Development and validation of a CT-based model for noninvasive prediction of T stage in gastric cancer: A multicenter study

Tao Jin, Dan Liu, Fubi Hu, Xiao Zhang, Hong-Kun Yin, Hui-Ling zhang, Kai Zhang, Zi-Xing Huang, Kun Yang

Submitted to: Journal of Medical Internet Research
on: January 31, 2024

Table of Contents

Original Manuscript **5**
Supplementary Files **23**
Figures **24**
Figure 1 ..... **25**
Figure 2 ..... **26**
Figure 3 ..... **27**
Figure 4 ..... **28**
Figure 5 ..... **29**
Figure 6 ..... **30**
Related publication(s) - for reviewers eyes onlies **31**
Related publication(s) - for reviewers eyes only 0 ..... **32**
Related publication(s) - for reviewers eyes only 0 ..... **32**

Development and validation of a CT-based model for noninvasive prediction of T stage in gastric cancer: A multicenter study

Tao Jin[1]; Dan Liu[1*] PhD; Fubi Hu[2] PhD; Xiao Zhang[3]; Hong-Kun Yin[4] PhD; Hui-Ling zhang[5] PhD; Kai Zhang[5] PhD; Zi-Xing Huang[5] PhD; Kun Yang[1*] PhD

[1]West China Hospital Chengdu CN
[2]The First Affiliated Hospital of Chengdu Medical College Chengdu CN
[3]People's Hospital of Leshan Leshan CN
[4]6. Institute of Advanced Research, Infervision Medical Technology Co., Ltd, Beijing, China. Beijing CN
[5]Institute of Advanced Research, Infervision Medical Technology Co., Ltd, Beijing, China. Beijing CN
[*]these authors contributed equally

Corresponding Author:
Kun Yang PhD
West China Hospital
Guo Xue street
Chengdu
CN

Abstract

Background: No studies have reported the use of deep learning radiomics to predict T staging in gastric cancer via integrating radiomics and deep learning.

Objective: To develop a computed tomography (CT)-based model for the automatic prediction of the T stage of gastric cancer (GC) via radiomics and deep learning.

Methods: A total of 771 GC patients from 3 centers were retrospectively enrolled and divided into training, validation, and testing cohorts. GC patients were classified into mild (stage T1 and 2), moderate (stage T3) and severe (stage T4) groups. Three predictive models based on the labelled CT images were constructed by using the radiomics features (Radiomics model), deep features (Deep learning model) and the combination of both (Hybrid model).

Results: The overall classification accuracy of the radiomics model was 64.3% in the internal testing dataset. The deep learning model and hybrid model showed better performance than the radiomics model, with overall classification accuracies of 75.7% (p = 0.037) and 81.4% (p = 0.001), respectively. On the subtasks of binary classification of tumor severity, the AUCs of the radiomics model, deep learning model and hybrid model were 0.875, 0.866 and 0.886 in the internal testing dataset and 0.820, 0.818 and 0.972 in the external testing dataset for differentiating mild (stage T1~2) from nonmild (stage T3~4) patients, while yielding 0.815, 0.892 and 0.894 in the internal testing dataset and 0.685, 0.808 and 0.897 in the external testing dataset for differentiating nonsevere (stage T1~3) from severe (stage T4) patients, respectively.

Conclusions: The hybrid model integrating radiomics features and deep features shows favorable performance in diagnosing the pathological stage of gastric cancer. Clinical Trial: None

(JMIR Preprints 31/01/2024:56851)
DOI: https://doi.org/10.2196/preprints.56851

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Development and validation of a CT-based model for noninvasive prediction of T stage in gastric cancer: A multicenter study

Tao Jin[1,2#], Dan Liu[3#], Fubi Hu[4], Xiao Zhang[5], Hongkun Yin[6], Huiling Zhang[6], Kai Zhang[6], Zixing Huang[3*], Kun Yang[1,2*]

1. Gastric Cancer Center, West China Hospital, Sichuan University, China;

2. Laboratory of Gastric Cancer, State Key Laboratory of Biotherapy/Collaborative Innovation Center of Biotherapy and Cancer Center, West China Hospital, Sichuan University, China;

3. Department of Radiology, West China Hospital, Sichuan University, Chengdu, China;

4. Department of Radiology, The First Affiliated Hospital of Chengdu Medical College, 278# Baoguang Road, Xindu District, Chengdu, Sichuan, China;

5. Department of Radiology, People's Hospital of Leshan, Leshan, China;

6. Institute of Advanced Research, Infervision Medical Technology Co., Ltd, Beijing, China.

Correspondence to

Prof. Zixing Huang, Department of Radiology, West China Hospital, Sichuan University, Chengdu, China, email: zixinghuang@hotmail.com

Prof. Kun Yang, Gastric Cancer Center, and Laboratory of Gastric Cancer, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, No. 37 Guo Xue Xiang Street, Chengdu 610041, Sichuan Province, China, e-mail: yangkun068@163.com

#: Co-first author

*: Corresponding author

Abstract

Objectives

To develop a computed tomography (CT)-based model for the automatic prediction of the T stage of gastric cancer (GC) via radiomics and deep learning.

Methods

A total of 771 GC patients from 3 centers were retrospectively enrolled and divided into training, validation, and testing cohorts. GC patients were classified into mild (stage T1 and 2), moderate (stage T3) and severe (stage T4) groups. Three predictive models based on the labelled CT images were constructed by using the radiomics features (Radiomics model), deep features (Deep learning model) and the combination of both (Hybrid model).

Results

The overall classification accuracy of the radiomics model was 64.3% in the internal testing dataset. The deep learning model and hybrid model showed better performance than the radiomics model, with overall classification accuracies of 75.7% (p = 0.037) and 81.4% (p = 0.001), respectively. On the subtasks of binary classification of tumor severity, the AUCs of the radiomics model, deep learning model and hybrid model were 0.875, 0.866 and 0.886 in the internal testing dataset and 0.820, 0.818 and 0.972 in the external testing dataset for differentiating mild (stage T1~2) from nonmild (stage T3~4) patients, while yielding 0.815, 0.892 and 0.894 in the internal testing dataset and 0.685, 0.808 and 0.897 in the external testing dataset for differentiating nonsevere (stage T1~3) from severe (stage T4) patients, respectively.

Conclusions

The hybrid model integrating radiomics features and deep features shows favorable performance in diagnosing the pathological stage of gastric cancer.

Highlights

• We develop a CT based model for the automatic prediction of the T stage of GC via radiomics and deep learning.

• The hybrid model had higher clinical benefit than the radiomics model and the deep learning model.

• The hybrid model shows favorable performance in diagnosing the pathological stage of gastric cancer.

