Dovepress

ORIGINAL RESEARCH

Construction of a 2.5D Deep Learning Model for Predicting Early Postoperative Recurrence of Hepatocellular Carcinoma Using Multi-View and Multi-Phase CT Images

Yu-Bo Zhang[1,2], Zhi-Qiang Chen[1,3], Yang Bu [ID][2], Peng Lei [ID][2], Wei Yang[4], Wei Zhang[2]

[1]School of Clinical Medicine, Ningxia Medical University, Yinchuan, 750004, People's Republic of China; [2]Department of Hepatobiliary Surgery, General Hospital of Ningxia Medical University, Yinchuan, 750004, People's Republic of China; [3]Department of Hepatobiliary Surgery, People's Hospital of Ningxia Hui Autonomous Region, Yinchuan, 750002, People's Republic of China; [4]Department of Radiology, General Hospital of Ningxia Medical University, Yinchuan, 750004, People's Republic of China

Correspondence: Peng Lei, Department of Hepatobiliary Surgery, General Hospital of Ningxia Medical University, No. 804 Shengli Nan Street, Yinchuan, 750004, People's Republic of China, Email leipengnx@126.com; Yang Bu, Department of Hepatobiliary Surgery, General Hospital of Ningxia Medical University, No. 804 Shengli Nan Street, Yinchuan, 750004, People's Republic of China, Email buyang1976@163.com

Purpose: To construct a 2.5-dimensional (2.5D) CT radiomics-based deep learning (DL) model to predict early postoperative recurrence of hepatocellular carcinoma (HCC).

Patients and Methods: We retrospectively analyzed the data of patients who underwent HCC resection at 2 centers. The 232 patients from center 1 were randomly divided into the training (162 patients) and internal validation cohorts (70 patients); 91 patients from center 2 formed the external validation cohort. We developed a 2.5D DL model based on a central 2D image with the maximum tumor cross-section and adjacent slices. Multiple views (transverse, sagittal, and coronal) and phases (arterial, plain, and portal) were incorporated. Multi-instance learning techniques were applied to the extracted data; the resulting comprehensive feature set was modeled using Logistic Regression, RandomForest, ExtraTrees, XGBoost, and LightGBM, with 5-fold cross validation and hyper-parameter optimization with Grid-search. Receiver operating characteristic curves, calibration curves, DeLong test, and decision curve analysis were used to evaluate model performance.

Results: The 2.5D DL model performed well in the training (AUC: 0.920), internal validation (AUC: 0.825), and external validation cohorts (AUC: 0.795). The 3D DL model performed well in the training cohort and poorly in the internal and external validation cohorts (AUCs: 0.751, 0.666, and 0.567, respectively), indicating overfitting. The combined model (2.5D DL+clinical) performed well in all cohorts (AUCs: 0.921, 0.835, 0.804). The Hosmer-Lemeshow test, DeLong test, and decision curve analysis confirmed the superiority of the combined model over the other signatures.

Conclusion: The combined model integrating 2.5D DL and clinical features accurately predicts early postoperative HCC recurrence.

Keywords: hepatocellular carcinoma, liver resection, deep learning, computed tomography, recurrence

Introduction

Liver resection and liver transplantation are preferred surgical treatments for early hepatocellular carcinoma (HCC), with comparable outcomes;[1] however, these treatments are associated with a high postoperative recurrence rate of approximately 50%–70%, which significantly impacts the postoperative survival rate and contributes to a poor prognosis.[2,3] Studies have shown that early recurrence, occurring within 2 years after surgery, accounts for 61.4%–83.3% of all postoperative recurrences,[4,5] and is associated with worse postoperative survival rates.[6,7] Therefore, the early identification of patients who are at high risk for early recurrence and the implementation of appropriate preventive measures has become a research hotspot in the field of liver cancer research.

2223

Numerous studies have identified risk factors for postoperative HCC recurrence, such as microvascular invasion (MVI), pathological differentiation, alpha-fetoprotein level, and tumor size.[8–11] However, the evaluation of these factors requires invasive procedures or postoperative pathological specimens; moreover, these factors do not fully reflect the heterogeneity of solid tumors.[12–14] Consequently, conventional imaging examinations and oncological markers have limitations in promptly diagnosing HCC recurrence and guiding individualized precise prevention and treatment strategies against early recurrence. In recent years, radiomics has emerged as a non-invasive technique that shows great potential in predicting cancer outcomes based on quantitative analysis.[15] Radiomics techniques can efficiently extract a wide range of imaging features from medical images in a high-throughput manner. The extracted features, such as texture, gray level, intensity, and morphological information, cannot be evaluated visually, but can reflect tumor heterogeneity at the cellular level. Thus, they transform digital medical images into quantitative features that reveal pathophysiological characteristics.[16–18] Furthermore, with the continuous advancement of computer technology and the emergence of artificial intelligence, machine learning and deep learning techniques have enabled the direct analysis of large-scale, complex, and diverse data. This capability provides new possibilities for processing and mining medical big data.[19–21] The above advances offer the possibility of multi-directional and multi-dimensional analyses for the clinical application of radiomics, and greatly promote the application of radiomics technology in the precise diagnosis and treatment of solid tumors. In previous studies of liver cancer, the processing of radiological image data was largely based on the extraction of two-dimensional (2D) features from the region of interest (ROI) with the maximum tumor section, followed by analysis and modeling of the extracted data.[22,23] However, relying solely on single 2D slices may result in the loss of inter-slice contextual information, leading to the incomplete representation of the overall tumor characteristics. Nevertheless, the use of three-dimensional (3D) images that comprehensively capture tumor information is challenging due to the significant memory consumption and computational burden associated with their analysis. Additionally, the analysis of 3D image data presents difficulties in network training, resulting in complex engineering or modeling failures.[24,25]

The present study introduces a novel approach that utilizes "2.5-dimensional" (2.5D) imaging data derived from the cross-section with the maximum tumor area and its adjacent slices, encompassing the transverse, coronal, and sagittal perspectives. By integrating multi-phase imaging modalities, namely, the arterial, plain, and portal phases, this method enhances the dimensionality of the data beyond conventional 2D slices without fully extending into 3D imaging. We employed convolutional neural networks (CNNs) to train on this enriched dataset, effectively leveraging its spatial coherence and diverse viewing angles. Furthermore, we aggregated features using multi-instance learning techniques and subsequently applied machine learning algorithms to efficiently model the data.

The above methodology has the potential to enhance diagnostic accuracy by leveraging the depth of 2.5D data while effectively managing computational efficiency. Consequently, we believe that this technique will be a viable option for advanced medical imaging applications. The aim of this study is to employ the above technique to predict the early postoperative recurrence of HCC. The workflow of this study is depicted in Figure 1.

