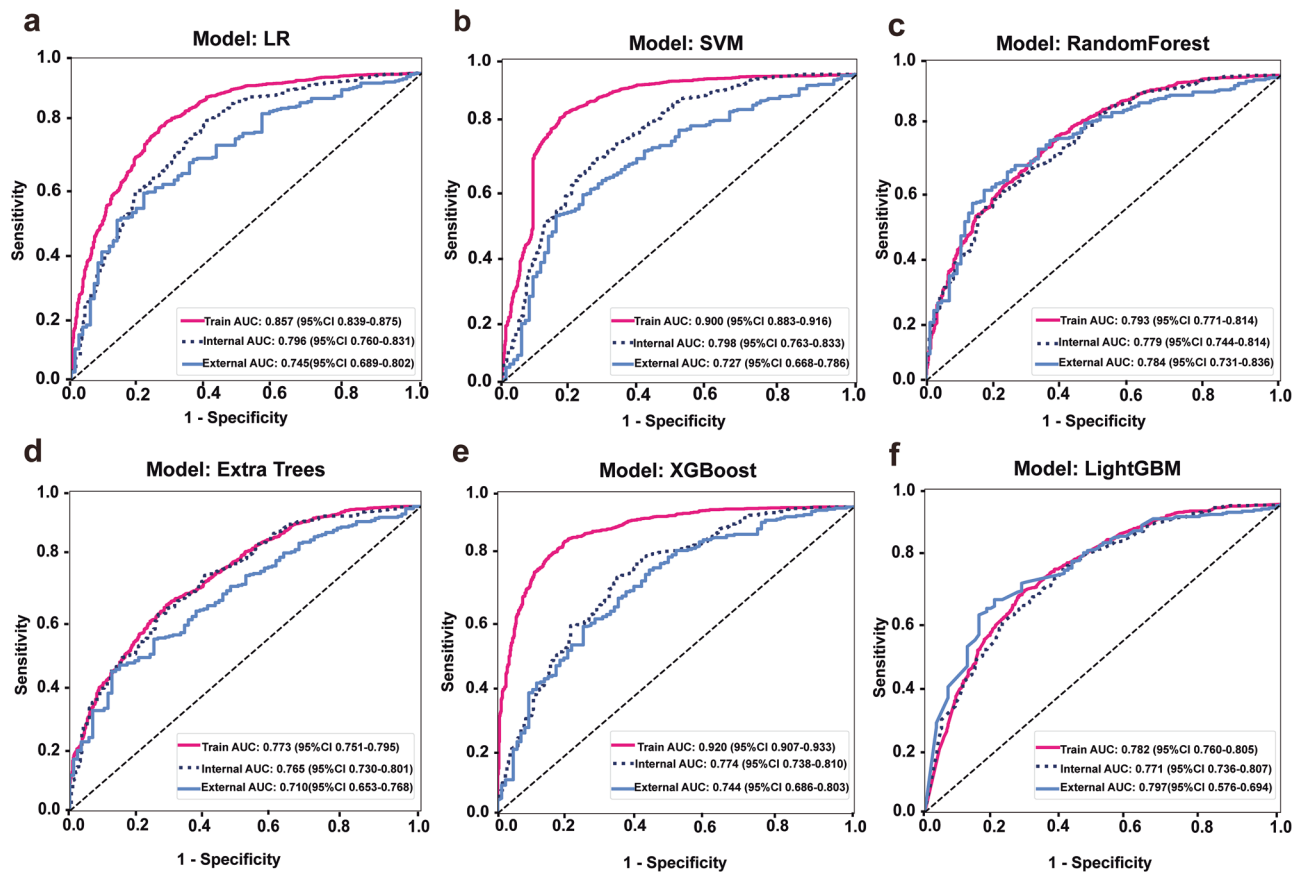


Fig. 2 | ROC curves of six 2.5D MIL-based machine learning models in each cohort. a Logistic Regression (LR); **b** Support Vector Machine (SVM); **c** Random Forest; **d** ExtraTrees; **e** XGBoost; **f** LightGBM.

Comparison among 2.5D MIL-based, traditional clinical, and radiomics models

In addition, we also compared the diagnostic performance of the 2.5D MIL-based model with both clinical models and traditional deep learning models. We observed that the 2.5D MIL-based model exhibited significantly better discriminatory ability than traditional deep learning radiomics models and the clinical models, with AUCs of 0.942, 0.900, and 0.767 on the training cohort (Fig. 3a), with AUCs of 0.841, 0.798, and 0.707 on the internal validation cohort (Fig. 3b), with AUCs of 0.822, 0.727, and 0.530 on the external validation cohort (Fig. 3c). Furthermore, the 2.5D MIL-based model had an accuracy of 68.5%, a sensitivity of 66.4%, and a specificity of 73.3%, in the train cohort, an accuracy of 76.7%, a sensitivity of 76.4%, and a specificity of 77.7%, in the internal validation cohort, meanwhile, the external validation cohort had an accuracy of 74.9%, a sensitivity of 72.0%, and a specificity of 83.7% (Supplementary Table 2). The confusion matrices presented also demonstrated the best diagnostic performance of the 2.5D MIL-based model across the three cohorts (Fig. 3d–f). The decision curve analysis (DCA) revealed that the 2.5D MIL-based model significantly enhanced the predicted probabilities for pT1N0 GC (Fig. 3g–i). Univariate and multivariate logistic regression analyses revealed that the 2.5D MIL-based model had much higher odds ratios (ORs) in predicting pT1N0 GC than traditional radiomics and clinical models in three cohorts (Table 3). This may suggest that the 2.5D MIL-based model possesses a greater potential for achieving net benefits, thereby underscoring its effectiveness in clinical decision-making contexts.