Abbreviations

CT	Computed tomography
GC	Gastric cancer
ROC	Receiver operating characteristic
DCA	Decision curve analysis
UICC	Union for International Cancer Control
AJCC	American Joint Committee on Cancer
TNM	Tumor-node-metastasis
ROIs	Tumor regions of interest
ICC	Intraclass correlation coefficient
NLP	Natural language processing
AUC	Area under the curve
PPV	Positive predictive value
NPV	Negative predictive value
SVM	Support vector machine
ViT	Vision transformer
t-SNE	t-distributed stochastic neighbor embedding

Keywords

Gastric cancer; computed tomography; radiomics; T stage; deep learning

Introduction

Gastric cancer (GC) is one of the most prevalent cancers and ranks as the fourth most common cause of cancer-related death globally[1]. Although systemic therapy can increase the survival rate and improve quality of life, the prognosis remains poor due to diagnosis at an advanced stage[2]. Currently, gastric cancer staging is performed according to the Union for International Cancer Control (UICC) and American Joint Committee on Cancer (AJCC) tumor-node-metastasis (TNM) staging system[3]. This could be used to stratify cancer prognosis. As part of the TNM staging system, T staging based on the tumor depth of gastric cancer is crucial for developing treatment plans. For patients with clinical T1, endoscopic procedures were considered the first choice, while for patients with clinical T2, T3 or T4, surgery and perioperative chemotherapy were recommended[4]. However, the clinical and pathological diagnosis of tumor depths can sometimes vary[5]. In terms of preoperative T staging, computed tomography (CT) has a sensitivity and specificity of 80% to 90% for discriminating between early gastric and advanced gastric cancers[6],

and precise prediction of preoperative T staging by CT plays a vital role in the treatment of gastric cancer. A retrospective study[7] involving 244 gastric cancer patients showed that the performance of single-phase CT radiomics models was favorable in the differentiation between T2 and T3/4 stage tumors[7]. By extracting intricate information that is imperceptible to human eyes in medical imaging and transforming it into quantitative data, the radiomics and deep learning approaches had shown potential in improving the diagnostic capability of current imaging. In radiomics, medical images are transformed into mineable high-dimensional data, which can be used to quantify lesion heterogeneity that cannot be seen in the images. Furthermore, deep learning is the state-of-the-art machine learning approach that uses multiple processing layers and connections to learn complex relationships between input data and desired outputs from a large number of labelled examples, which could provide clinicians with decision support and improve diagnostic and treatment accuracy and efficiency[8, 9]. Previous studies have constructed a deep learning model based on the CT imaging radiomics signature to predict the number of lymph node metastases and survival for patients with resected GC[10, 11]. Our team has previously studied occult peritoneal metastasis of gastric cancer by using imaging radiomics and deep learning[12, 13]. In addition, the feasibility and promising performance of machine learning approaches in assessing T staging in lung cancer has been demonstrated[14]. However, few studies have reported the combination of deep learning and radiomics in predicting T staging in gastric cancer. To improve the diagnostic accuracy of T staging to better develop treatment strategies, the aim of this research was to focus on preoperatively discriminating pathologic T staging by deep learning and combining single-phase CT radiomics models with deep learning parameters, which can potentially be helpful and assist in guiding clinicians in regard to personalized medicine.

Materials and methods

Patient enrollment and eligibility criteria

The inclusion criteria were as follows:

(1) Primary gastric adenocarcinoma diagnosed by endoscopy-biopsy pathology; (2) venous images of the whole abdomen (with a slice thickness of 2 mm) were obtained preoperatively, and later laparoscopy or surgery was performed within 2 weeks; (3) no typical peritoneal metastasis (PM) findings, such as omental nodules or omental cake, extensive ascites, or irregular thickening with high peritoneal enhancement, on CT; and (4) no indications of distant metastasis or other tumors.

The exclusion criteria were as follows:

(1) Previous abdominal surgery; (2) previous abdominal malignancies or inflammatory diseases; (3) insufficient distention of the stomach; (4) poor imaging quality due to artifacts; and (5) indiscernible primary GC tumor on CT images.

Ethical approval

This research was granted approval by the Ethics Committee of West China Hospital, Sichuan University (2019-1158). Given the retrospective nature of the study design and the anonymous analysis of all data, the necessity to obtain informed consent was waived.

Segmentation and preprocessing of tumor regions

Given the T-stage in GC, we divided the images into three grading for this study: mild group (stage T1 and stage T2), moderate group (stage T3) and severe group (stage T4a and stage T4b). The open-source software ITK-SNAP (version 3.8.0, www.itksnap.org) was utilized for image segmentation. The outline for the tumor regions of interest (ROIs) was manually drawn by a junior radiologist with 6 years of experience in reading abdominal CT images, and meanwhile was revised by a senior radiologist with 17 years of experience. When delineating the tumor ROI, the radiologist referred to the results of gastroscopy to determine the location of the tumor. Besides, 40 CT images of GC lesions were randomly selected and delineated again by the junior radiologist to assess the test–retest reliability. Blinded to segmentations delineated by the junior radiologist, these 40 CT images of GC lesions were delineated again by the senior radiologist to assess intraclass correlation coefficient (ICC). Features with an inter-observer ICC > 0.85 were retained, and the ICC values of the selected radiomics were presented in Table S2 in the Supplementary Material

Extraction and selection of the radiomics features

The radiomics features were automatically extracted from the ROI on the maximum cross-section layer of the NCCT images by using the PyRadiomics package (version 3.0). A set of filters (Wavelet, Square, SquareRoot, Logarithm) were applied to highlight particular image properties[15]. Finally, a total of 1183 radiomics features were extracted from each of the patients. The principal component analysis (PCA) method was used for dimension reduction of the features and to reduce computation complexity and prevent overfitting. PCA was an unsupervised method that transformed complex and high-dimension original radiomics data into a dimensionally reduced set of uncorrelated features named principal components. The calculation of the principal components was carried out by singular value decomposition of the standardized radiomics features, and the top 15 components that explained most of the variance in the training dataset were finally selected for model development.

CT Image Acquisition

The details of the CT protocol are presented in **Section 2** and **Table S1** in the **Supplementary Material**. Segmentation and preprocessing of tumor regions, Extraction and selection of the radiomics features were shown **in supplementary material.** To reduce the radiologists' workload of depicting the ROIs, this study attempted to apply partial markers to automatically generate 3D patches. The radiologists manually delineated the ROIs at the initial image layer, the final image layer and the maximum cross-section layer with the largest boundaries of the tumor. According to ROI information of the three key layers, 3D cubes including the whole tumor and the surrounding structures were automatically generated (**Figure 1**).

Development of the Radiomics model

Based on the selected radiomics features, a radiomics model was constructed by using the Support Vector Machine (SVM) classifier. As a supervised learning method which was very effective in linear or nonlinear classification tasks, the SVM classifier was widely used in radiomics analysis. The development and validation of the radiomics model was performed on the InferScholar (version 3.5) platform, and the parameters were set as follows: C = 1.0, kernel = 'Sigmoid', gamma = 'auto', Tol = 0.001, class_weight = 'balanced', and other parameters were set as default.

Development of the deep learning models

Vision transformer-based deep learning model

Vision transformer (ViT) is derived by the pure transformer, which was first proposed for natural language processing (NLP) by Vaswani and his team[16]. In this study, we applied Vision Transformer to grade GC severity based on CT images. The current implementation is inspired by the work of Dosovitskiy and his team, which applied ViT to the ImageNet dataset and showed excellent performance in image classification compared with state-of-the-art Convolutional Neural Networks (CNNs)[17]. Due to the advantage of checking the overall relationship by modelling the dependence between different parts of an image, ViT showed strength in global image classification, which was very suitable for our case[18]. The conceptual architecture of the ViT used in our study is illustrated in **Figure 2**. Briefly, the 3D patches containing GC regions were cropped from CT images. After resizing, the 3D patch was split into small patches, and these small patches were converted to a sequence of patch embeddings by flattening and linear projection. Then, the patch embeddings together with the positional embedding were fed into the transformer encoder to obtain the final representation. Consequently, the learnable features in the input images were fed into the classifier head to identify GC severity. In this study, the embedding dimension of the ViT models was fixed to 768, both numbers of the encoder layers and the attention heads were 12, and the dimensionality of the expanded representation in predicting head was 3072.