Material and Methods
Patient Selection and Ethics Statement
This retrospective study was conducted at Ningxia Medical University General Hospital (Center 1) from January 1, 2018 to May 2023, and involved a total of 439 patients with HCC. The patients were randomly divided into a training cohort and an internal validation cohort at a ratio of 7:3. Additionally, an external validation cohort consisting of 91 hCC patients who underwent radical resection of liver cancer at the People's Hospital of Ningxia Hui Autonomous Region (Center 2) between January 2018 and May 2023 was used to evaluate the effectiveness of the predictive model. Patients meeting any of the following criteria were excluded from the study: (1) patients who received anti-tumor treatment such as radiofrequency ablation, arterial chemoembolization, or radiotherapy before the operation, (2) patients without preoperative CT images or poor image quality, (3) patients with confirmed distant metastases before surgery, and (4) patients with missing clinical and follow-up data. A flow chart of the inclusion and exclusion criteria for the study patients is shown in Figure 2. For the purpose of this study, we defined early recurrence as a time from curative treatment to the first recurrence of less than 2 years.[10] HCC recurrence refers to the appearance of a new tumor in or outside the
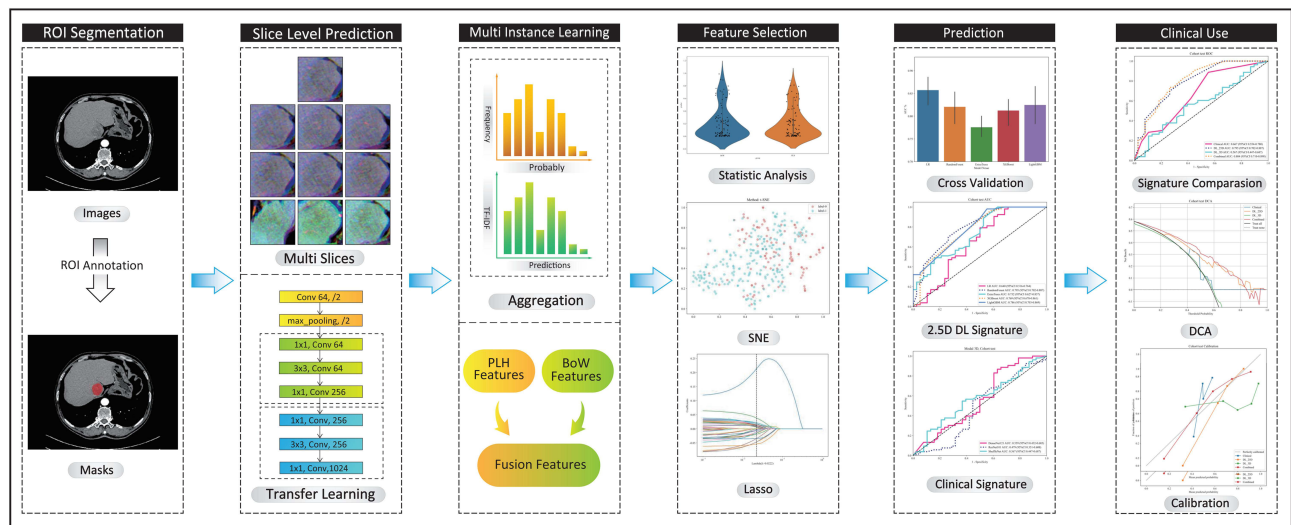
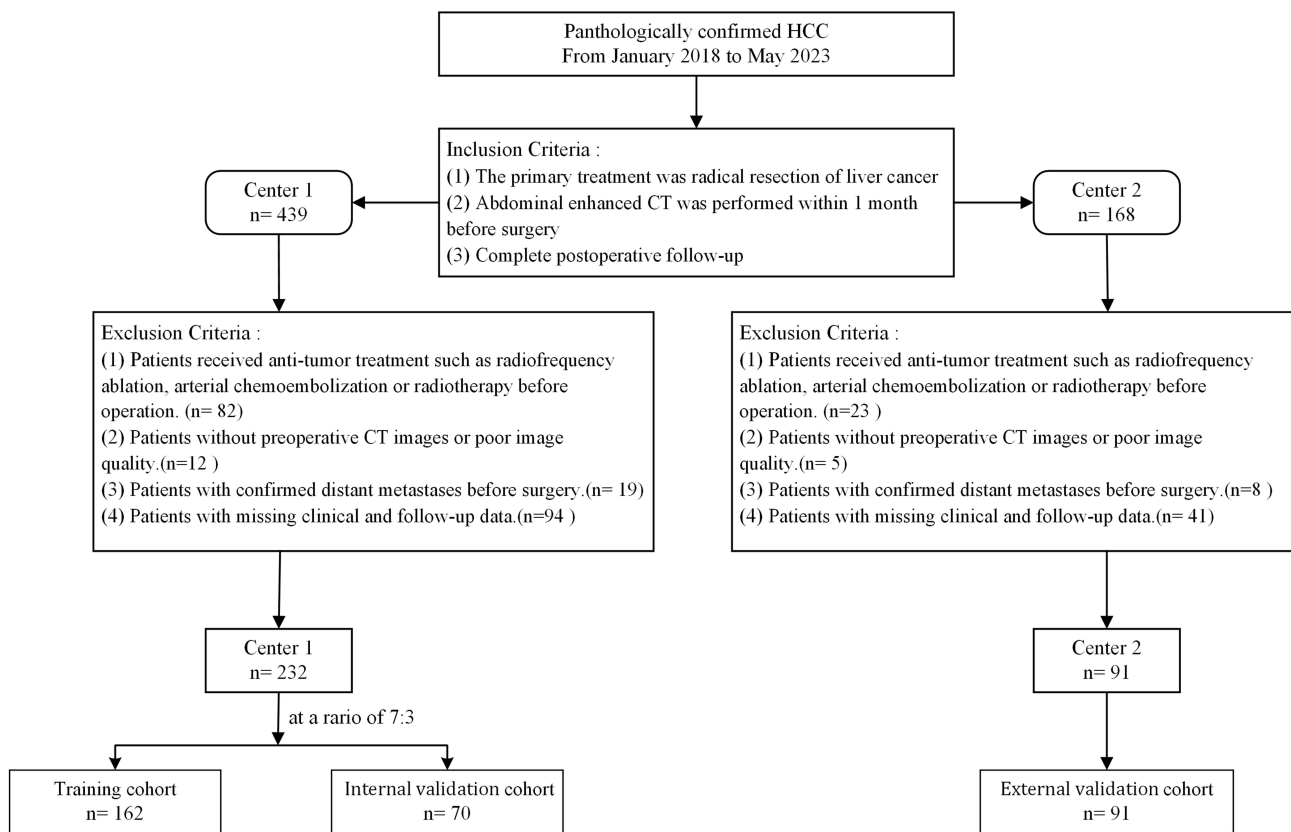**Figure 1** Illustration of the workflow of our study.

Figure 2 Flowchart of the criteria for patient selection.

liver after treatment, as determined using imaging or pathological findings. This study adhered to the ethical principles for medical research outlined in the World Medical Association Declaration of Helsinki. Ethics approval was obtained from the ethics committees of both the General Hospital of Ningxia Medical University (approval number: KYLL-2023-0232) and People's Hospital of Ningxia Hui Autonomous Region (approval number: 2023-LL-057). Due to the retrospective nature of this study the use of anonymous data collection, the requirement for written informed consent was waived.

Image Acquisition

A 256-slice spiral CT scanner (Brilliance iCT, Philips, the Netherlands) or a 64-slice spiral CT scanner (SOMATOM Definition, Siemens, Germany) was used for CT image acquisition at Center 1. At Center 2, a GE Revolution Apex CT (General Electric Medical Systems, Milwaukee, Wisconsin) scanner (512 rows) or a 256-row spiral CT scanner (Brilliance iCT, Philips, the Netherlands) was used. The patient was positioned in a supine posture, and scanning was performed from the dome of the diaphragm to the lower edge of the pubic symphysis plane within one breath hold. The scanning parameters were as follows: tube voltage, 100–120 kV; tube current, 150–250 mA; matrix size, 512*512 pixels; and slice thickness and interslice gap, approximately 1 mm. A total injection volume of contrast material (ioversol, 300 mgI/mL) equivalent to 1.5–2 mL/kg body weight was administered through an elbow vein at an injection rate of 2.5–3.0 mL/s, followed by a flush of normal saline solution totaling approximately 20 mL. Scanning was performed at approximately 30–35 s after contrast injection in the arterial phase and 60–65 s after contrast injection in the portal phase.

Image Segmentation

To enhance the precision and consistency of our medical image analysis, we established a standardized voxel spacing protocol. This protocol facilitated accurate comparisons across various volumes of interest by employing a fixed-resolution resampling technique. We uniformly adjusted the resolution to optimize the image quality for subsequent analytical processes. Additionally, we optimized the imaging parameters by setting the window width to 250 hounsfield units and the window level to 50 hounsfield units, ensuring optimal image contrast and clarity for further analysis.

We utilized ITK-SNAP software (version 3.6, www.itk-snap.org) to delineate the ROI of the tumor. All CT images were independently evaluated by 2 senior abdominal radiologists. Any discrepancies in their annotations were resolved through consultation with an expert radiologist with 20 years of experience who manually segmented the ROI layer by layer. All the radiologists were blinded to the clinical and pathological information.

2.5D Deep Learning Procedure

The contemporary literature on deep learning frequently utilizes the largest cross-section of the ROI, an approach that may overlook contextual information within the ROI.[26–28] To address this limitation, our model design incorporates the 3D characteristics of the ROI. We have developed a 2.5D deep learning model that enhances depiction by integrating several layers surrounding the central slice as well as data from multiple perspectives. This allows for a more precise representation. Additionally, we conducted comparative assessments with 3D models (details in Supplementary Material 1C).

2.5D Image cropping

We formulated a methodology to construct a series of 2D images by extracting a central slice as well as adjacent slices along both the superior-inferior and anterior-posterior axes. The selected range for the adjacent slices was set to ±1, ±2, and ±4, resulting in an ensemble of seven 2D images per patient. These images, centered on the maximal cross-sectional slice of the ROI, partially capture 3D structural data, thus constituting 2.5D data. Cropping was performed using the OKT-crop_max_roi tool of the OnekeyAI Platform, with parameters configured to encompass extended cross-sectional contexts by including slices at positions +1, +2, −1, −2, +4, and −4. Furthermore, we incorporated 3 different perspectives of the ROI area—transverse, sagittal, and coronal—yielding a total of 10 distinct 2D regions. The dataset also included images from 3 distinct imaging phases: arterial, plain, and portal. All these images were combined into a 3-channel input format to form our 2.5D data as shown in Figure 3.

Slice-level model training

In the training phase, we incorporated the generated 2.5D data into a transfer learning framework to assess its effectiveness. We evaluated the performance of several prominent deep learning architectures, including DenseNet121, ResNet101 ResNet50, and VGG19, which were all pre-trained on the ImageNet Large Scale Visual Recognition Challenge 2012 dataset. Additional details on model configurations and training procedures are provided in Supplementary Material 1A.
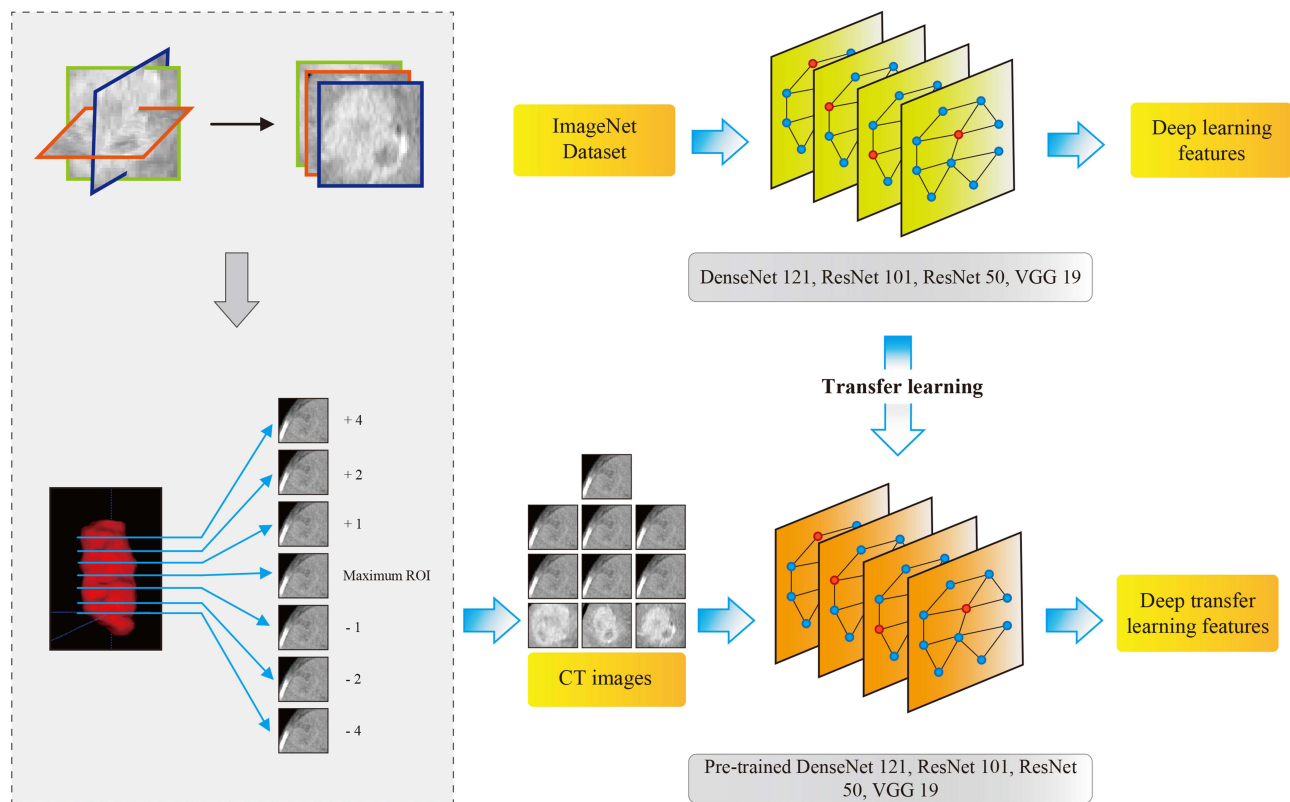
Figure 3 Procedure of 2.5-dimensional deep learning.