Visualization and clinical significance of the 2.5D MIL-based model

The aforementioned results had demonstrated that the 2.5D MIL-based model constructed in this study could effectively screen for patients with

pT1N0 GC using preoperative enhanced CT images. Based on the model visualization of Grad-CAM, we found that in a pT1N + /pT2-4Nx GC with obvious lesions, the model paid more attention to the lesions themselves (Fig. 4a). For pT1N0 GC, the model's focus was more on the area around the ROI. To further analyze the clinical significance of this model, we found that there were significant differences in the model scores among patients with different pT stages within the whole cohort, with scores significantly decreasing as the pT stage advanced (Fig. 4b). As for pN stage, the scores of pN+ patients were significantly lower than those with pN0 stage (Fig. 4c). Furthermore, patients were stratified into three equal groups based on their model scores, categorized as high, moderate, and low-scoring cohort. The risk assessment revealed that the probability of diagnosing pT1, pN0, and pT1N0 was 72.9%, 75.1%, and 69.0%, respectively, in the high-scoring cohort, all of which were significantly higher than those in the other two groups (Fig. 4d). Simultaneously, we also found that the prognosis of the high-scoring cohort was significantly better than that of the moderate and low scoring cohorts (Fig. 4e, $P < 0.001$, 5-year OS: 89.9% vs. 70.0% vs. 51.9%), and this trend was also observed in the population of patients with EGC (Fig. 4e, $P = 0.002$, 93.1% vs. 90.0% vs. 80.1%). In further analysis focusing on EGC, we similarly found that patients with LNM had higher scores from the model (Fig. 4f). And EGC patients with low scores had a high risk of LNM, reaching up to 23.8%, which was significantly higher than the other two groups (Fig. 4g).

Discussion

While it has been demonstrated that combined perioperative therapy can enhance the prognosis of GC patients, surgical or endoscopic resection still remains an essential component of the treatment strategy^{8,19}. Currently, minimally invasive, stomach-preserving, function-preserving, and

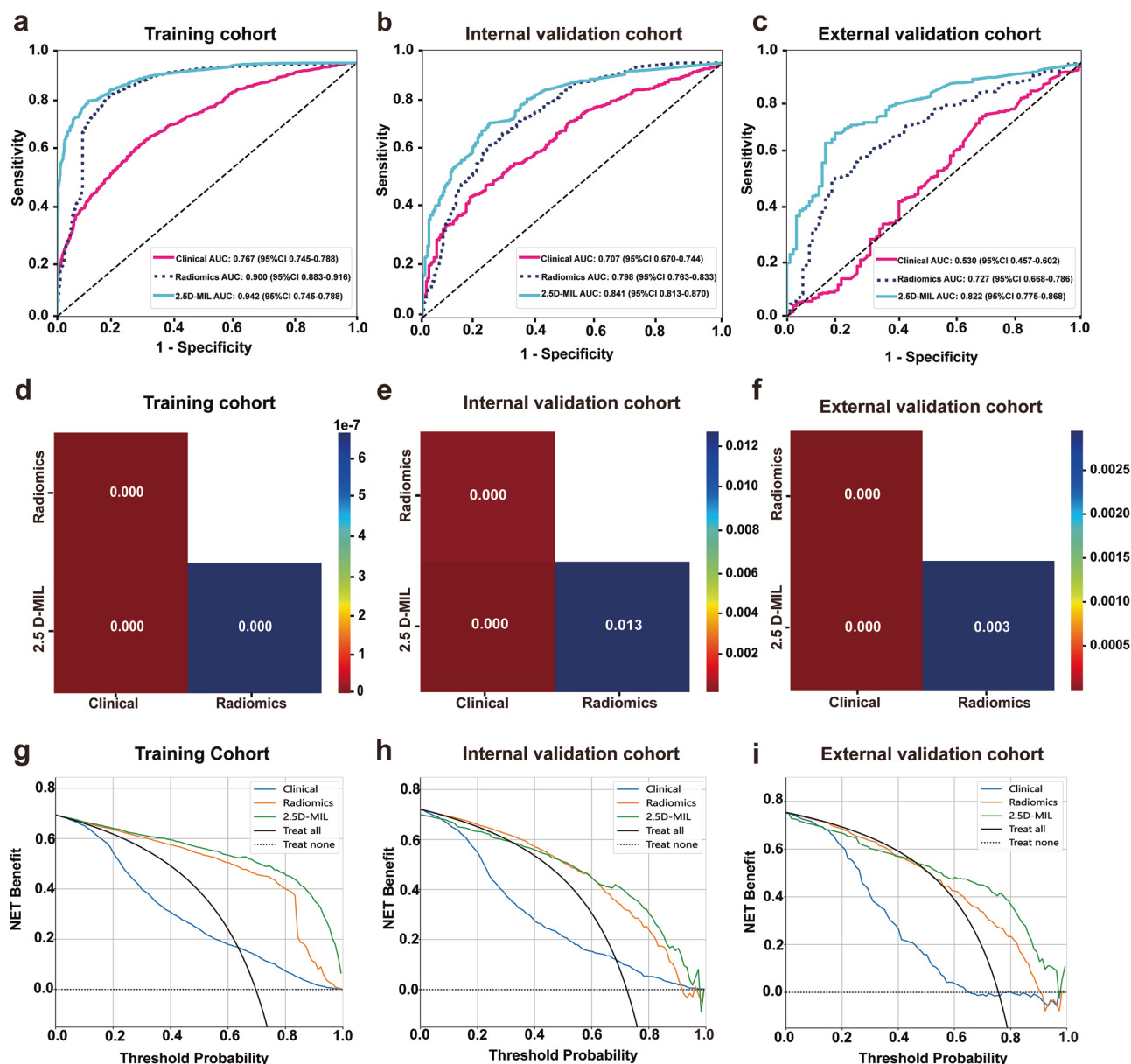


Fig. 3 | Different models' ROC analyses and the DeLong test of each cohort. ROC curves of the clinical model, traditional radiomics model and 2.5D MIL-based model for predicting pT1N0 GC in training cohort (a), internal validation cohort (b), external validation cohort (c); the DeLong test among the clinical model, traditional radiomics model and 2.5D MIL-based model in the training cohort (d), internal

validation cohort (e), and external validation cohort (f); Decision Curve Analysis of the clinical, traditional radiomic model and 2.5D MIL-based model for predicting pT1N0 GC in the training cohort (g), the internal validation cohort (h), and the external validation cohort (i).