Deep learning-based Hybrid model

To integrate radiomics features and deep learning features, the ViT architecture-based hybrid model was proposed in this study, namely, the ViT-Radiomics model[19, 20], which combined both the ViT model and radiomics models for the grading of GC severity. In brief, the deep features extracted by ViT were first transformed into a 1280-bit vector, while the selected radiomics features were also transformed into a 512-bit vector by using a vectorization transform approach. Then, the vectorized deep features and radiomics features were further concatenated into a 1792-bit vector and used to predict the severity of gastric cancer (mild, moderate or severe).

Data augmentation

To meet the high demand for large amounts of training data in ViT-based deep learning models, several sophisticated data augmentation techniques were applied in the training dataset to avoid the overfitting problem and promote the prediction performance during the model training process. Data augmentation techniques in this study mainly referred to a series of geometric transformations including scaling (0.9 times or 1.1 times the image size), translation (up and down, left and right, front and back), mirroring, rotation and flip (horizontal and vertical) on the

generated 3D patches of the original images. After data augmentation, the sample size in the training set increased to 5 times that of the original training dataset.

Training of the deep learning models

The deep learning models (deep learning model and hybrid model) were trained by maximizing the identification performance (i.e., accuracy) for grading GC severity and minimizing the categorical cross-entropy loss. This study used stochastic gradient descent (SGD) with a momentum of 0.95, a weight decay of 0.0001 and an initial learning rate of 0.001 to optimize the model parameters. The number of modelling epochs was set to 150, and the mini-batch size was set to 32. The model development process was implemented using four GeForce RTX 2080ti GPUs on Ubuntu 18.04.4 LTS, Python 3.7.11 and PyTorch 1.7.1. No samples overlapped at the patient level in the development and independent datasets.

Evaluation of model performance

The trinary diagnostic capability of the predictive models was assessed by overall accuracy and Cohen's kappa coefficient[21], and the chi-square test was used for the comparison of accuracies between different predictive models. Because of the imbalanced distribution of the three GC severity categories, the per-class F1-score and weighted average F1-score were also calculated[22].

In addition, two clinically important subtasks were also evaluated: (i) binary classification of mild (stage T1~2) GCs and nonmild GCs (stage T3~4) and (ii) binary classification of nonsevere GCs (stage T1~3) and severe GCs (stage T4). The discriminative efficacy of binary classification was evaluated by receiver operating characteristic (ROC) analysis with respect to the area under the curve (AUC). The sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were also calculated under the optimal threshold according to the maximum Youden index[23]. Furthermore, decision curve analysis (DCA) was used to evaluate the clinical utility of the radiomics model, deep learning model and hybrid model for binary classification of mild/nonmild and nonsevere/severe GCs by comparing the net benefit across a range of threshold probabilities in the training and validation datasets[24].

Statistical analysis

Statistical analysis was performed with R project (v. 3.3.1), SPSS software (version 23.0) and MedCalc software (version 20.0). The differences in continuous variables with normal or non-normal distributions were evaluated through Student's t test or the Mann–Whitney U test, respectively. Categorical variables were compared with the

chi-square test. The difference between the two AUCs of different models was assessed by using DeLong's test[25]. The decision curve was plotted using the "rmda" package in R language. A two-sided p value less than 0.05 was considered statistically significant.

Results

Patient characteristics

Ultimately, a total of 706 GC individuals were enrolled in this retrospective study, and the patients were randomly divided into a development dataset (n=566) and an independent testing dataset (n=140) at a ratio of 4:1. Sixty-five GC patients from two external centers were also recruited as the external testing dataset (**Figure S1 in the Supplementary Material**).

Clinical characteristics in the development, internal and external sets are listed in **Table 1**. Age, peritoneum metastasis, and CA199 differed significantly among the three datasets. No significant difference was observed among the three datasets in terms of sex, size, site, Boorman type, grade, Lauren type, adjacent tissue invasion, pathological T stage, TNM, CA724, CA125 and CEA.

Performance evaluation of the trinary classification of GC severity

The three-class confusion matrixes of the predictive models in the development, internal testing and external testing datasets are shown in Figure 3. The overall accuracies of the radiomics model, deep learning model and hybrid model were 69.8%, 75.1% and 83.7% in the development dataset, 64.3%, 75.7% and 81.4% in the internal testing dataset, and 53.8%, 70.8% and 81.5% in the external testing dataset, respectively. The accuracy of deep learning model and hybrid model were statistically different from the accuracy of radiomics model in the development dataset t (deep learning model *vs* radiomics model, $p = 0.046$; hybrid model *vs* radiomics model, $p < 0.001$), internal testing dataset (deep learning model *vs* radiomics model, $p = 0.037$; hybrid model *vs* radiomics model, $p = 0.001$) and external testing dataset (deep learning model *vs* radiomics model, $p = 0.047$; hybrid model *vs* radiomics model, $p < 0.001$). The overall accuracy of the hybrid model was statistically different from that of the deep learning model in the development dataset ($p < 0.001$); however, no significant difference was found in the internal testing dataset ($p = 0.245$) or external testing dataset ($p = 0.150$). Similarly, the predicted results of both the deep learning model and hybrid model showed good

agreement with the pathologically confirmed ground truth in the development dataset (deep learning model kappa = 0.660, hybrid model kappa = 0.767), internal testing dataset (deep learning model kappa = 0.665, hybrid model kappa = 0.735) and external testing dataset (hybrid model kappa = 0.689), while the agreement for the radiomics model was only moderate (development kappa = 0.579, internal testing kappa = 0.518, external testing kappa = 0.400). These results, as well as the per-class F1-score and weighted average F1-score, are presented in Table 2.

Evaluation on two binary classification subtasks

As depicted in Figure 4, the binary classification performance of the predictive models for mild vs nonmild and nonsevere vs severe GCs was evaluated in both internal and external testing datasets. For the binary classification of mild (stage T1~2) GCs and nonmild GCs (stage T3~4), the radiomics model, deep learning model and hybrid model showed similar performance, with AUCs of 0.875 (95% CI, 0.809-0.925), 0.866 (95% CI, 0.799-0.918) and 0.886 (95% CI, 0.822-0.934) in the internal testing dataset, respectively (Delong's test, all p values > 0.05). Meanwhile, The AUC values of the hybrid model, radiomics model, and deep learning in the external testing dataset were 0.972 (95% CI, 0.897-0.997), 0.820 (95% CI, 0.704-0.904; p = 0.002), and 0.818 (95% CI, 0.703-0.903; p = 0.025), respectively. For the binary classification of nonsevere GCs (stage T1~3) and severe GCs (stage T4), the AUC values of deep learning model and hybrid model were 0.892 (95% CI, 0.829-0.938) and 0.894 (95% CI, 0.831-0.940), respectively. And the AUC value of the radiomics model was 0.815(95% CI, 0.740-0.875) (p = 0.024 radiomics model vs deep learning model and p = 0.025 radiomics model vs hybrid model). In addition, The area under the receiver operating characteristic curve (AUC) values for the deep learning model, hybrid model, and radiomics model were 0.808 (95% CI, 0.691-0.895), 0.897 (95% CI, 0.797-0.959), and 0.685 (95% CI, 0.558-0.795), respectively (p = 0.032 radiomics model vs deep learning model and p = 0.002 radiomics model vs hybrid model) in the external testing dataset, while no significant difference was found between them (p = 0.098). The detailed performance of these models in the internal and external testing datasets is listed in Table 3 and Table 4, respectively.