Multi-instance learning fusion

We utilized 2 multi-instance learning fusion techniques (detailed in Supplementary Material 1B):

1. Predict Likelihood Histogram: Using 2.5D deep learning models, we generated histograms to display the distribution of predictive probabilities and labels for each slice in the 2.5D images, providing a probabilistic summary of the prediction landscape.
2. Bag of Words: We analyzed each image by segmenting it into slices, and extracting probabilities and predictions from each slice. For each sample, 7 predictive results were compiled from the 2.5D and multi-model analyses. These results were treated like word frequencies in a document, and the Term Frequency-Inverse Document Frequency method was employed to characterize these features effectively.
3. Feature Fusion: We combined features from both the above techniques with radiomics features to enhance the dataset.

Signature Building
2.5D Deep Learning Signature
We applied dimensionality-reduction techniques, such as *t*-tests, correlation coefficients, and Lasso regularization, to the aggregated multi-instance learning features. We then modeled these features using prevalent machine learning algorithms, including Logistic Regression, RandomForest, ExtraTrees, XGBoost, and LightGBM. To ensure model robustness, we utilized 5-fold cross-validation within the training dataset and optimized hyperparameters through Grid Search.

3D Deep Learning Signature
We selected the model with the best performance in the internal validation set to as our 3D deep learning signature.

Clinical Signature

We conducted univariate analyses of clinical features by using the same models applied to the 2.5D deep learning data. This approach facilitated the development of a robust clinical model.

Metrics

We assessed the diagnostic performance of our deep learning model in the test cohort through the construction of receiver operating characteristic (ROC) curves. Additionally, we evaluated the calibration performance of the model by using calibration curves, and tested its calibration capabilities with the Hosmer-Lemeshow (HL) goodness-of-fit test. Decision curve analysis was also performed to ascertain the clinical utility of the predictive models.

Statistical Analysis

We assessed the normality of clinical data by using the Shapiro–Wilk test. Continuous variables were evaluated for statistical significance by employing the *t*-test or the Mann–Whitney *U*-test, depending on their distribution. Categorical variables were analyzed using the chi-square test. All statistical analyses were conducted in Python (version 3.7.12). The Python package Statsmodels (version 0.13.2) was utilized. Radiomics feature extraction was performed with PyRadiomics (version 3.0.1). Machine learning algorithms were implemented using Scikit-learn (version 1.0.2). Our deep learning models were developed using PyTorch (version 1.11.0), and were optimized for performance with CUDA (version 11.3.1) and cuDNN (version 8.2.1).

Results

Baseline Characteristics of Patients

During the study period, a total of 607 patients were found to have pathologically confirmed HCC after surgery in both centers. A total of 323 patients were enrolled in the study, according to the inclusion and exclusion criteria. Center 1 enrolled 232 patients, of whom 162 patients were assigned to the training cohort, and 70 patients were assigned to the internal validation cohort. Center 2 enrolled 91 patients, who served as the external validation cohort. The baseline characteristics of all the patients are presented in Table 1.

Clinical Signature

Univariable and multivariable analyses were conducted on all clinical features, and odds ratios (ORs) and corresponding *P*-values were calculated for each variable (Table 2). Notably, the features "pathological differentiation" and "Ki67 index" yielded *P*-values below 0.05, signifying statistical significance. Hence, these variables were chosen for inclusion in the development of the combined model. The detailed results of the clinical models can be found in Supplementary Material 2A.

Results of MIL_2.5D Signature

Slice-Level Results

We assessed the performance of 4 deep learning architectures—DenseNet121, ResNet101, ResNet50, and VGG19—across the training, internal validation, and external validation cohorts, with a primary focus on area under the curve (AUC) to measure model effectiveness. The results are shown in Figure 4 and Table 3. DenseNet121 exhibited a considerable decrease in AUC from 0.882 in the training cohort to 0.685 in the external validation cohort, indicating potential overfitting. ResNet101 showed strong performance in the training cohort with an AUC of 0.960, but its performance dropped in the internal validation cohort (AUC = 0.628) and external validation cohort (AUC = 0.739), suggesting variability in generalization. ResNet50 maintained a stable AUC across all cohorts (training cohort: 0.923, external validation cohort: 0.725). VGG19 was associated with an AUC of 0.768 in the training cohort, which decreased to 0.667 in the external validation cohort; however, with its AUC of 0.698 in the internal validation cohort, this architecture was considered competitive enough to ensure model reliability and prevent data leakage. Therefore, we selected the VGG19 model for inputting into our multi-instance learning framework, even though the other models

Table 1 Baseline Characteristics of the Study Cohorts

Feature	Training cohort			Internal validation cohort			External validation cohort		
	No recurrence	**Recurrence**	**P value**	**No recurrence**	**Recurrence**	**P value**	**No recurrence**	**Recurrence**	**P value**
TBIL (μmol/L)	19.15±19.55	20.44±24.26	0.989	18.43 ± 5.72	15.59 ± 8.29	0.046	19.79±20.01	21.51±24.62	0.629
Pathological differentiation			0.003			0.845			0.733
Poor	41 (71.93%)	49 (46.67%)		9 (64.29%)	40 (71.43%)		30 (78.95%)	39 (73.58%)	
Moderate or well	16 (28.07%)	56 (53.33%)		5 (35.71%)	16 (28.57%)		8 (21.05%)	14 (26.42%)	
Liver cirrhosis			0.419			0.951			0.222
No	21 (36.84%)	47 (44.76%)		6 (42.86%)	21 (37.50%)		23 (60.53%)	24 (45.28%)	
Yes	36 (63.16%)	58 (55.24%)		8 (57.14%)	35 (62.50%)		15 (39.47%)	29 (54.72%)	
Tumor size, cm			0.022			0.671			0.229
<5	37 (64.91%)	47 (44.76%)		7 (50.00%)	22 (39.29%)		21 (55.26%)	37 (69.81%)	
≥5	20 (35.09%)	58 (55.24%)		7 (50.00%)	34 (60.71%)		17 (44.74%)	16 (30.19%)	
Number of tumors			0.78			0.484			0.027
Solitary	50 (87.72%)	89 (84.76%)		11 (78.57%)	36 (64.29%)		37 (97.37%)	42 (79.25%)	
Multiple	7 (12.28%)	16 (15.24%)		3 (21.43%)	20 (35.71%)		1 (2.63%)	11 (20.75%)	
Intratumoral necrosis			1.0			0.671			0.115
No	26 (45.61%)	49 (46.67%)		7 (50.00%)	22 (39.29%)		15 (39.47%)	31 (58.49%)	
Yes	31 (54.39%)	56 (53.33%)		7 (50.00%)	34 (60.71%)		23 (60.53%)	22 (41.51%)	
Tumor margin			0.623			0.003			0.342
No	34 (59.65%)	57 (54.29%)		4 (28.57%)	42 (75.00%)		19 (50.00%)	20 (37.74%)	
Yes	23 (40.35%)	48 (45.71%)		10 (71.43%)	14 (25.00%)		19 (50.00%)	33 (62.26%)	
Adjacency to major vessels			1.0			1.0			0.526
No	35 (61.40%)	64 (60.95%)		9 (64.29%)	35 (62.50%)		24 (63.16%)	38 (71.70%)	
Yes	22 (38.60%)	41 (39.05%)		5 (35.71%)	21 (37.50%)		14 (36.84%)	15 (28.30%)	
Age, years			0.98			0.421			<0.001
<60	34 (59.65%)	61 (58.10%)		7 (50.00%)	37 (66.07%)		13 (34.21%)	39 (73.58%)	
≥60	23 (40.35%)	44 (41.90%)		7 (50.00%)	19 (33.93%)		25 (65.79%)	14 (26.42%)	
Gender			0.815			0.443			0.27
Male	14 (24.56%)	29 (27.62%)		5 (35.71%)	12 (21.43%)		11 (28.95%)	9 (16.98%)	
Female	43 (75.44%)	76 (72.38%)		9 (64.29%)	44 (78.57%)		27 (71.05%)	44 (83.02%)	
HBV infection			0.44			0.285			<0.001
No	18 (31.58%)	41 (39.05%)		7 (50.00%)	17 (30.36%)		24 (63.16%)	12 (22.64%)	
Yes	39 (68.42%)	64 (60.95%)		7 (50.00%)	39 (69.64%)		14 (36.84%)	41 (77.36%)	
HBV DNA, IU/mL			0.527			0.095			0.012
<500	41 (71.93%)	69 (65.71%)		12 (85.71%)	32 (57.14%)		33 (86.84%)	32 (60.38%)	
≥500	16 (28.07%)	36 (34.29%)		2 (14.29%)	24 (42.86%)		5 (13.16%)	21 (39.62%)	
PLT, ×10⁹ cells/L			0.54			0.355			0.026
<125	18 (31.58%)	27 (25.71%)		5 (35.71%)	11 (19.64%)		6 (15.79%)	21 (39.62%)	
≥125	39 (68.42%)	78 (74.29%)		9 (64.29%)	45 (80.36%)		32 (84.21%)	32 (60.38%)	
AFP, ng/mL			0.116			0.095			0.234
<400	43 (75.44%)	65 (61.90%)		12 (85.71%)	32 (57.14%)		29 (76.32%)	33 (62.26%)	
≥400	14 (24.56%)	40 (38.10%)		2 (14.29%)	24 (42.86%)		9 (23.68%)	20 (37.74%)	