Table 3 | Univariate and multivariate logistic regression analysis of the high and low risk scores of each model and pT1N0 GC

	Training cohort		Internal validation cohort		External validation cohort	
	OR (95%CI)	P-value	OR (95%CI)	P-value	OR (95%CI)	P-value
Univariate analysis						
2.5D MIL-based model (high vs. low)	77.702 (47.915–126.008)	<0.001	10.393 (7.166–15.073)	<0.001	11.947 (6.214–22.971)	<0.001
Clinical model (high vs. low)	10.414 (8.114–13.366)	<0.001	2.991 (2.173–4.117)	<0.001	1.390 (0.755–2.560)	0.291
Radiomics model (high vs. low)	29.270 (20.956–40.884)	<0.001	6.164 (4.302–8.830)	<0.001	4.902 (2.870–8.372)	<0.001
Multivariate analysis						
2.5D MIL-based model (high vs. low)	30.124 (18.130–50.053)	<0.001	6.291 (4.220–9.380)	<0.001	8.705 (4.405–17.203)	<0.001
Clinical model (high vs. low)	5.974 (4.336–8.230)	<0.001	1.897 (1.313–2.742)	0.001	1.034 (0.519–2.060)	0.924
Radiomics model (high vs. low)	12.107 (8.267–17.731)	<0.001	3.266 (2.195–4.860)	<0.001	2.626 (1.458–4.727)	0.001

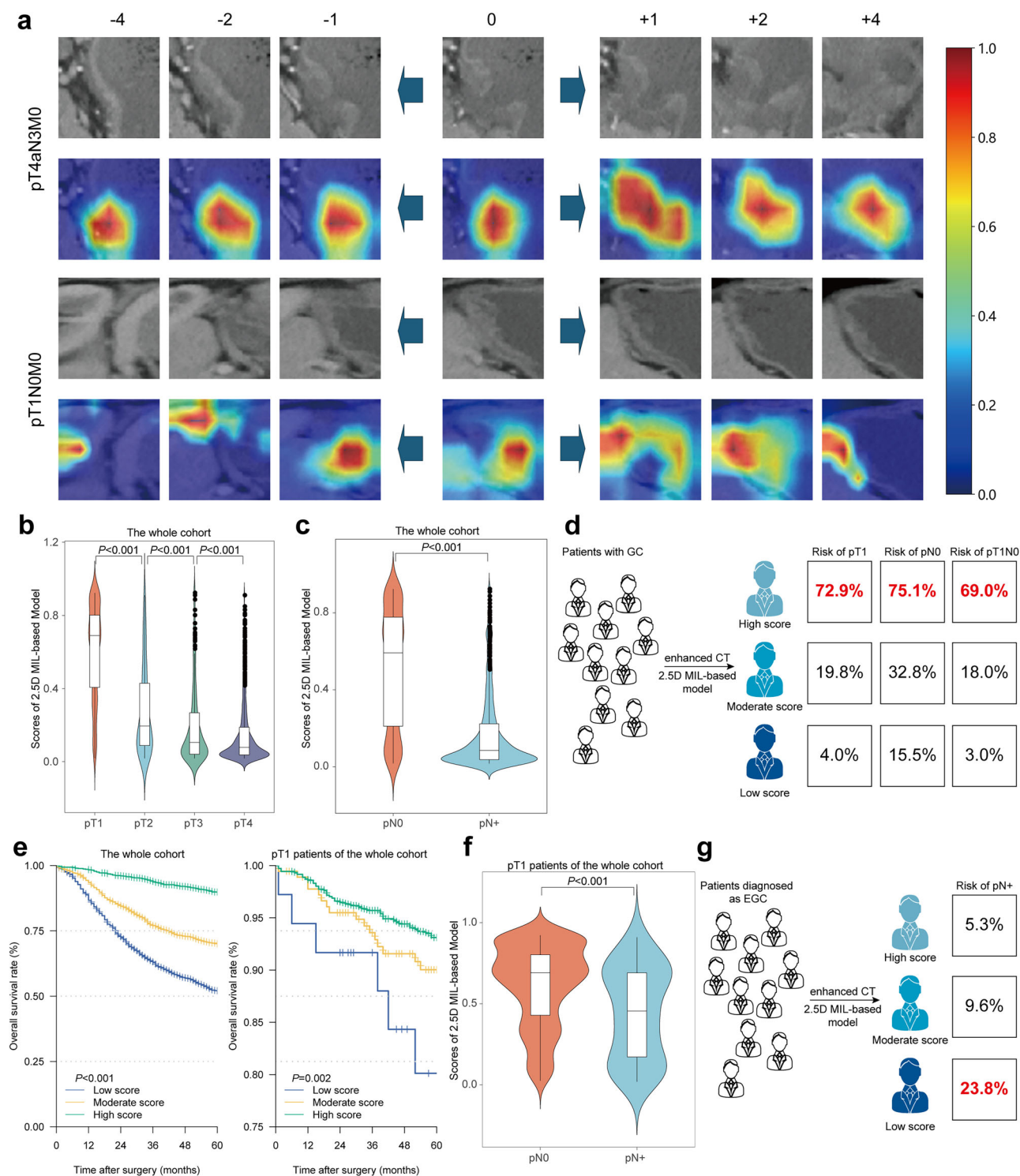


Fig. 4 | Visualization and clinical significance of the 2.5D MIL-based Model. Visualization of areas that the model deemed most relevant for making its predictions based on Grad-CAM (a) the model scores of patients with different pT stages (b) the model scores of patients with different pN stages (c) risk assessment in groups

with high, moderate or low scores (d) prognostic analysis of patients in different scoring groups of the whole cohort and the pT1 patients (e) the model scores of pT1 patients with different pN stages (f) risk of lymph node metastasis in patients with EGC in different scoring groups (g).

individualized treatments have become trends in global GC treatment^{20,21}. Clinically, determining the depth of invasion and lymph node status is crucial for assessing the curative potential of endoscopic treatment and surgery, especially in patients with EGC without LNM. Although only patients with GC in situ (Tis stage) and T1a GC without LNM can be successfully treated with endoscopic mucosal or submucosal resection, the incidence of regional LNM is limited in patients with T1b GC, in

which D2 gastrectomy might be excessively invasive and could lead to a significant waste of medical resources^{22,23}.