Clinical utility analysis

The decision curve analysis for the two binary classification subtasks of the different predictive models in the internal and external datasets are presented in Figure 5. For the binary classification of mild (stage T1~2) GCs and nonmild GCs (stage T3~4), the hybrid model had a slightly higher overall net benefit than the radiomics model and

the deep learning model across the majority range of reasonable threshold probabilities in the internal testing dataset. For the binary classification of nonsevere GCs (stage T1~3) and severe GCs (stage T4), the deep learning model and hybrid model showed obviously higher net benefit than the radiomics model across the majority range of reasonable threshold probabilities in the internal testing dataset. A similar tendency was also observed in the external dataset, as shown in Figure 5C and 5D. These results were consistent with the ROC analysis.

Visualization of the internal features learned by the neural networks

The internal features learned by the hybrid model were examined in the internal testing dataset by using the t-distributed stochastic neighbor embedding (t-SNE) method[26]. Each point represented an input NCCT image of a patient projected from the high-dimensional vector of the neural network's last hidden layer into two dimensions. The mild group (blue point cloud) and severe group (red point cloud) showed clear clustering patterns and were split across the moderate group (green point cloud).

Discussion

Principal Findings and Comparison With Prior Work

A preoperative evaluation of tumor invasion depth determines the individual treatment plan for gastric cancer. CT has been the first choice preoperatively for GC evaluation and is important for clinical practice. However, the routine preoperative determination method is still not accurate enough, especially when radiologists look at it with the naked eye. This study makes full use of the CT radiomics signature and the signature extracted by deep learning to make a better prediction model to overcome the heterogeneity caused by the naked eye. To our knowledge, this is the first article that demonstrates a CT based model utilizing radiomics and deep learning techniques for the automated prediction of the T stage of gastric cancer. Our hybrid model demonstrated superior clinical utility compared to both the radiomics and deep learning models individually. Moreover, the hybrid model exhibited promising diagnostic performance in determining the pathological stage of gastric cancer.

Over the past decade, radiomics research has gained great attention, and a growing number of studies have been published on radiomics in oncology[27]. This could be at least an important complement

to subjective evaluation by radiologists. Previous studies have illustrated that CT-based radiomics used alone or by deep learning or other statistical methods could predict the number of lymph node metastases[11, 28], chemotherapy treatment response, and survival[10, 29] for gastric cancer patients. Radiomics are mathematically defined descriptors, while deep learning features are less intuitive due to the complexity of deep neural networks. The predictive values of the two are different and stackable[30]. Sun et al[31] conducted a pilot study based on radiomics analysis and found that the model had a predictive AUC of 0.852 in rectal cancer for the diagnostic T stage. Yang and his colleagues[32] demonstrated that CT radiomics signatures exhibited favorable predictive performance for esophageal carcinoma T stage with an AUC of 0.86. Furthermore, our study also demonstrated that the hybrid model (83.7%) and deep learning model (75.1%) had higher accuracy than the radiomics model (69.8%) in evaluating three grade (mild, moderate and severe group)-based T staging in the development dataset and internal and external datasets. In addition, the hybrid model and deep learning model showed better performance than the radiomics model, with good agreement between the model's prediction and ground truth. Thus, using the hybrid model and deep learning model to predict pathology is eligible and reliable. Regarding the overall accuracy of the hybrid model and the deep learning model, the hybrid model had higher accuracy than the deep learning model in the development dataset ($p < 0.001$); however, no significant difference was found in the internal testing dataset ($p = 0.245$) or external testing dataset ($p = 0.150$), which was possibly due to the limited sample size in the internal and external datasets. By using the hybrid model, the diagnostic accuracy of T stage could be improved, which can help clinicians to make more precise treatment for GC patients to avoid delays in treatment or mistreatment. Furthermore, deep learning can reduce the workload of doctors and improve work efficiency.

Additionally, we aimed to better explore the predictive models constructed by the deep learning model and hybrid model. The tumor invasion depth divided into mild vs nonmild and nonsevere vs severe GCs was evaluated separately. The radiomics model, deep learning model and hybrid model

all had good performance in differentiating mild and nonmild GC, as well as severe and nonsevere GC, in both the internal and external datasets. Meanwhile, the hybrid model (AUC, 0.972; 95% CI, 0.897-0.997) outperformed both the radiomics model (AUC, 0.820; 95% CI, 0.704-0.904; p = 0.002) and deep learning (AUC, 0.818; 95% CI, 0.703-0.903; p =0.025) in the external testing dataset. Based on the hybrid model, for patients with T1 or T2, endoscopic resection or surgery was considered in combination with other examinations, while in patients who had T3 or T4 lesions, adjuvant therapy was recommended. One retrospective study involving 572 gastric cancer patients diagnosed at T3 or T4 stage pathologically showed that a radiomics model based on CT images by deep learning is effective at discriminating serosa invasion in gastric cancer[33], which is consistent with our results. GC patients with T1~T3 disease have a lower risk of peritoneal metastasis than those with T4 disease, while in patients with T4 disease, the hybrid model could assist clinicians in improving the detection accuracy of peritoneal metastasis, especially for regions where staging laparoscopy was not applied widely. Furthermore, the results of DCA showed that the deep learning model and hybrid model showed obviously higher net benefit than the radiomics model across the majority range of reasonable threshold probabilities, which means that the deep learning model and hybrid model have certain clinical application value. T-SNE analysis also showed that different GC risk stratifications have their own distinct clusters. Based on the above results, the deep learning model and hybrid model have good performance in distinguishing T staging. Additionally, the hybrid model might have better performance in distinguishing T staging than the deep learning model and radiomics model.

Clinical implication

The hybrid model has the capability to accurately predict the T stage of patient outcomes directly from conventional CT images through automated processes. Furthermore, the hybrid model shows promise in aiding clinicians by offering a more dependable and accurate preoperative T staging diagnosis which can influence real-world clinical decision-making. For example, by providing

accurate T-staging, neoadjuvant therapy can be arranged for patients with advanced gastric cancer. By pretreatment evaluation of T stage, it could help clinicians to choose the correct treatment method. Following surgical procedures, we offer individualized treatment plans for patients through precise pathology diagnosis and selection of suitable adjuvant therapies.

Ethical implication

A prior investigation identified three key domains of ethical concerns related to the utilization of artificial intelligence: algorithms, data, and practices. These domains highlight the importance of obtaining informed consent and establishing data use agreements between data providers and third-party data aggregators. Furthermore, ensuring the quality of data utilized in AI algorithms, especially in the context of enhancing patient treatment decisions, is a significant area of concern. Furthermore, the integration of AI technology should not be viewed as a substitute for the collaborative decision-making process that is integral to optimal patient care. It is important to juxtapose these findings with the standard clinical practices currently employed for T staging in gastric cancer.

Limitations

There are several limitations in the study. First, it was a retrospective study that had selection bias, and limited sample size images still caused bias in model construction. Thus, for further validation and to enhance the generalizability of our results, a prospective design with a large sample size is necessary for further investigation. Second, our method requires manual segmentation of the tumor, which is time consuming. Automated segmentation methods, as well as fully automated models, could be valuable in the future. Additionally, when tumor ROIs were outlined manually, there was some heterogeneity in the radiologists' experience, and the radiologists needed to be trained uniformly. In addition, our study population included only Asians, further research from multiple centers is necessary to confirm the validity of this model prior to its potential clinical implementation. Furthermore, artificial intelligence lacks the capacity to participate in complex dialogues with patients, as well as the ability to establish the necessary trust and empathy essential

for fostering the therapeutic alliance crucial to the patient-physician relationship and favorable treatment outcomes.