(*Continued*)

Table 1 (Continued).

Feature	Training cohort			Internal validation cohort			External validation cohort		
	No recurrence	Recurrence	*P* value	No recurrence	Recurrence	*P* value	No recurrence	Recurrence	*P* value
ALB, g/L			1.0			1.0			0.08
<40	37 (64.91%)	69 (65.71%)		8 (57.14%)	33 (58.93%)		18 (47.37%)	36(67.92%)	
≥40	20 (35.09%)	36 (34.29%)		6 (42.86%)	23 (41.07%)		20 (52.63%)	17(32.08%)	
AST, U/L			0.092			0.951			0.718
<40	42 (73.68%)	62 (59.05%)		8 (57.14%)	35 (62.50%)		28 (73.68%)	36 (67.92%)	
≥40	15 (26.32%)	43 (40.95%)		6 (42.86%)	21 (37.50%)		10 (26.32%)	17 (32.08%)	
ALT, U/L			0.638			1.0			0.182
<50	45 (78.95%)	78 (74.29%)		10 (71.43%)	39 (69.64%)		32 (84.21%)	37 (69.81%)	
≥50	12 (21.05%)	27 (25.71%)		4 (28.57%)	17 (30.36%)		6 (15.79%)	16 (30.19%)	
Ki67			0.003			0.35			0.091
Low	28 (49.12%)	26 (24.76%)		7 (50.00%)	18 (32.14%)		22 (57.89%)	20 (37.74%)	
High	29 (50.88%)	79 (75.24%)		7 (50.00%)	38 (67.86%)		16 (42.11%)	33 (62.26%)	
MVI			0.007			0.101			0.463
No	40 (70.18%)	49 (46.67%)		9 (64.29%)	20 (35.71%)		28 (73.68%)	34 (64.15%)	
Yes	17 (29.82%)	56 (53.33%)		5 (35.71%)	36 (64.29%)		10 (26.32%)	19 (35.85%)	
Child-Pugh			1.0			0.789			0.646
A/B	48 (84.21%)	87 (82.86%)		13 (92.86%)	48 (85.71%)		35 (92.11%)	46 (86.79%)	
C	9 (15.79%)	18 (17.14%)		1 (7.14%)	8 (14.29%)		3 (7.89%)	7 (13.21%)	
ALBI			0.754			0.763			0.114
1	22 (38.60%)	38 (36.19%)		7 (50.00%)	23 (41.07%)		20 (52.63%)	17 (32.08%)	
2	34 (59.65%)	63 (60.00%)		7 (50.00%)	33 (58.93%)		18 (47.37%)	35 (66.04%)	
3	1 (1.75%)	4 (3.81%)		Null	Null		Null	1 (1.89%)	

Abbreviations: TBIL, total bilirubin; PLT, platelet count; AFP, alpha-fetoprotein; ALB, albumin; AST, aspartate transaminase; ALT, alanine aminotransferase; MVI, microvascular invasion; ALBI, albumin-bilirubin; HBV, hepatitis B virus.

demonstrated superior performance in the test dataset. This decision was made to prevent data leakage and ensure the integrity of our predictive modeling process (Table 3).

Grad-CAM

To investigate the recognition capabilities of the deep learning models on varied samples, we employed the gradient-weighted class activation mapping (Grad-CAM) technique for visualizing the activations in the final convolutional layer associated with cancer-type predictions. Figure 5 illustrates the application of Grad-CAM, highlighting the image regions that significantly influenced the decision-making process of the models. This enhances our understanding of model interpretability.

MIL Fusion Results

The logistic regression model exhibited good performance in the training cohort (AUC = 0.964), but this significantly declined in the internal and external validation cohorts (AUC = 0.762 and 0.640, respectively), indicating potential overfitting or limited generalizability (Figure 6 and Table 4). The RandomForest algorithm maintained its robust performance across all cohorts, achieving the highest AUC of 0.920 in the training cohort and a strong AUC of 0.795 in the external validation cohort, with a high specificity of 0.929 despite lower validation accuracy. The ExtraTrees model showed good results in the training cohort (AUC = 0.945) and reasonable performance in the external validation cohort (AUC = 0.732), though it exhibited variability in validation. Both XGBoost and LightGBM performed well in the training cohort but showed variable outcomes in the other cohorts, with XGBoost demonstrating a notable drop in sensitivity during testing.

Table 2 Univariable and Multivariable Analyses of Clinical Features

Feature	Univariable analysis			Multivariable analysis		
	OR	95% CI	*P*-value	OR	95% CI	*P*-value
Liver cirrhosis	0.928	0.819–1.053	0.332			
HBV infection	0.929	0.816–1.058	0.349			
Gender	0.965	0.838–1.112	0.676			
Intratumoral necrosis	0.990	0.874–1.123	0.899			
ALB	0.992	0.870–1.131	0.919			
TBIL	1.001	0.998–1.003	0.731			
Adjacency to major vessels	1.004	0.883–1.141	0.955			
Age	1.015	0.894–1.151	0.849			
Child-Pugh class	1.022	0.865–1.209	0.827			
ALBI	1.036	0.922–1.164	0.615			
Tumor margin	1.051	0.927–1.191	0.514			
Number of tumors	1.057	0.884–1.264	0.609			
ALT	1.060	0.916–1.226	0.511			
HBV DNA	1.067	0.934–1.220	0.422			
PLT	1.069	0.930–1.229	0.429			
AFP	1.149	1.008–1.310	0.082			
AST	1.156	1.016–1.315	0.064			
Tumor size	1.202	1.063–1.359	<0.05	1.108	0.971–1.264	0.197
MVI	1.242	1.099–1.404	<0.05	1.15	1.007–1.314	0.084
Pathological differentiation	1.263	1.117–1.426	<0.05	1.229	1.093–1.383	<0.05
Ki67 index	1.284	1.130–1.459	<0.05	1.212	1.069–1.374	<0.05

Abbreviations: OR, odds ratio; CI, confidence interval; HBV, hepatitis B virus; TBIL, total bilirubin; PLT, platelet count; AFP, alpha-fetoprotein; ALB, albumin; AST, aspartate transaminase; ALT, alanine aminotransferase; MVI, microvascular invasion; ALBI, albumin-bilirubin.

The above analysis underscored the potential of the RandomForest algorithm, particularly when augmented with multi-instance learning fusion, for delivering reliable and robust predictions. The ability of this model to effectively generalize across diverse datasets makes it highly suitable for applications where accuracy and reliability are critical. The fusion approach significantly elevated model performance, recommending its use in scenarios demanding high dependability (Supplementary Material 2B).

Signature Comparison
Predictive performance
The comparative analysis of the clinical, 2.5D, 3D, and combined models revealed distinct performance patterns across the training, internal validation, and external validation cohorts, which were primarily assessed through AUC. The
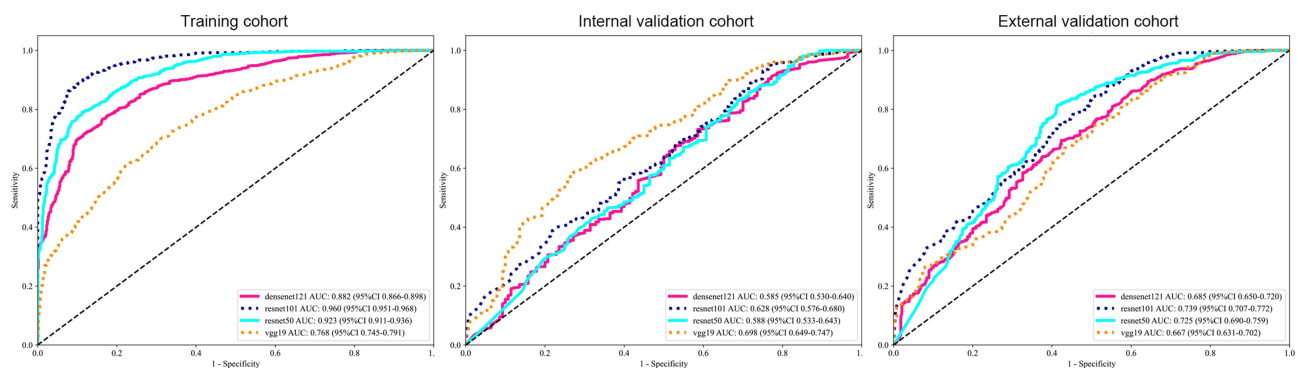

Figure 4 Receiver operating characteristic curves of different models in slice-level prediction.