Previously, most studies have focused on the diagnosis of invasive depth or LNM^{24,25}. Accurate evaluation of the T stage is a crucial factor in determining treatment options and assessing prognosis. Under-staging may lead to incomplete tumor resection, while over-staging may result in unnecessary over-treatment. Previous studies^{26,27} have reported that the

accuracy of CT in diagnosing the T1 stage of GC ranges from 63.0% to 82.7%, while radiologists with varying levels of experience had different rates of misdiagnosing T1 as T2. Guan et al.²⁸ constructed an enhanced CT-based deep learning system to preoperatively predict the stage with an accuracy of 0.909. Zeng et al.²⁹ constructed a CNN prediction model for diagnosing EGC from GC patients based on the ResNet101 model, with a perfect accuracy rate of 91.4% in the external validation cohort. In addition, several studies explored the risk factors of LNM in EGC, which is another important factor affecting treatment choice. Zhu et al.³⁰ conducted a study on 2348 patients across five major tertiary medical centers, comparing six machine learning technologies, and reported that XGBoost demonstrated the best performance in predicting LNM in EGC. However, although the CT models constructed in the aforementioned studies all suggested a prominent predictive ability than manual work, their techniques were traditional and exclusively predicted invasive depth or LNM. With the ongoing advancements in deep learning technology, the utilization of micro-information from CT images to predict macro or micro-information of the disease has increasingly become a mainstream approach in precision medicine³¹. Several studies have successfully predicted microsatellite instability status (AUC = 0.78)³², Ki-67 expression (AUC = 0.84)³³, programmed cell death-ligand 1 expression (AUC = 0.98)³⁴, and other microscopic variables using radiomic data.

In this study, unlike previous studies, the depth of invasion and LNM were considered at the same time, and we constructed a deep learning prediction model based on 2.5D radiomics and MIL, which were novel technologies to get more information, for pT1N0 GC screening. We extracted adjacent slices in both the superior-inferior and anterior-posterior directions from the central slice; the range of adjacent slices was established at ± 1 , ± 2 , and ± 4 slices, resulting in a 2.5D sight (7 slices per patient). The aim was to maximize the possibility of finding early gastric cancer and to get more information from the model. This approach allows our 2.5D model to capture more dimensions than 2D. Choi et al.³⁵ developed a non-invasive model for predicting glioma IDH status using deep learning and radiomics, incorporating ± 2 and ± 4 slices, totaling 5 slices. The hybrid model achieved an accuracy of 93.8% in the training group. Li et al.³⁶ constructed a 2.5D model using the ResNet18 deep learning architecture, incorporating ± 2 slices, totaling 3 slices. This model also demonstrated superior prediction performance compared to traditional radiomics. The 3D model has the advantage of comprehensive information; however, it also suffers from challenges of poor model fitting and external validation. Prior research had indicated that 2.5D models might exhibit superior performance compared to 3D models. Kim et al.³⁷ developed deep learning models using 2.5D and 3D CT images. They observed that the 2.5D DenseNet achieved a significantly higher AUC of 0.921 compared to 0.835 of the 3D DenseNet ($P = 0.037$). Another study for predicting subsequent fracture risk in patients with hip fracture also showed a higher C-index of the 2.5D model than the 3D model (training cohort: 0.73 vs. 0.60; test cohort: 0.73 vs. 0.59)³⁸. The 2.5D model leverages both the spatial information from 2D slices and the additional depth information from multiple slices, without the computational intensity and data requirements of full 3D models. This approach allows for a more efficient and practical application in clinical settings, where rapid and accurate diagnosis is crucial. It has the potential to capture the complexity of anatomical structures and disease patterns that may not be fully appreciated in 2D images, while avoiding the challenges associated with 3D data, such as increased computational load and the need for more data to train the model.

Our 2.5D MIL-based model showed a significant performance in recognizing pT1N0 GC patients, which achieved model effectiveness exceeding a 0.7 score, especially in the XGBoost model. The predictive performance of this model was significantly superior to that of models developed using traditional radiomic techniques and those based exclusively on clinical data, with AUCs of 0.942, 0.900, and 0.767 on the training cohort, 0.841, 0.798, and 0.707 on the internal validation cohort, and 0.822, 0.727, and 0.530 on the external validation cohort. The consistent statistical analysis results between the two centers showed that there were some

distributional differences in patient characteristics between the two medical centers. However, we believe these differences do not significantly conflict with the positive results of model construction or external validation. In real-world clinical environments, patient characteristics may vary across different centers. However, despite this variability, our model still demonstrated a relatively high AUC, which further highlights the generalizability and high clinical value of our model. Moreover, the overall accuracy of our 2.5D model surpassed that of both clinical and traditional models, and it demonstrated a superior capability in accurately identifying pT1N0 patients. Kim et al.³⁸ showed that for 2-, 3-, and 5-year fracture predictions, the 2.5-dimensional ensemble model showed AUC of 0.78, 0.76, and 0.77, respectively, which was significantly higher than the other five image-based traditional radiomics models (AUC 0.57–0.71), and the combined model obtained a higher AUC than the clinical model containing well-known risk factors (AUC of 0.58–0.70,) ($P < 0.001$). Furthermore, the confusion matrices presented also demonstrated the best diagnostic performance of the 2.5D MIL-based model across the three cohorts. The DCA also revealed that the 2.5D MIL-based model significantly enhanced the predicted probabilities for pT1N0 GC, and this suggests that the 2.5D MIL-based model possesses a greater potential for achieving net benefits, thereby underscoring its effectiveness in clinical decision-making contexts. Firstly, we emphasize the model's ability to provide consistent and objective assessments, which may complement the subjective interpretations of radiologists, and its potential to assist in identifying patterns that may be difficult for the human eye to detect, in cases where expert radiologists are not readily available or where a second opinion is needed, our model could support radiologists in making more informed decisions. Our 2.5D MIL-based model, on the other hand, offers several advantages over the traditional methods. It leverages machine learning algorithms to analyze complex datasets, potentially providing a more nuanced and personalized diagnostic approach. Additionally, our model can be trained on a diverse range of data types and readily adapt to new data, a capability not always feasible with traditional methods.