Conclusion

The hybrid model might be able to distinguish T staging more effectively than deep learning models or radiomics models, which could be applied in clinical practice.

Conflict of interest

The authors have declared that no competing interest exists.

Data availability

The data and code are available by contacting the corresponding authors.

Funding

This study has received funding by (1) 1. 3. 5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYJC21006); (2) Foundation of Science & Technology Department of Sichuan Province (2023YFS0060)

Informed Consent

Written informed consent was waived by the Institutional Review Board.

Ethical Approval

Institutional Review Board approval was obtained.

Reference

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, 2021. **71**(3): p. 209-249.
2. Ajani, J.A., et al., *Gastric Cancer, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology.* J Natl Compr Canc Netw, 2022. **20**(2): p. 167-192.
3. Edge, S.B. and C.C. Compton, *The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM.* Ann Surg Oncol, 2010. **17**(6): p. 1471-4.
4. Joshi, S.S. and B.D. Badgwell, *Current treatment and recent progress in gastric cancer.* CA Cancer J Clin, 2021. **71**(3): p. 264-279.
5. Nanishi, K., et al., *Diagnostic accuracy of the gastric cancer T-category with respect to tumor localization.* Langenbecks Arch Surg, 2020. **405**(6): p. 787-796.
6. Ahn, H.S., et al., *Diagnostic accuracy of T and N stages with endoscopy, stomach protocol CT, and endoscopic ultrasonography in early gastric cancer.* J Surg Oncol, 2009. **99**(1): p. 20-7.
7. Wang, Y., et al., *Prediction of the Depth of Tumor Invasion in Gastric Cancer: Potential*

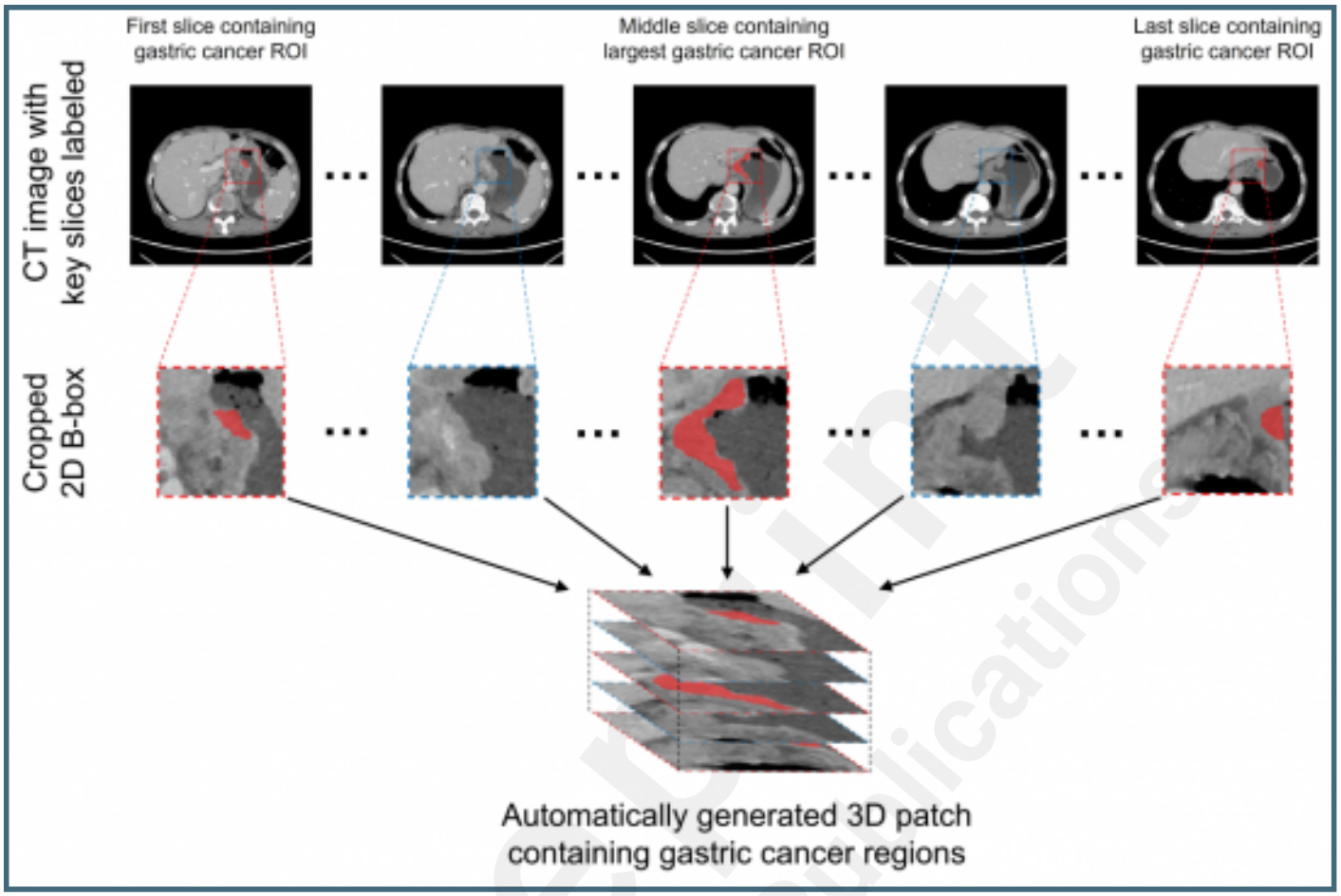
Role of CT Radiomics. Acad Radiol, 2020. **27**(8): p. 1077-1084.
8. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* Nature, 2015. **521**(7553): p. 436-44.
9. Chan, H.P., et al., *Deep Learning in Medical Image Analysis.* Adv Exp Med Biol, 2020. **1213**: p. 3-21.
10. Jiang, Y., et al., *Development and Validation of a Deep Learning CT Signature to Predict Survival and Chemotherapy Benefit in Gastric Cancer: A Multicenter, Retrospective Study.* Ann Surg, 2021. **274**(6): p. e1153-e1161.
11. 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, 2020. **31**(7): p. 912-920.
12. Liu, D., et al., *A Bounding Box-Based Radiomics Model for Detecting Occult Peritoneal Metastasis in Advanced Gastric Cancer: A Multicenter Study.* Front Oncol, 2021. **11**: p. 777760.
13. Huang, Z., et al., *Deep Convolutional Neural Network Based on Computed Tomography Images for the Preoperative Diagnosis of Occult Peritoneal Metastasis in Advanced Gastric Cancer.* Front Oncol, 2020. **10**: p. 601869.
14. Kirienko, M., et al., *Convolutional Neural Networks Promising in Lung Cancer T-Parameter Assessment on Baseline FDG-PET/CT.* Contrast Media Mol Imaging, 2018. **2018**: p. 1382309.
15. van Griethuysen, J.J.M., et al., *Computational Radiomics System to Decode the Radiographic Phenotype.* Cancer Res, 2017. **77**(21): p. e104-e107.
16. Vaswani, A., et al., *Attention is all you need.* NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: p. 6000-6010.
17. Alexey Dosovitskiy, L.B., Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* Proceedings of the Ninth International Conference on Learning Representations (ICLR), 2021: p. 1–12.
18. Kim, S., J. Nam, and B.C. Ko, *Facial Expression Recognition Based on Squeeze Vision Transformer.* Sensors (Basel), 2022. **22**(10).
19. Han, K., et al., *A Survey on Vision Transformer.* IEEE Trans Pattern Anal Mach Intell, 2022. **Pp**.
20. Leamons, R., Cheng, H., Al Shami, A. , *Vision Transformers for Medical Images Classifications. In: Arai, K. (eds) Intelligent Systems and Applications.* IntelliSys 2023. **Lecture Notes in Networks and Systems, vol 544. Springer, Cham. https://doi.org/10.1007/978-3-031-16075-2_22**.
21. McHugh, M.L., *Interrater reliability: the kappa statistic.* Biochem Med (Zagreb), 2012. **22**(3): p. 276-82.
22. Zhao, W., et al., *3D Deep Learning from CT Scans Predicts Tumor Invasiveness of Subcentimeter Pulmonary Adenocarcinomas.* Cancer Res, 2018. **78**(24): p. 6881-6889.
23. Yin, J., H. Samawi, and L. Tian, *Joint inference about the AUC and Youden index for paired biomarkers.* Stat Med, 2022. **41**(1): p. 37-64.
24. Vickers, A.J. and E.B. Elkin, *Decision curve analysis: a novel method for evaluating prediction models.* Med Decis Making, 2006. **26**(6): p. 565-74.
25. DeLong, E.R., D.M. DeLong, and D.L. Clarke-Pearson, *Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.* Biometrics, 1988. **44**(3): p. 837-45.
26. Laurens, V.D.M. and G. Hinton, *Visualizing Data using t-SNE.* Journal of Machine Learning Research, 2008. **9**(2605): p. 2579-2605.
27. Chen, Q., et al., *Radiomics in precision medicine for gastric cancer: opportunities and challenges.* Eur Radiol, 2022.
28. Jin, C., et al., *Deep learning analysis of the primary tumour and the prediction of lymph node metastases in gastric cancer.* Br J Surg, 2021. **108**(5): p. 542-549.
29. Zhang, L., et al., *A deep learning risk prediction model for overall survival in patients with gastric cancer: A multicenter study.* Radiother Oncol, 2020. **150**: p. 73-80.
30. Hao, D., et al., *Identifying Prognostic Markers From Clinical, Radiomics, and Deep Learning Imaging Features for Gastric Cancer Survival Prediction.* Front Oncol, 2021. **11**: p. 725889.
31. Sun, Y., et al., *Radiomic features of pretreatment MRI could identify T stage in patients*