Table 3 Slice-Level Results of Different CNN Models

Model	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
Densenet121	0.794	0.882	0.8656–0.8983	0.778	0.823	0.890	0.668	Training
Densenet121	0.633	0.585	0.5299–0.6399	0.673	0.471	0.836	0.265	Internal
Densenet121	0.642	0.685	0.6496–0.7197	0.689	0.576	0.694	0.570	External
Resnet101	0.894	0.960	0.9508–0.9682	0.896	0.891	0.938	0.823	Training
Resnet101	0.804	0.628	0.5764–0.6802	0.946	0.236	0.832	0.524	Internal
Resnet101	0.702	0.739	0.7071–0.7718	0.875	0.461	0.694	0.726	External
Resnet50	0.824	0.923	0.9107–0.9362	0.790	0.888	0.928	0.696	Train
Resnet50	0.724	0.588	0.5326–0.6432	0.827	0.314	0.828	0.312	Internal
Resnet50	0.718	0.725	0.6904–0.7595	0.811	0.587	0.733	0.690	External
VGG19	0.686	0.768	0.7448–0.7913	0.667	0.721	0.815	0.540	Training
VGG19	0.611	0.698	0.6486–0.7468	0.580	0.736	0.898	0.305	Internal
VGG19	0.635	0.667	0.6312–0.7023	0.738	0.492	0.670	0.574	External

Abbreviations: CNN, convolutional neural network; AUC, area under the curve; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value.

clinical model showed moderate effectiveness in the training cohort (AUC = 0.719), but poor performance in the internal validation (AUC = 0.614) and external validation cohorts (AUC = 0.667), suggesting limited generalizability. The 2.5D model excelled with high efficacy in the training cohort (AUC = 0.920) and maintained a strong performance in the internal validation (AUC = 0.825) and external validation cohorts (AUC = 0.795), indicating robust capability in integrating contextual information. The 3D model, while adequate in the training cohort (AUC = 0.751), showed a noticeable drop in performance in the internal validation cohort (AUC = 0.666) and a further decline in the external validation cohort (AUC = 0.567), confirming its susceptibility to overfitting due to its complexity and the absence of pre-trained parameters. The combined model leveraged the strengths of the individual models, demonstrating superior integration and generalizability with impressive AUCs in the training (0.921), internal validation (0.835), and external validation (0.804) cohorts. The fusion of the clinical and 2.5D models to create the combined model notably enhanced prediction accuracy by effectively capturing detailed imaging data alongside relevant clinical contexts. Thus, the above results indicated that the 2.5D model, which leverages the contextual information around the primary image slice, performed robustly across all datasets, with particularly strong results when combined with clinical features. The 3D model, while theoretically promising due to its depth of analysis, showed a clear tendency towards overfitting, which was likely exacerbated by the absence of pre-trained parameters. The combined model not only improved upon the robustness of the 2.5D model but also enhanced its generalizability, as evidenced by its consistently high AUCs across all cohorts, particularly in the internal and external validation cohorts. These findings suggest that integrating 2.5D deep learning with clinical features effectively captures both the detailed imaging data and relevant clinical context, thereby significantly enhancing prediction accuracy (Table 5 and Figure 7).

The HL test quantifies the discrepancy between predicted probabilities and observed outcomes. A higher HL statistic indicates superior calibration, showing that the model's predictions align closely with actual results. In this study, the combined model exhibited the highest calibration performance, with HL test statistics of 0.716, 0.297, and 0.408 in the training, internal validation, and external validation cohorts, respectively, all of which were significantly greater than 0.05 (Figure 8A). The DeLong test was applied to both the training and validation sets. The results highlighted the superior performance of the combined model, which integrated clinical and deep learning results (Figure 8B). This model not only showed a marked improvement in performance but also significantly outperformed the clinical-only approach, with P-values < 0.05. The results of the decision curve analysis for both the training and validation sets indicated that our combined model provided considerable advantages in terms of predicted probabilities (Figure 8C). Furthermore, it consistently offered a greater potential for net benefit as compared to the other signatures, underscoring its effectiveness. A nomogram based on the combined model was constructed in the training cohort to predict the early recurrence of liver cancer after surgery (Figure 9).
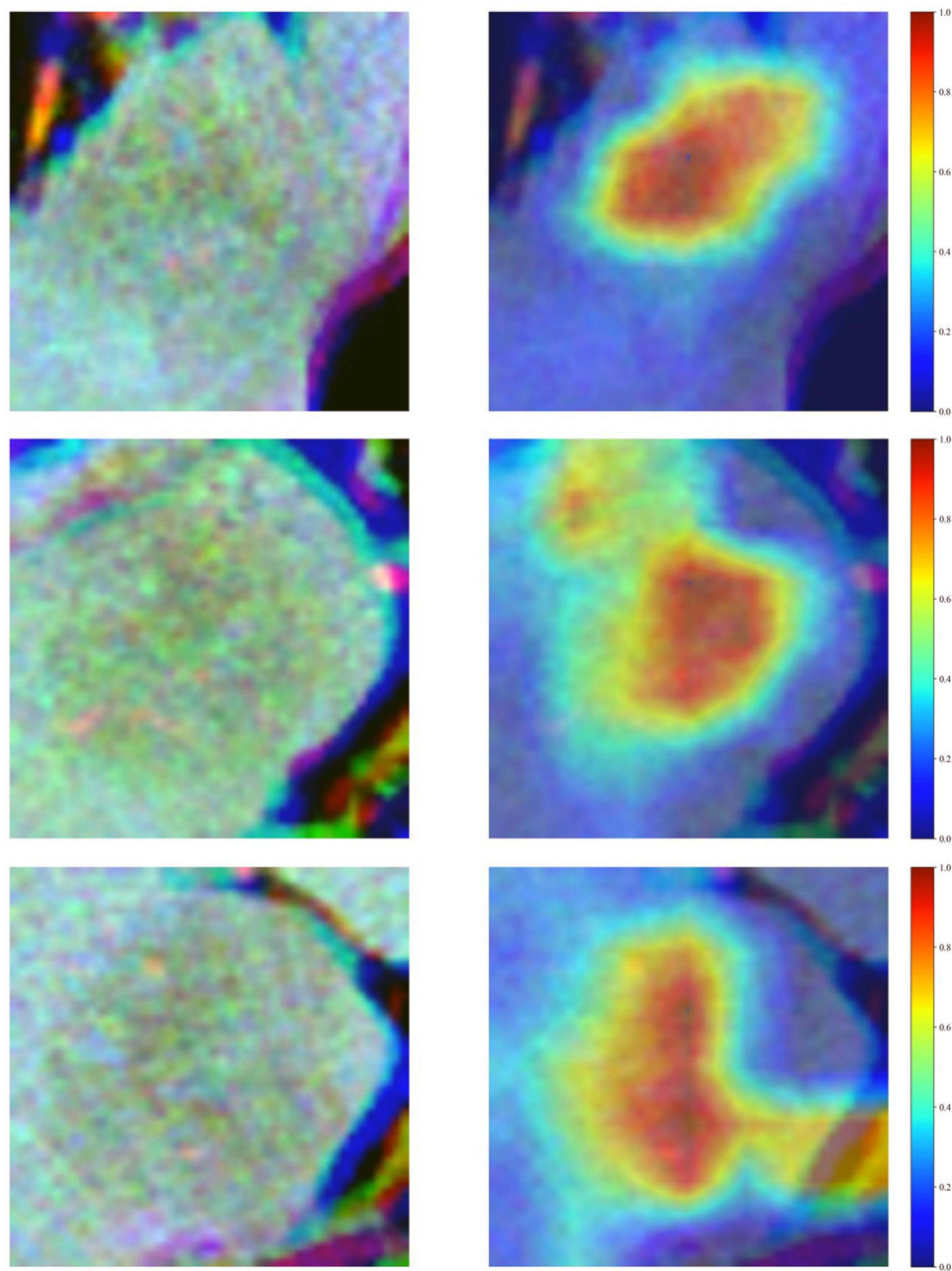
Figure 5 Grad-CAM visualizations for 2 representative samples, demonstrating how the 2.5D VGG19 model selectively focuses on different regions of the images to make its predictions. This visualization is crucial for understanding the attention mechanism of model in practical applications.

Discussion

The prediction of early recurrence of HCC after surgery can facilitate early intervention, reduce unnecessary adjuvant therapy, and optimize the treatment plan. This approach ultimately prolongs patient survival, improves their quality of life, and saves significant medical costs. There is evidence that radiomics has great potential in predicting HCC recurrence.[29,30] Studies have demonstrated that a deep learning model based on multilayer CT images outperforms a deep learning model based on single CT images.[20,31] In this study, we collected data from patients who underwent radical resection of HCC from 2 centers. We introduced an innovative method for the processing of liver cancer radiomics data. Specifically, we utilized a 2.5D imaging data approach that involved obtaining maximum cross-
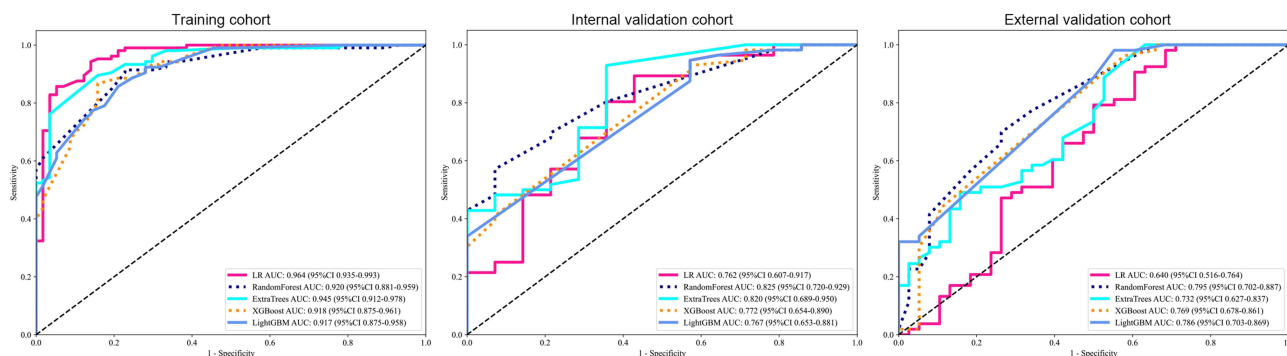
Figure 6 Cross validation results for parameter grid search. Receiver operating characteristic curves of different models in patient-level prediction.

sections and surrounding slices in multiple planes (transverse, coronal, and sagittal planes) as well as integrating multiphase imaging modes (arterial phase, plain phase, and portal phase).