Finally, we have undertaken a comprehensive examination of the clinical utility of this 2.5D MIL-based model. Our findings indicated that the model was not only proficient in predicting pT1N0 GC but also demonstrated exceptional screening capabilities for pT1 or pN0 individually, with high-scoring cohorts comprising 72.9% and 75.1%, respectively. Additionally, the Japanese gastric cancer guidelines indicate that the risk of LNM can be predicted based on post-endoscopic submucosal dissection (ESD) pathological findings, categorizing patients into eight distinct types according to a scoring system, with the risk ranging from 1.6% to 27.3%^{39,40}. In our model, the risk of LNM could be assessed exclusively using pre-operative CT data, with the incidence of LNM in the low-scoring cohort reaching up to 23.8%. The accuracy of radiologists in diagnosing the T1 stage ranged from 63.0% to 82.7%. The overall accuracy of EUS in diagnosing the depth of invasion in early gastric cancer was 75.7%⁴¹; however, its ability to detect lymph node metastasis was limited. Our model exhibits a higher detection rate than conventional CT interpretation, with early detection rates that are equal to or exceed those of EUS, and it can predict lymph node metastases. In contrast, our model provides superior screening diagnostic capabilities, offering greater convenience and ease of implementation. More accurate prediction of the depth of invasion and LNM can guide surgeons in planning more appropriate surgical interventions, potentially reducing the risk of recurrence and improving survival rates. This model can also assist in personalized treatment planning and monitoring, which are essential components of modern oncological care.

However, this study has several limitations. Firstly, despite the large and multicenter sample size, the retrospective nature and distribution changes may introduce inherent patient selection bias. The results of the concordance analysis showed some differences in distribution between multicenter patient characteristics, which may have affected the validation results of our model. This uncertainty hampers the generalizability of our findings across different populations. Consequently, prospective studies with careful design are essential to further validate the clinical applicability

of the model. Secondly, all patients in this study were diagnosed with gastric cancer, and their lesions were detectable by CT images. Consequently, this model is not applicable to EGC patients whose lesions are undetectable by CT images preoperatively, potentially limiting its clinical utility in certain scenarios.

In summary, we first constructed a deep learning prediction model based on 2.5D radiomics and MIL for effectively diagnosing pT1N0 GC patients, which provides valuable information for the individualized treatment selection. However, future prospective studies are needed to confirm the clinical utility of this model.

Methods

Patients

This study included patients from two separate large medical centers in China, including those who underwent radical gastrectomy at Zhejiang Cancer Hospital between January 2006 and December 2019 and those who underwent radical gastrectomy at Zhejiang Hospital of Traditional Chinese Medicine between January 2014 and December 2016. Patients were enrolled according to the following criteria: (1) pathological diagnosis was gastric adenocarcinoma or signet ring cell carcinoma; (2) patients with radical gastrectomy combined with R0 resection and D2 lymphadenectomy; (3) abdominal enhanced CT examination was performed within one week before surgery; (4) complete clinicopathological data and follow-up data (regular outpatient follow-up and telephone follow-up was up to May 2024); (5) image quality meets analysis requirements: lesions can be seen on CT images, the gastric cavity is sufficiently dilated, and no breathing and peristaltic artifacts are observed in the images. Patients were excluded according to the following criteria: (1) incomplete clinicopathological data; (2) combined with other malignant tumors; (3) gastric distention was not ideal, and it was difficult to segment the tumor; (4) artifacts on CT images seriously affected the observation of LNs. Clinicopathological data of the patients were collected, including age, gender, height, weight, BMI, CEA, CA199, CA125, AFP, CA242, CA724, CA50, tumor location, pathological type, and postoperative pathological stage. All patients were staged according to the latest version of the International Union for Cancer Control and the American Joint Committee on Cancer.

This study was designed as a multicenter, retrospective study, and all patients were given informed consent. This study has been approved by the Ethics Committee of Zhejiang Cancer Hospital (IRB-2024-220) and Zhejiang Hospital of Traditional Chinese Medicine (2022-KL-102-04), and the study was conducted in accordance with the principles of the Declaration of Helsinki. In order to protect patients' privacy and comply with relevant privacy protection regulations, we performed metadata cleaning on all DICOM files of CT images used in this study. The cleaning process aims to remove all sensitive personal information while retaining technical information related to medical image processing. First, use the PyDICOM library to read the original DICOM file. Remove sensitive information; meanwhile, the technical information related to the imaging parameters, such as the scanning sequence, pixel pitch, etc., is retained. Then save the cleaned DICOM file to the specified directory. Through the above cleaning process, we successfully removed all sensitive personal information while retaining the necessary technical information. The cleaned DICOM files were used for subsequent medical image analysis, ensuring the compliance of the research and the protection of patient privacy.

Tumor segmentation

All patients underwent enhanced CT examination within one week before surgery. Tumor segmentation was conducted by three experienced radiologists using ITK-SNAP software (version 3.8, <http://www.itksnap.org>). Since GC can be clearly distinguished from surrounding normal gastric tissue in enhanced portal phase CT images, all mapping was performed along tumor boundaries in these images. Additionally, considering the holistic nature of the tumor, this study mapped tumor lesions at all levels on CT images. The workload for tumor segmentation was divided equally

among the three experienced radiologists. After each radiologist completed the segmentation for their assigned cases, a second radiologist independently reviewed and verified the segmentation. Any discrepancies or disagreements were discussed and resolved through consensus. In cases where there were disagreements or discrepancies during the internal verification process, a third radiologist was involved to provide an additional opinion and help reach a consensus.