with rectal cancer: Preliminary findings. J Magn Reson Imaging, 2018.
32. Yang, M., et al., *Computed Tomography-Based Radiomics in Predicting T Stage and Length of Esophageal Squamous Cell Carcinoma.* Front Oncol, 2021. **11**: p. 722961.
33. Sun, R.J., et al., *CT-based deep learning radiomics analysis for evaluation of serosa invasion in advanced gastric cancer.* Eur J Radiol, 2020. **132**: p. 109277.

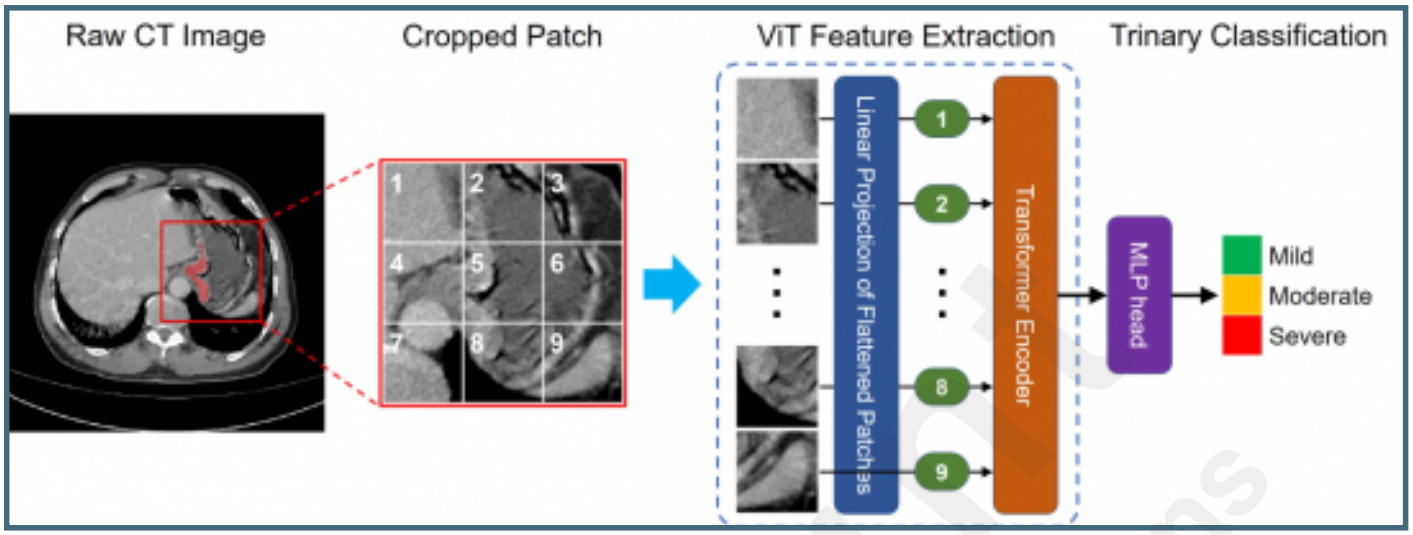
Supplementary Files

Figures

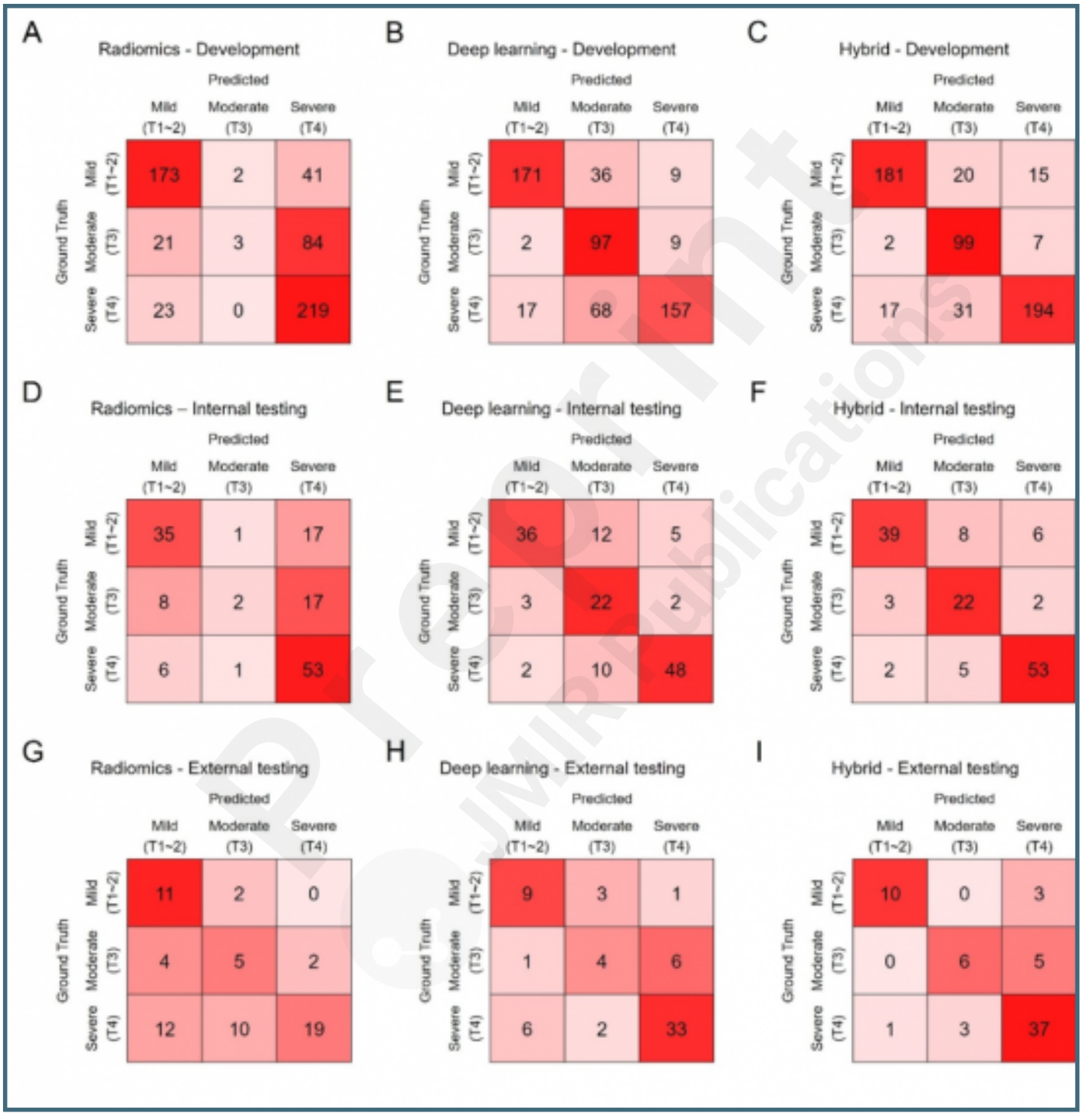
Example of the key slice-based automatic generation of a 3D cube including the whole GC region.



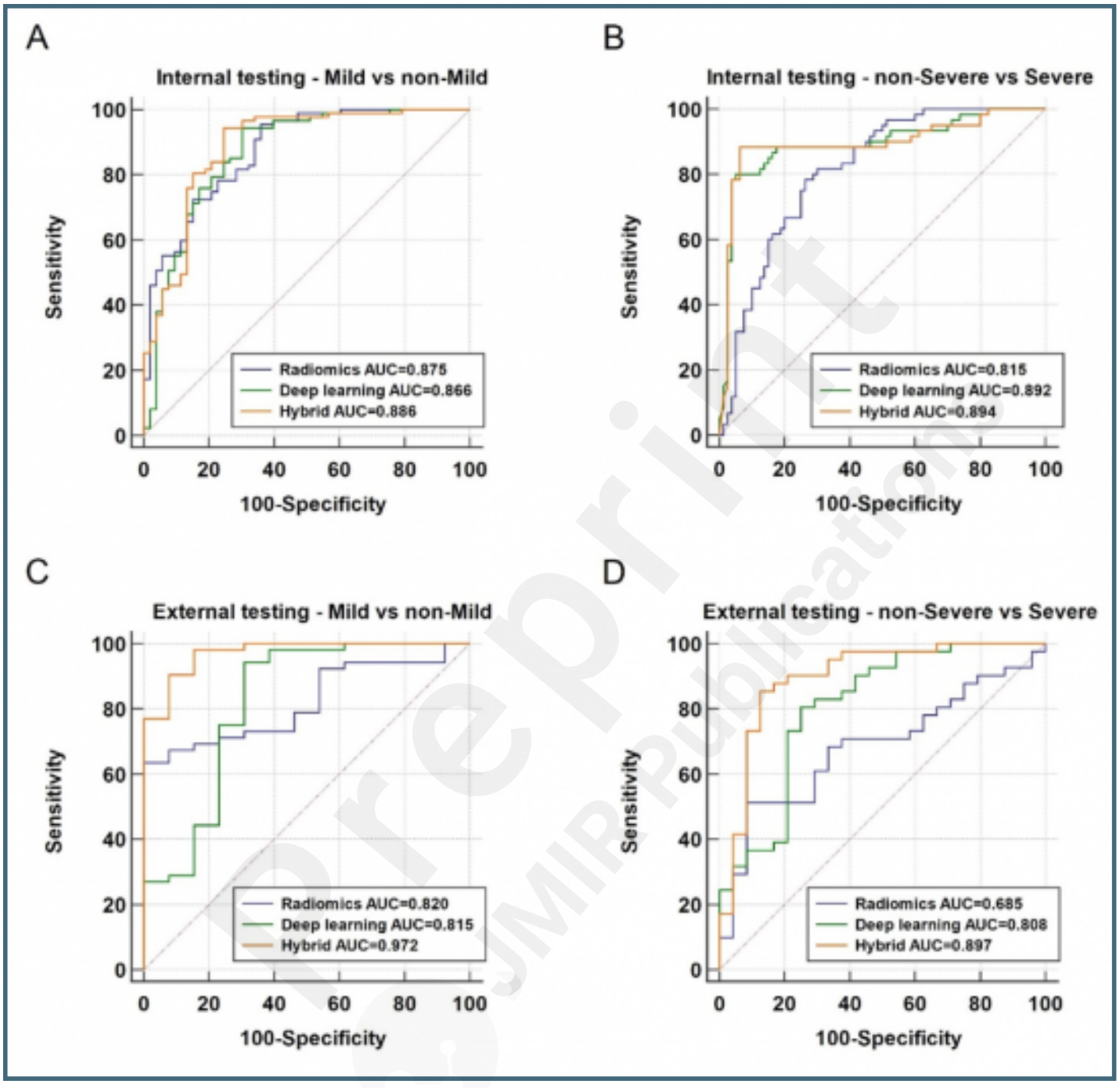
Overview of the Vision Transformer for the pathological severity classification of gastric cancer on NCCT images.