CT images play an important role in the study of the prognosis of HCC patients. With its high resolution and wide application, CT imaging is an important tool for clinical diagnosis and evaluation. Moreover, CT scans are highly standardized, and the image quality is relatively consistent between different institutions, making it a convenient test for multi-center studies. In the diagnosis and treatment of liver cancer, CT scans are one of the most widely used examinations, with lower costs than those of PET-CT and MRI, making them suitable for routine examinations. Therefore, in this study, we used CT images. We incorporated the generated 2.5D data into the transfer learning framework, which is a common technique in deep learning to solve the problem of limited training data hindering the generalizability of deep learning algorithms.[32,33] The implementation of CNNs for classification tasks usually requires training a large number of parameters, and usually consumes a large number of strongly labeled samples when training from scratch.[34,35] In transfer learning, trained model parameters are transferred to a new model to help the new model train; thus, the new model can benefit from what was learned in the previous task and learn new tasks faster.[34,36,37]

We also utilized a multi-instance learning approach to enhance predictive accuracy, integrating various data points from a single sample to create a comprehensive feature set. The process involved several steps. First, for slice-level predictions, the deep learning models were used to make predictions based on each slice individually and obtain the corresponding probabilities and labels. Next, multi-instance learning feature aggregation was performed using 2

Table 4 Metrics of Different Machine Learning Methods in Multi-Instance Learning Models

| Model	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
LR	0.883	0.964	0.935–0.993	0.848	0.947	0.967	0.771	Training
LR	0.814	0.762	0.607–0.917	0.875	0.571	0.891	0.533	Internal
LR	0.681	0.640	0.516–0.764	0.887	0.395	0.671	0.714	External
RandomForest	0.858	0.920	0.881–0.959	0.905	0.772	0.880	0.815	Training
RandomForest	0.571	0.825	0.720–0.929	0.482	0.929	0.964	0.310	Internal
RandomForest	0.703	0.795	0.702–0.887	0.679	0.737	0.783	0.622	External
ExtraTrees	0.833	0.945	0.912–0.978	0.762	0.965	0.976	0.687	Training
ExtraTrees	0.857	0.820	0.689–0.950	0.911	0.643	0.911	0.643	Internal
ExtraTrees	0.714	0.732	0.627–0.837	0.887	0.474	0.701	0.750	External
XGBoost	0.852	0.918	0.875–0.961	0.857	0.842	0.909	0.762	Training
XGBoost	0.814	0.772	0.654–0.890	0.911	0.429	0.864	0.545	Internal
XGBoost	0.626	0.769	0.678–0.861	0.434	0.895	0.852	0.531	External
LightGBM	0.802	0.917	0.875–0.958	0.790	0.825	0.892	0.681	Training
LightGBM	0.814	0.767	0.653–0.881	0.911	0.429	0.864	0.545	Internal
LightGBM	0.725	0.786	0.703–0.869	0.887	0.500	0.712	0.760	External

Abbreviations: AUC, area under the curve; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; LR, logistic regression.

Table 5 Metrics on Different Signatures

Signature	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Cohort
Clinical	0.568	0.719	0.6414–0.7966	0.362	0.947	0.927	0.446	Training
DL_2.5D	0.858	0.920	0.8806–0.9590	0.905	0.772	0.880	0.815	Training
DL_3D	0.704	0.751	0.6741–0.8283	0.638	0.825	0.870	0.553	Training
Combined	0.846	0.921	0.8806–0.9608	0.876	0.789	0.885	0.776	Training
Clinical	0.200	0.614	0.4581–0.7689	0.000	1.000	0.000	0.200	Internal
DL_2.5D	0.571	0.825	0.7202–0.9290	0.482	0.929	0.964	0.310	Internal
DL_3D	0.614	0.666	0.4855–0.8461	0.571	0.786	0.914	0.314	Internal
Combined	0.700	0.835	0.7235–0.9474	0.661	0.857	0.949	0.387	Internal
Clinical	0.604	0.667	0.5536–0.7801	0.623	0.579	0.673	0.524	External
DL_25D	0.703	0.795	0.7019–0.8875	0.679	0.737	0.783	0.622	External
DL_3D	0.571	0.567	0.4467–0.6874	0.528	0.632	0.667	0.490	External
Combined	0.714	0.804	0.7098–0.8984	0.698	0.737	0.787	0.636	External

Abbreviations: AUC, area under the curve; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; DL_25D, 2.5D deep learning signature; DL_3D, 3D deep learning signature.

technique. In the first technique—histogram feature aggregation—each distinct number was treated as a "bin" to count occurrences across types. We utilized 2.5D deep learning models to create histograms that visually illustrated the spread of predictive probabilities and labels across individual slices within the 2.5D images. This method offered a probabilistic summary of the predicted situation. The second multi-instance learning technique we utilized was the bag of words feature aggregation. We analyzed each image by segmenting it into slices, and extracted probabilities and predictions from each slice. For each sample, 7 predictive results were compiled from the 2.5D and multi-model analyses. These results were treated like word frequencies in a document, and the Term Frequency-Inverse Document Frequency method was used to effectively characterize these features. We combined features from both the above multi-instance learning techniques with radiomics features to enhance the dataset. This approach leveraged diverse data sources to boost the representational power of image attributes, significantly improving model accuracy in classification tasks. Finally, machine learning algorithms were applied to build models using the obtained data. To construct the model, we applied 4 deep learning architectures to the multi-instance learning framework. We found that the VGG19 model showed superior performance, with AUCs of 0.768, 0.698, and 0.667 in the training, internal validation, and external validation cohorts, respectively, which supported its selection for reasons of model reliability and prevention of data leakage. We used popular machine learning algorithms (Logistic Regression, RandomForest, ExtraTrees, XGBoost, and LightGBM) to model the aggregated multi-instance learning features after dimensionality reduction. The results showed that the random forest algorithm had considerable potential. Random forest is an integrated learning technique that generates predictions by building and merging many decision trees.[38] When augmented with multi-instance learning fusion, this algorithm delivered reliable and robust predictions, with AUCs of 0.920, 0.825, and 0.795 in the training, internal validation, and external validation cohorts, respectively. The analysis of clinical features showed that the degree of tumor
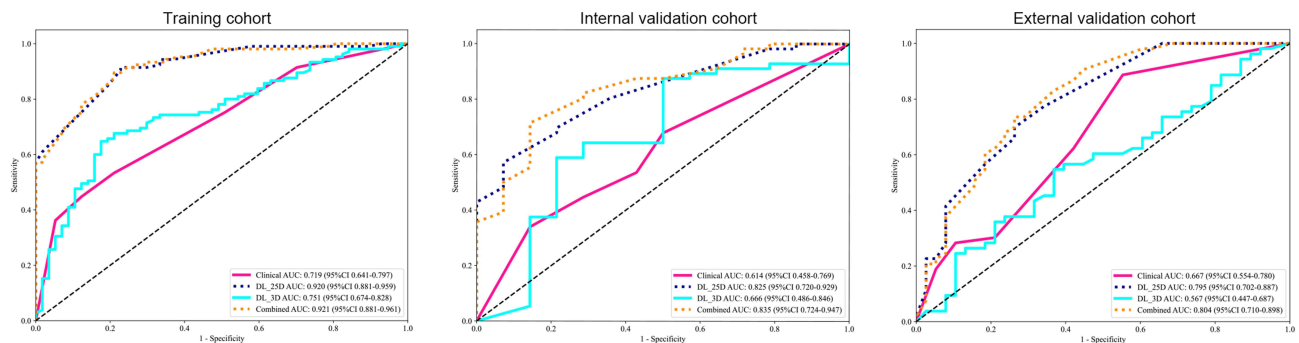
Figure 7 Receiver operating characteristic curves of different signatures in different cohorts.
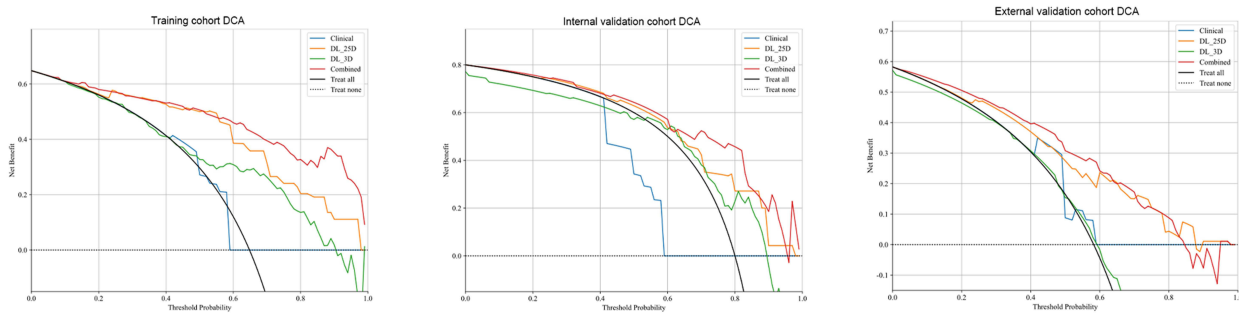
A
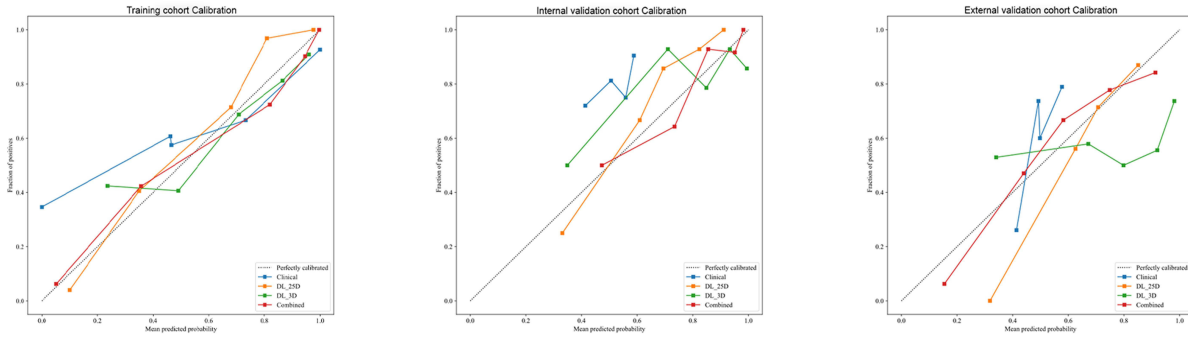
B
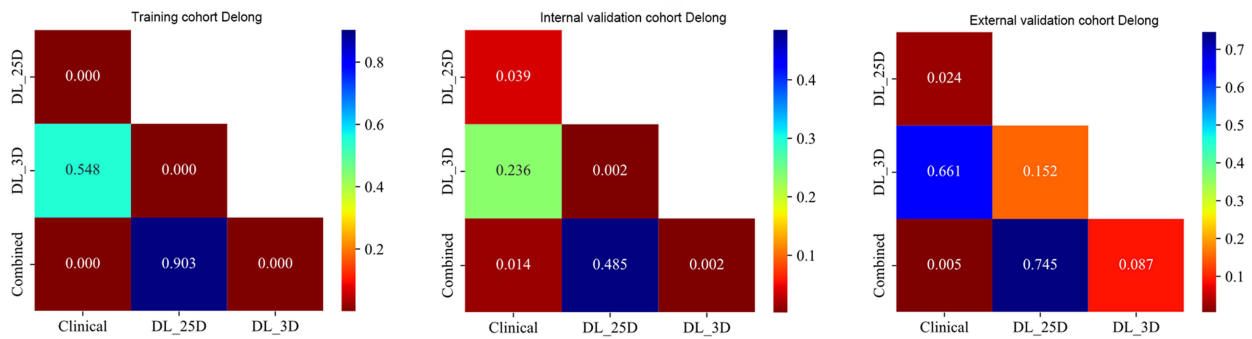
C