Data preprocessing

To mitigate the impact of extreme values and outliers, pixel intensities were limited to a range of -200 to 400 . This normalization was essential for homogenizing the dataset, thereby improving its consistency and reliability for subsequent analyses. We also addressed the issue of variable voxel spacing in different volumes of interest (VOIs). A fixed resolution resampling technique was employed to standardize voxel spacing to a uniform size of $1\text{ mm} \times 1\text{ mm} \times 1\text{ mm}$. This approach ensured uniform spatial dimensions across images, facilitating more accurate and consistent comparisons and analyses.

2.5D deep learning radiomics

Our study presents a novel approach called 2.5D deep learning radiomics, which differs from traditional 2D deep learning models that typically use only the maximal cross-sectional slice for model training with 2D convolution. In our research, we adopted the 2.5D concept, which includes the maximal cross-sectional slice and samples several layers above and below it. This approach not only generates a larger set of training samples but also utilizes additional contextual information to achieve improved outcomes.

Data generation

The process started by identifying the Region of Interest (ROI) within the raw 3D image data, specifically focusing on the largest cross-sectional slice. To construct a series of 2D images, adjacent slices were extracted in both the superior-inferior and anterior-posterior directions from this central slice. The range of these adjacent slices was set at ± 1 , ± 2 , and ± 4 slices, resulting in a set of 2D images (7 slices per patient) centered around the maximal cross-sectional slice of the ROI. This approach, which effectively captures partial 3D structural information, is referred to as 2.5D data. During the cropping process, the ROI was expanded by 10 mm to include peri-tumoral information. Cropping was performed using the OKT-crop_max_roi tool with parameters set to surrounds +1, +2, -1, -2, +4, -4, and enlarge 10, effectively capturing an enhanced cross-sectional context of the ROI.

Model training

For model training, the generated 2.5D data were integrated into a transfer learning framework. We evaluated the performance of several well-known deep learning models: DenseNet121, InceptionV3, ResNet50, ResNet101, and VGG19, all pre-trained on the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC-2012) dataset. In our study, we implemented Z-score normalization to ensure a uniform intensity distribution across RGB channels in the images. The normalized images were then utilized as input for our model. During the training phase, we applied real-time data augmentation techniques, including random cropping, horizontal flipping, and vertical flipping. For the test images, we conducted only normalization, without any augmentation. To achieve consistency across the dataset, we normalized the grayscale values of the slices using a min-max transformation, setting the range to $[-1, 1]$. Each cropped subregion image was resized to 224×224 pixels (or 299×299 pixels for InceptionV3), employing nearest neighbor interpolation. This resizing was necessary to align with the input specifications of the selected deep learning models.

Considering the specific characteristics of our image dataset, we carefully adjusted the learning rate to enhance model generalization. For this, we employed a cosine decay strategy for the learning rate. Learning rate

is details as follows:

$$\eta_t = \eta_{\min}^i + \frac{1}{2}(\eta_{\max}^i - \eta_{\min}^i) \left(1 + \cos\left(\frac{T_{cur}}{T_i}\pi\right) \right) \quad (1)$$

The minimum learning rate, $\eta_{\min}^i$, is set to 0, while the maximum learning rate, $\eta_{\max}^i$, is set at 0.01. The parameter T_i denotes the number of iterations. Other hyperparameters are configured as follows: the optimizer is SGD (Stochastic Gradient Descent), and the loss function used is softmax cross-entropy.

Multi-instance learning

We employed a multi-instance learning-based approach for feature fusion, aiming to enhance the predictive accuracy of our models. This method involves integrating various data points or instances from a single sample to formulate a comprehensive feature set. Rather than averaging probabilities, we treated each of the 7 slices representing a patient as an individual instance and applied MIL techniques to aggregate the recognition results of these slices. Given the variety of models used in our 2D deep learning approach, we fused only the results from the best-performing models in the validation set. Then, we utilized the deep learning model to predict each patch, obtaining corresponding probabilities and labels, denoted as $Patch_{prob}$ and $Patch_{pred}$ respectively.

Multi-instance learning feature aggregation

Histogram feature aggregation: (1) We treated each distinct number as a “bin” and counted the occurrence of each type of data across these bins. (2) The frequencies of $Patch_{prob}$ and $Patch_{pred}$ falling into each bin were tallied. (3) All features underwent min-max normalization. (4) This process resulted in the generation of $Histo_{prob}$ and $Histo_{pred}$.

Bag of Words (BoW) Feature Aggregation: (1) Initially, a dictionary was created by identifying unique elements within $Patch_{prob}$ and $Patch_{pred}$. (2) Each patch was then represented as a vector, where the frequency of each dictionary element in the patch was noted. (3) We applied Term Frequency-Inverse Document Frequency (TF-IDF) transformation to these vectors, emphasizing the importance of less frequent but more informative features. (4) This resulted in a BoW feature representation for each patch, encapsulating both the presence and significance of features within a patch. (5) The final BoW features, denoted as BoW_{prob} and BoW_{pred} , offered a comprehensive and weighted representation of the patches, suitable for subsequent analytical processes.

The final stage in our multi-instance learning-based feature fusion involves the integration of the previously derived features: $Histo_{prob}$, $Histo_{pred}$, BoW_{prob} , and BoW_{pred} . To achieve this, we employ a feature concatenation method, symbolized by $\oplus$, which combines these individual feature sets into a single, comprehensive feature vector. The specific formula for this concatenation is as follows:

$$feature_{fusion} = Histo_{prob} \oplus Histo_{pred} \oplus Bow_{prob} \oplus Bow_{pred} \quad (2)$$

2.5D deep learning signature

For the features derived from MIL, we first performed feature selection and dimensional reduction, followed by modeling using machine learning algorithms. We standardized features using Z-scores and conducted t-tests to calculate P-values, retaining features with P-values < 0.05. LASSO regression was applied to finalize the feature set of the radiomic signature by pushing the coefficients of irrelevant features towards zero. The optimal regularization parameter λ was determined through 10-fold cross-validation. The final features were applied to six machine learning models, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest, ExtraTrees, XGBoost, and LightGBM, to construct the prediction model. These models were independently trained, utilizing a GridSearch algorithm for

hyperparameter tuning, with the best model parameters retained based on validation set performance. Our approach is referred to as 2.5D MIL.