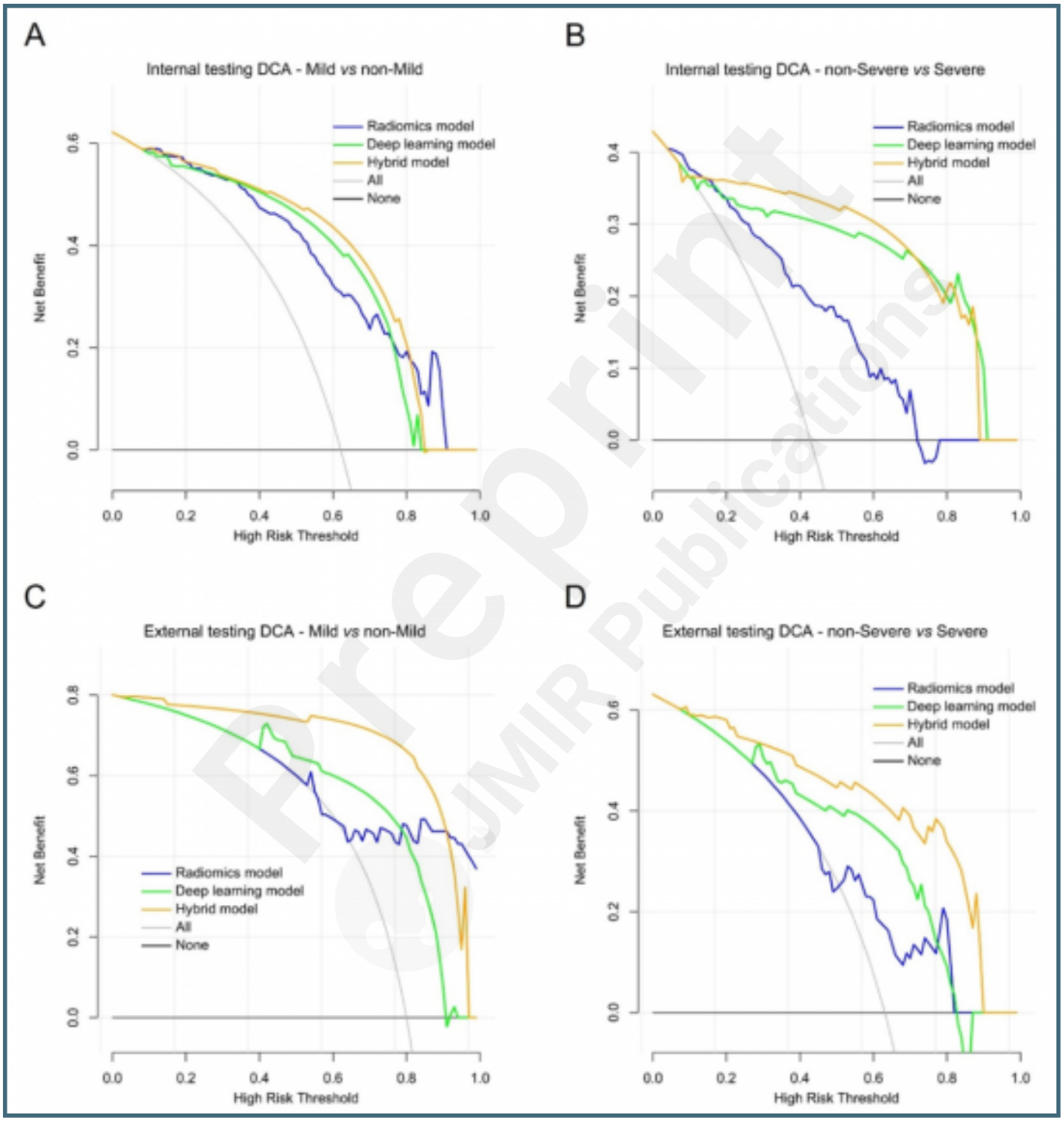
Trinary classification performance of the predictive models. A-C: The three-class confusion matrix of the radiomics model, deep learning model and hybrid model in the development dataset. D-F: The three-class confusion matrix of the radiomics model, deep learning model and hybrid model in the internal testing dataset. G-I: The three-class confusion matrix of the radiomics model, deep learning model and hybrid model in the external testing dataset.



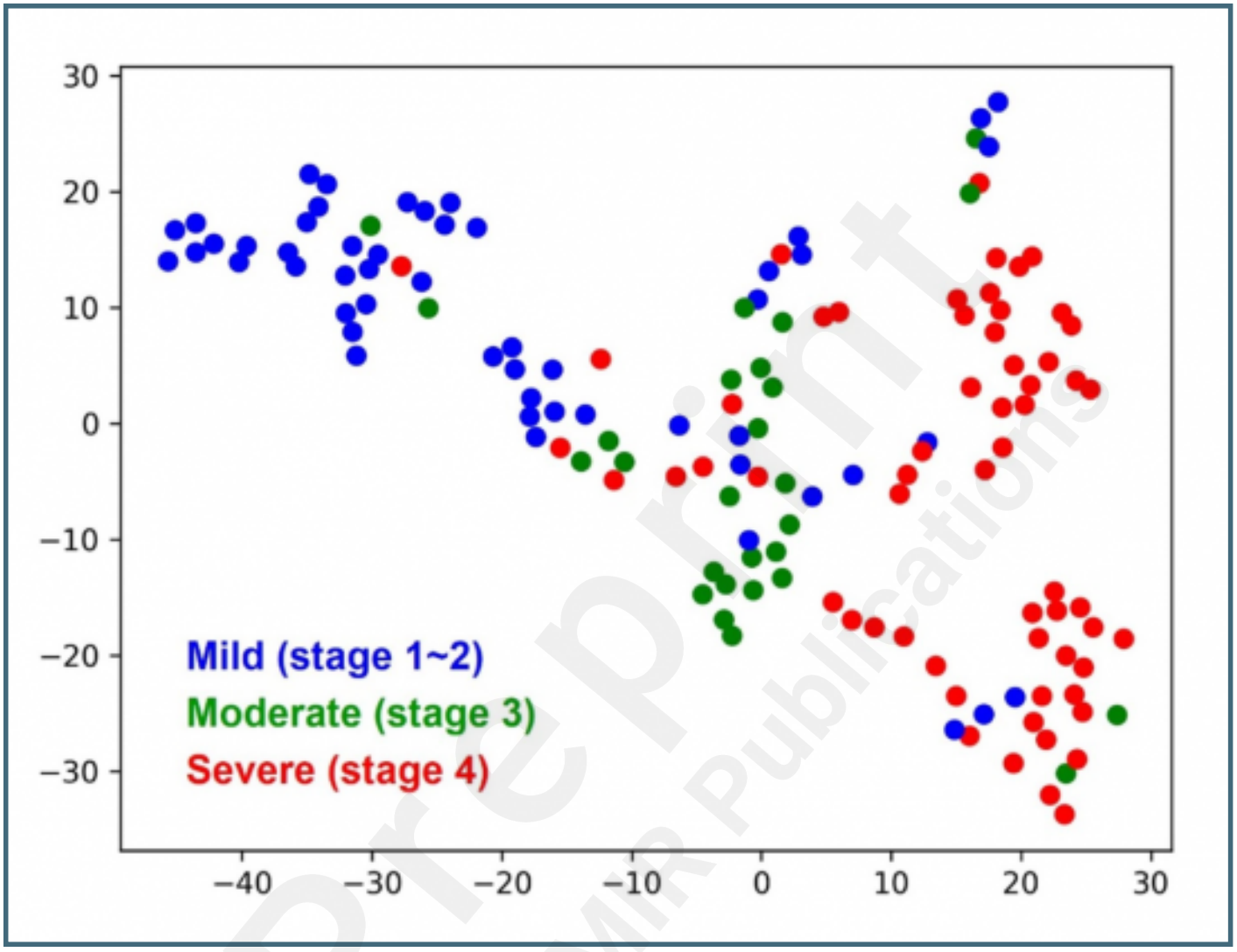
Comparison between the predictive models on two binary subtasks.



Decision curve analysis for the predictive models in the internal testing dataset on (A) binary classification subtask 1 (stage T1~2 vs stage T3~4) and (B) binary classification subtask 2 (stage T1~3 vs stage T 4) and for the predictive models on binary classification subtask 1 (C) and 2 (D) in the external testing dataset. The net benefit is depicted on the y-axis. The gray line and black line represent situations in which all patients and no patients underwent biopsy/surgery, respectively.



Visualization of the internal representations of the hybrid model for three severity classes by t-SNE. Colored point clouds represented the different severity categories, showing how the neural network clustered the diseases.



Related publication(s) - for reviewers eyes onlies

Untitled.
URL: http://asset.jmir.pub/assets/8ae1e6f134254c184785e7a14b141ad5.pdf

Table1-4.
URL: http://asset.jmir.pub/assets/9de088b22a29851add480419f6973606.pdf