Figure 8 (A) Calibration curves of different signatures in different cohorts. **(B)** Heatmaps of *P*-values on the DeLong test for different signatures. **(C)** Decision curves of different signatures in the study cohorts.
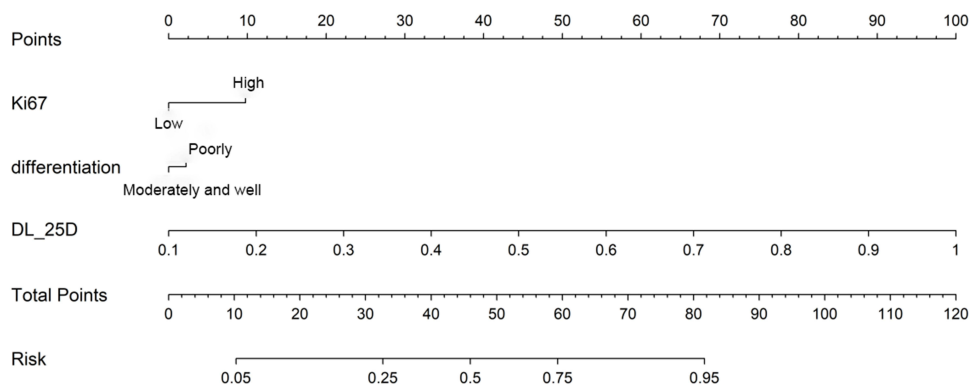

Figure 9 The constructed nomogram for the combined model.

differentiation and the Ki67 index were closely related to the early recurrence of HCC after surgery. Ki-67 is a nuclear antigen related to cell proliferation activity, which is commonly used to reflect the level of cell proliferation.[39] Pathological differentiation degree is an important factor influencing the recurrence of HCC; the lower the degree of pathological differentiation, the higher the risk of postoperative HCC recurrence.[40] The above findings are consistent with those of previous studies.[41,42] To improve the accuracy of prediction, we combined 2.5D deep learning with clinical features to construct a combined model. We compared the AUCs of the clinical, 2.5D, 3D, and combined models in the training, internal validation, and external validation cohorts to evaluate their respective predictive performances. The results showed that the clinical model showed moderate predictive performance in the training cohort (AUC = 0.719), but poor performance in the internal validation cohort (AUC = 0.614) and external validation cohort (AUC = 0.667), suggesting limited generalizability. The 2.5D model performed well in the training, internal validation, and external validation cohorts, with AUCs of 0.920, 0.825 and 0.795, respectively, indicating a strong ability to integrate contextual information. Although the AUC of the 3D model was promising in the training cohort (AUC = 0.751), it significantly decreased in the internal validation cohort (AUC = 0.666) and further declined in the external validation cohort (AUC = 0.567). These results confirm that the complexity and absence of pre-training parameters rendered the 3D model susceptible to overfitting, thereby impacting its overall predictive performance. In contrast, the combined model capitalized on the individual strengths of each component model, showcasing superior integration capabilities and generalizability with impressive AUCs across all cohorts (training cohort: 0.921, internal validation cohort: 0.8305, external validation cohort: 0.804). Moreover, the combined model consistently outperformed the other models in both the HL test and DeLong test, demonstrating its superiority. Additionally, decision curve analysis consistently revealed a greater potential for benefit for the combined model. These findings strongly underscore the effectiveness of the combined model.

The aforementioned findings show the exceptional performance of 2.5D image data in radiomics-based deep learning models. This can be attributed to the fact that CT scans are 3D images comprised of multiple 2D slices, which can be transformed into several so-called 2.5D images. The 2.5D image retains a subset of adjacent 2D slices while maintaining identical pixel height and width as a 2D image, thereby preserving certain original 3D features of CT scans. By leveraging the contextual information surrounding the primary image slice, the 2.5D DL model exhibited robustness across all datasets and particularly excelled when combined with clinical features for prediction purposes. Although theoretically promising due to their comprehensive analysis capabilities, 3D deep learning models often encounter limitations in performance due to the use of small and non-diverse datasets, displaying a clear inclination towards overfitting, which is potentially exacerbated by the absence of pretrained parameters.[25,43] In contrast, our constructed combined model not only enhanced the resilience of the 2.5D model but also amplified its versatility as evidenced by consistently high AUCs across all cohorts, especially the internal and external validation cohorts. These results demonstrate that integrating clinical features with 2.5D deep learning effectively captures intricate imaging data along with the relevant clinical context, leading to significantly improved prediction accuracy. However, certain limitations should be considered. First, the manual whole-tumor ROI sketching process was time-consuming; therefore, developing an accurate and stable automatic segmentation method for tumor ROIs would greatly enhance the clinical utility of our model. Second, as our study is retrospective in nature, a potential data-selection bias may exist; thus, larger multicenter prospective studies are required to further validate the clinical value of our model.

Conclusion

The combined model, integrating 2.5D deep learning and clinical features, demonstrated excellent applicability in predicting the early recurrence of HCC after surgery. This will not only optimize postoperative treatment and monitoring plans while reducing the wastage of medical resources but will also enable personalized interventions for high-risk patients, thereby potentially improving patient prognosis.

Funding

Disclosure

The authors report no conflicts of interest in this work.