Traditional radiomics and clinical models

In our research, handcrafted radiomic features were classified into three primary categories: (I) Geometry, (II) Intensity, and (III) Texture. For the extraction of texture features, we employed various techniques including the gray-level co-occurrence matrix (GLCM), gray-level run length matrix (GLRLM), gray-level size zone matrix (GLSZM), and neighborhood gray-tone difference matrix (NGTDM). Feature extraction for each subregion was conducted using the pyradiomics tool (version 3.0.1), in compliance with the Imaging Biomarker Standardization Initiative (IBSI) guidelines. The process of feature selection for our radiomic signature was similar to that employed in the construction of the 2.5D Deep Learning Signature.

Moreover, we also conducted an analysis of the correlations among various clinical features. And we proceeded to incorporate all the identified features into the construction of a comprehensive clinical model. This model aims to integrate and utilize the diverse range of clinical features to enhance the understanding and predictive capability in the clinical context.

Model comparison

To validate the performance of our proposed 2.5D deep learning framework, we included clinical models (Clinical) and traditional radiomic models (Radiomics). The diagnostic efficacy of the 2.5D MIL model was evaluated in the test cohort using the receiver operator characteristic (ROC) curves, DeLong test and univariate and multivariate logistic regression analyses, while DCA was used to assess the clinical utility of the predictive models.

Gradient-weighted class activation mapping (Grad-CAM)

To enhance the interpretability of our 2.5D deep learning model, we employed Grad-CAM. This technique produces a localization map highlighting the regions of significance for the classification task. We rendered the last convolutional layer of the final residual block transparent with respect to the prediction task. This approach enabled us to visualize and identify the areas that the model deemed most relevant for making its predictions.

Statistical analysis

We used various statistical tests to compare the clinical characteristics of patients. The χ^2 test was used for discrete variables, and the independent sample t-test was applied to continuous variables. All data analyses were performed using Python version 3.7.12. For statistical analyses, statsmodels version 0.13.2 was used. The pyradiomics package version 3.0.1 was employed for extracting radiomic features. Machine learning algorithms were implemented using scikit-learn version 1.0.2. Deep learning models were developed based on torch version 1.11.0, with cuda11.3.1 and cudnn8.2.1.

Data availability

Due to the privacy of patients, the data related to patients cannot be available for public access but can be obtained from the corresponding author on reasonable request approved by the institutional review board of all enrolled centers.

Code availability

The underlying code for this study is available on GitHub and can be accessed via this link: <https://github.com/OnekeyAI-Platform/2.5DMulti3>.

Received: 17 January 2025; Accepted: 13 July 2025;

Published online: 23 July 2025

References

1. Sung, H. et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Yao, K. et al. Guidelines for endoscopic diagnosis of early gastric cancer. *Dig. Endosc.* **32**, 663–698 (2020).
3. Lordick, F. et al. Gastric cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann. Oncol.* **33**, 1005–1020 (2022).
4. Liu, Q. et al. Updated evaluation of endoscopic submucosal dissection versus surgery for early gastric cancer: a systematic review and meta-analysis. *Int. J. Surg.* **73**, 28–41 (2020).
5. Songun, I. et al. Surgical treatment of gastric cancer: 15-year follow-up results of the randomised nationwide Dutch D1D2 trial. *Lancet Oncol.* **11**, 439–449 (2010).
6. Wu, C. ZHU Z. Diagnosis and evaluation of gastric cancer by positron emission tomography. *World J. Gastroenterol.* **20**, 4574–4585 (2014).
7. Lee, K. et al. Can endoscopic ultrasonography (EUS) improve the accuracy of clinical T staging by computed tomography (CT) for gastric cancer?. *Eur. J. Surg. Oncol.* **47**, 1969–1975 (2021).
8. Ajani, J. A. et al. Gastric Cancer, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Cancer Netw.* **20**, 167–192 (2022).
9. Guo, J. et al. A prospective analysis of the diagnostic accuracy of 3 T MRI, CT and endoscopic ultrasound for preoperative T staging of potentially resectable esophageal cancer. *Cancer Imaging* **20**, 64 (2020).
10. Arslan, H. et al. Contribution of diffusion weighted MRI to diagnosis and staging in gastric tumors and comparison with multi-detector computed tomography. *Radio. Oncol.* **51**, 23–29 (2017).
11. Ba-Ssalamah, A. et al. Accuracy of hydro-multidetector row CT in the local T staging of oesophageal cancer compared to postoperative histopathological results. *Eur. Radio.* **21**, 2326–2335 (2011).
12. Gai, Q. et al. Clinical significance of multi-slice spiral CT, MRI combined with gastric contrast-enhanced ultrasonography in the diagnosis of T staging of gastric cancer. *Clin. Transl. Oncol.* **23**, 2036–2045 (2021).
13. Mocellin, S. & PASQUALI, S. Diagnostic accuracy of endoscopic ultrasonography (EUS) for the preoperative locoregional staging of primary gastric cancer. *Cochrane Database Syst. Rev.* **2015**, CD009944 (2015).
14. Dong, D. et al. Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. *Ann. Oncol.* **31**, 912–920 (2020).
15. Cui, Y. et al. A CT-based deep learning radiomics nomogram for predicting the response to neoadjuvant chemotherapy in patients with locally advanced gastric cancer: A multicenter cohort study. *EClinicalMedicine* **46**, 101348 (2022).
16. Gao, X. et al. A CT-based radiomics model for prediction of lymph node metastasis in early stage gastric cancer. *Acad. Radio.* **28**, e155–e164 (2021).
17. Lee, I. J. et al. Diagnostic performance of 64-channel multidetector CT in the evaluation of gastric cancer: differentiation of mucosal cancer (T1a) from submucosal involvement (T1b and T2). *Radiology* **255**, 805–814 (2010).
18. Makino, T. et al. Preoperative T staging of gastric cancer by multi-detector row computed tomography. *Surgery* **149**, 672–679 (2011).
19. Zhang, X. et al. Perioperative or postoperative adjuvant oxaliplatin with S-1 versus adjuvant oxaliplatin with capecitabine in patients with locally advanced gastric or gastro-oesophageal junction adenocarcinoma undergoing D2 gastrectomy (RESOLVE): an open-label, superiority and non-inferiority, phase 3 randomised controlled trial. *Lancet Oncol.* **22**, 1081–1092 (2021).
20. Choi, J. et al. Safety of pylorus-preserving gastrectomy for gastric cancer combined with antral high-risk lesions: a comparison with endoscopic submucosal dissection. *Surg. Endosc.* **37**, 2987–2996 (2023).
21. Xiao, X. M. et al. Pylorus-preserving versus distal subtotal gastrectomy for surgical treatment of early gastric cancer: a meta-analysis. *Hepatogastroenterology* **61**, 870–879 (2014).
22. Vos, E. L. et al. Risk of lymph node metastasis in T1b gastric cancer: an international comprehensive analysis from the Global Gastric Group (G3) Alliance. *Ann. Surg.* **277**, e339–e345 (2023).
23. Akiyama, Y. et al. Frequency of lymph node metastasis according to tumor location in clinical T1 early gastric cancer: supplementary analysis of the Japan Clinical Oncology Group study (JCOG0912). *J. Gastroenterol.* **58**, 519–526 (2023).
24. Jiang, M. et al. Value of multi-slice spiral computed tomography in the diagnosis of metastatic lymph nodes and N-stage of gastric cancer. *J. Int. Med. Res.* **47**, 281–292 (2019).
25. Liu, B. et al. The predictive potential of contrast-enhanced computed tomography based radiomics in the preoperative staging of cT4 gastric cancer. *Quant. Imaging Med. Surg.* **12**, 5222–5238 (2022).
26. Wang, Z. L. et al. Utility of the gastric window in computed tomography for differentiation of early gastric cancer (T1 stage) from muscularis involvement (T2 stage). *Abdom. Radio.* **46**, 1478–1486 (2021).
27. Xie, Z. Y. et al. T and N staging of gastric cancer using dual-source computed tomography. *Gastroenterol. Res. Pract.* **2018**, 5015202 (2018).
28. Guan, X., Lu, N. & Zhang, J. Accurate preoperative staging and HER2 status prediction of gastric cancer by the deep learning system based on enhanced computed tomography. *Front. Oncol.* **12**, 950185 (2022).
29. ZENG, Q. et al. Deep learning model for diagnosing early gastric cancer using preoperative computed tomography images. *Front. Oncol.* **12**, 1065934 (2022).
30. Zhu, H. et al. Preoperative prediction for lymph node metastasis in early gastric cancer by interpretable machine learning models: a multicenter study. *Surgery* **171**, 1543–1551 (2022).
31. Lambin, P. et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **14**, 749–762 (2017).
32. Bodalal, Z. et al. Non-invasive CT radiomic biomarkers predict microsatellite stability status in colorectal cancer: a multicenter validation study. *Eur. Radio. Exp.* **8**, 98 (2024).
33. Li, B. et al. Radiomics nomogram based on CT radiomics features and clinical factors for prediction of Ki-67 expression and prognosis in clear cell renal cell carcinoma: a two-center study. *Cancer Imaging* **24**, 103 (2024).
34. Liu, Z. et al. Radiomics models derived from arterial-phase-enhanced CT reliably predict both PD-L1 expression and immunotherapy prognosis in non-small cell lung cancer: a retrospective, multicenter cohort study. *Acad. Radiol.* **32**, 493–505 (2025).
35. Choi, Y. S. et al. Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics. *Neuro Oncol.* **23**, 304–313 (2021).
36. Li, Y. et al. Molecular subtyping of diffuse gliomas using magnetic resonance imaging: comparison and correlation between radiomics and deep learning. *Eur. Radiol.* **32**, 747–758 (2022).
37. Kim, H. et al. CT-based deep learning model to differentiate invasive pulmonary adenocarcinomas appearing as subsolid nodules among surgical candidates: comparison of the diagnostic performance with a size-based logistic model and radiologists. *Eur. Radiol.* **30**, 3295–3305 (2020).
38. Kim, Y. et al. A CT-based deep learning model for predicting subsequent fracture risk in patients with hip fracture. *Radiology* **310**, e230614 (2024).

39. Hatta, W. et al. A scoring system to stratify curability after endoscopic submucosal dissection for early gastric cancer: “eCura system”. *Am. J. Gastroenterol.* **112**, 874–881 (2017).
40. Japanese Gastric Cancer Association. Japanese Gastric Cancer Treatment Guidelines 2021 (6th edition). *Gastric Cancer* **26**, 1–25 (2023).
41. Tsujii, Y. et al. Diagnostic value of endoscopic ultrasonography for the depth of gastric cancer suspected of submucosal invasion: a multicenter prospective study. *Surg. Endosc.* **37**, 3018–3028 (2023).

Acknowledgements

We appreciate the great technical support from the Department of Radiology of Zhejiang Cancer Hospital. This study was supported by The National Key Research and Development Program of China (2021YFA0910100), Healthy Zhejiang One Million People Cohort (K-20230085), Post-doctoral Innovative Talent Support Program (BX2023375), National Natural Science Foundation of China (82304946, 82473489, 82403546), Natural Science Foundation of Zhejiang Province (LR21H280001, MS25H160146, TGY23H160038), and The Medicine and Health Science Fund of Zhejiang Province (2025KY047, 2023KY073).

Author contributions

Guarantors of integrity of entire study: C.H., S.W.P., G.L.Z.; study concepts/ study design: C.H., S.W.P., G.L.Z., J.Y.H., J.L.X.; data acquisition: all authors; contributed data or analysis tools: J.Y.H., J.L.X., W.J.C.; data analysis, data interpretation: J.Y.H., J.L.X., W.J.C.; manuscript drafting: J.Y.H., J.L.X.; manuscript revision for important intellectual content: C.H., S.W.P., G.L.Z., J.Y.H., J.L.X.; All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41698-025-01055-9>.

Correspondence and requests for materials should be addressed to Guoliang Zheng, Siwei Pan or Can Hu.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025