References

1. Sim YK, Chong MC, Gandhi M, et al. Real-world data on the diagnosis, treatment, and management of hepatocellular carcinoma in the Asia-Pacific: the INSIGHT study. *Liver Cancer*. 2024;13(3):298–313. doi:10.1159/000534513

2. Niu ZS, Wang WH, Niu XJ. Recent progress in molecular mechanisms of postoperative recurrence and metastasis of hepatocellular carcinoma. *World J Gastroenterol*. 2022;28(46):6433–6477. doi:10.3748/wjg.v28.i46.6433

3. Llovet JM, Kelley RK, Villanueva A, et al. Hepatocellular carcinoma. *Nat Rev Dis Primers*. 2021;7(1):6. doi:10.1038/s41572-020-00240-3

4. Zhang Y, Lei X, Xu L, Lv X, Xu M, Tang H. Preoperative and postoperative nomograms for predicting early recurrence of hepatocellular carcinoma without macrovascular invasion after curative resection. *BMC Surg*. 2022;22(1):233. doi:10.1186/s12893-022-01682-0

5. Zhang ZH, Jiang C, Qiang ZY, et al. Role of microvascular invasion in early recurrence of hepatocellular carcinoma after liver resection: a literature review. *Asian J Surg*. 2024;47(5):2138–2143. doi:10.1016/j.asjsur.2024.02.115

6. He W, Peng B, Tang Y, et al. Nomogram to predict survival of patients with recurrence of hepatocellular carcinoma after surgery. *Clin Gastroenterol Hepatol*. 2018;16(5):756–764.e710. doi:10.1016/j.cgh.2017.12.002

7. Wei T, Zhang XF, Bagante F, et al. Early versus late recurrence of hepatocellular carcinoma after surgical resection based on post-recurrence survival: an international multi-institutional analysis. *J Gastrointest Surg*. 2021;25(1):125–133. doi:10.1007/s11605-020-04553-2

8. Xia F, Zhang Q, Ndhlovu E, Zheng J, Gao H, Xia G. A nomogram for preoperative prediction of microvascular invasion in ruptured hepatocellular carcinoma. *Eur J Gastroenterol Hepatol*. 2023;35(5):591–599. doi:10.1097/MEG.0000000000002535

9. Shen J, Liu J, Li C, Wen T, Yan L, Yang J. The impact of tumor differentiation on the prognosis of HBV-associated solitary hepatocellular carcinoma following hepatectomy: a propensity score matching analysis. *Dig Dis Sci*. 2018;63(7):1962–1969. doi:10.1007/s10620-018-5077-5

10. Yan WT, Li C, Yao LQ, et al. Predictors and long-term prognosis of early and late recurrence for patients undergoing hepatic resection of hepatocellular carcinoma: a large-scale multicenter study. *Hepatobiliary Surg Nutr*. 2023;12(2):155–168. doi:10.21037/hbsn-21-288

11. Xu XF, Xing H, Han J, et al. Risk factors, patterns, and outcomes of late recurrence after liver resection for hepatocellular carcinoma: a multicenter study from China. *JAMA Surg*. 2019;154(3):209–217. doi:10.1001/jamasurg.2018.4334

12. Straś WA, Wasiak D, Łągiewska B, et al. Recurrence of hepatocellular carcinoma after liver transplantation: risk factors and predictive models. *Ann Transplant*. 2022;27:e934924.

13. Almqvist H, Crotty D, Nyren S, et al. Initial clinical images from a second-generation prototype silicon-based photon-counting computed tomography system. *Acad Radiol*. 2024;31(2):572–581. doi:10.1016/j.acra.2023.06.031

14. Koçak B, Durmaz E, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: a practical guide for beginners. *Diagn Interv Radiol*. 2019;25 (6):485–495. doi:10.5152/dir.2019.19321

15. Zhang X, Zhang Y, Zhang G, et al. Deep learning with radiomics for disease diagnosis and treatment: challenges and potential. *Front Oncol*. 2022;12:773840. doi:10.3389/fonc.2022.773840

16. Bera K, Braman N, Gupta A, Velcheti V, Madabhushi A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat Rev Clin Oncol*. 2022;19(2):132–146. doi:10.1038/s41571-021-00560-7

17. Xia TY, Zhou ZH, Meng XP, et al. Predicting microvascular invasion in hepatocellular carcinoma using CT-based radiomics model. *Radiology*. 2023;307(4):e222729. doi:10.1148/radiol.222729

18. Ji GW, Zhang YD, Zhang H, et al. Biliary tract cancer at ct: a radiomics-based model to predict lymph node metastasis and survival outcomes. *Radiology*. 2019;290(1):90–98. doi:10.1148/radiol.2018181408

19. Rezaeijo SM, Jafarpoor Nesheli S, Fatan Serj M, Tahmasebi Birgani MJ. Segmentation of the prostate, its zones, anterior fibromuscular stroma, and urethra on the MRIs and multimodality image fusion using U-Net model. *Quant Imaging Med Surg*. 2022;12(10):4786–4804. doi:10.21037/qims-22-115

20. Yao H, Tian L, Liu X, et al. Development and external validation of the multichannel deep learning model based on unenhanced CT for differentiating fat-poor angiomyolipoma from renal cell carcinoma: a two-center retrospective study. *J Cancer Res Clin Oncol*. 2023;149 (17):15827–15838. doi:10.1007/s00432-023-05339-0

21. Heo S, Park HJ, Lee SS. Prognostication of hepatocellular carcinoma using artificial intelligence. *Korean J Radiol*. 2024;25(6):550–558. doi:10.3348/kjr.2024.0070

22. Liu F, Liu D, Wang K, et al. Deep learning radiomics based on contrast-enhanced ultrasound might optimize curative treatments for very-early or early-stage hepatocellular carcinoma patients. *Liver Cancer*. 2020;9(4):397–413. doi:10.1159/000505694

23. Hectors SJ, Lewis S, Besa C, et al. MRI radiomics features predict immuno-oncological characteristics of hepatocellular carcinoma. *Eur Radiol*. 2020;30(7):3759–3769. doi:10.1007/s00330-020-06675-2

24. Kruthika KR, Rajeswari, Maheshappa HD, Maheshappa HD. Alzheimer's Disease Neuroimaging Initiative. CBIR system using Capsule Networks and 3D CNN for Alzheimer's disease diagnosis. *Inf Med Unlocked*. 2019;14:59–68. doi:10.1016/j.imu.2018.12.001

25. Singh SP, Wang L, Gupta S, Goli H, Padmanabhan P, Gulyás B. 3D deep learning on medical images: a review. *Sensors (Basel)*. 2020;20(18):5097. doi:10.3390/s20185097

26. Zheng X, Yao Z, Huang Y, et al. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. *Nat Commun*. 2020;11 (1):1236. doi:10.1038/s41467-020-15027-z

27. Tong T, Gu J, Xu D, et al. Deep learning radiomics based on contrast-enhanced ultrasound images for assisted diagnosis of pancreatic ductal adenocarcinoma and chronic pancreatitis. *BMC Med*. 2022;20(1):74. doi:10.1186/s12916-022-02258-8

28. Gu W, Chen Y, Zhu H, et al. Development and validation of CT-based radiomics deep learning signatures to predict lymph node metastasis in non-functional pancreatic neuroendocrine tumors: a multicohort study. *EClinicalMedicine*. 2023;65:102269. doi:10.1016/j.eclinm.2023.102269

29. Miranda J, Horvat N, Fonseca GM, et al. Current status and future perspectives of radiomics in hepatocellular carcinoma. *World J Gastroenterol*. 2023;29(1):43–60. doi:10.3748/wjg.v29.i1.43

30. Yao S, Ye Z, Wei Y, Jiang HY, Song B. Radiomics in hepatocellular carcinoma: a state-of-the-art review. *World J Gastrointest Oncol.* 2021;13 (11):1599–1615. doi:10.4251/wjgo.v13.i11.1599

31. La Greca Saint-Esteven A, Bogowicz M, Konukoglu E, et al. A 2.5D convolutional neural network for HPV prediction in advanced oropharyngeal cancer. *Comput Biol Med.* 2022;142:105215. doi:10.1016/j.compbiomed.2022.105215

32. Toseef M, Olayemi Petinrin O, Wang F, et al. Deep transfer learning for clinical decision-making based on high-throughput data: comprehensive survey with benchmark results. *Brief Bioinform.* 2023;24(4). doi:10.1093/bib/bbad254.

33. Xu H, Li C, Zhang L, Ding Z, Lu T, Hu H. Immunotherapy efficacy prediction through a feature re-calibrated 2.5D neural network. *Comput Methods Programs Biomed.* 2024;249:108135. doi:10.1016/j.cmpb.2024.108135

34. Liu L, Cai W, Tian H, et al. Ultrasound image-based nomogram combining clinical, radiomics, and deep transfer learning features for automatic classification of ovarian masses according to O-RADS. *Front Oncol.* 2024;14:1377489. doi:10.3389/fonc.2024.1377489

35. Feng B, Huang L, Liu Y, et al. A transfer learning radiomics nomogram for preoperative prediction of borrmann type iv gastric cancer from primary gastric lymphoma. *Front Oncol.* 2021;11:802205. doi:10.3389/fonc.2021.802205

36. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. *BMC Med Imaging.* 2022;22(1):69. doi:10.1186/s12880-022-00793-7

37. Zheng Q, Delingette H, Duchateau N, Ayache N. 3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. *IEEE Trans Med Imaging.* 2018;37(9):2137–2148. doi:10.1109/TMI.2018.2820742

38. Cai X, Zhang H, Wang Y, Zhang J, Li T. Digital pathology-based artificial intelligence models for differential diagnosis and prognosis of sporadic odontogenic keratocysts. *Int J Oral Sci.* 2024;16(1):16. doi:10.1038/s41368-024-00287-y

39. Li LT, Jiang G, Chen Q, Zheng JN. Ki67 is a promising molecular target in the diagnosis of cancer (review). *Mol Med Rep.* 2015;11(3):1566–1572. doi:10.3892/mmr.2014.2914

40. Wu M, Tan H, Gao F, et al. Predicting the grade of hepatocellular carcinoma based on non-contrast-enhanced MRI radiomics signature. *Eur Radiol.* 2019;29(6):2802–2811. doi:10.1007/s00330-018-5787-2

41. Mao B, Zhang L, Ning P, et al. Preoperative prediction for pathological grade of hepatocellular carcinoma via machine learning-based radiomics. *Eur Radiol.* 2020;30(12):6924–6932. doi:10.1007/s00330-020-07056-5

42. Martins-Filho SN, Paiva C, Azevedo RS, Alves VAF. Histological grading of hepatocellular carcinoma-a systematic review of literature. *Front Med Lausanne.* 2017;4:193. doi:10.3389/fmed.2017.00193

43. Minnema J, Wolff J, Koivisto J, et al. Comparison of convolutional neural network training strategies for cone-beam CT image segmentation. *Comput Methods Programs Biomed.* 2021;207:106192. doi:10.1016/j.cmpb.2021.106192