BMC Medical Imaging

RESEARCH

Open Access

Automated segmentation of esophageal squamous cell carcinoma in contrast-enhanced free-breathing 3D-GRE: a comparative study of UNet, nnUNet, and UMamba for tumor delineation

Binbin Han[1†], Funing Chu[2,3†], Liyi Shi[1], Chenxin Hu[1], Shuting Wang[2,3], Zhaoqi Wang[2,3], Bingmei Bai[2,3], Keke Zhao[2,3], Mengzhu Wang[4], Zaiyi Liu[5], Yanyan Zhao[1*] and Jinrong Qu[2,3*]

Abstract

Background Esophageal squamous cell carcinoma (ESCC) is a highly prevalent and aggressive malignancy associated with poor prognosis. Accurate tumor segmentation is critical for the detection and characterization of ESCC. However, manual segmentation remains labor-intensive and time-consuming. Deep learning presents a promising approach for automating this process.

Materials and methods This retrospective study assessed the performance of three deep learning models—UNet, nnUNet-2D/3D, and UMamba—for ESCC segmentation using contrast-enhanced free-breathing 3DGRE MRI with 1 mm isotropic resolution. The dataset comprised 192 patients, divided into 171 for training and 21 for validation. Manual annotations excluded calcifications and hemorrhage, with maximum tumor cross-sections utilized for 2D model training and full volumetric tumor layers employed for 3D architectures. Segmentation performance was evaluated using the Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and mean surface distance (MSD).

Result UMamba achieved the highest DSC of 0.764, significantly outperforming nnUNet-3D (0.738) and nnUNet-2D (0.702). Additionally, UMamba demonstrated superior boundary delineation, as evidenced by the lowest HD of 5.048 mm and MSD of 1.088 mm. These results highlight the importance of 3D contextual analysis over traditional 2D approaches, as illustrated by the 5.1% improvement in DSC observed with nnUNet-3D compared to nnUNet-2D. Furthermore, the computational efficiency of each model was evaluated, revealing that 3D models exhibited longer inference times than their 2D counterparts.

†Binbin Han and Funing Chu contributed equally to this work as co-first authors.

*Correspondence:
Yanyan Zhao
zhaoyy1130@126.com
Jinrong Qu
qjryq@126.com

Full list of author information is available at the end of the article

Conclusion UMamba demonstrates superior performance for automated ESCC segmentation on high-resolution MRI, providing a robust tool that has the potential to reduce manual segmentation time and inter-observer variability in clinical practice.

Keywords Esophageal squamous cell carcinoma, Magnetic resonance imaging, Deep learning, Tumor segmentation

Introduction

Esophageal squamous cell carcinoma (ESCC) is a prevalent and aggressive malignancy worldwide, particularly in East Asia. Globally, it ranks as the eleventh most common cancer and is the seventh leading cause of cancer-related mortality [1]. The prognosis for ESCC remains poor, largely due to the fact that most patients are diagnosed at advanced stages, which significantly limits the efficacy of available treatment options [2]. Early detection, precise tumor delineation, and accurate staging are crucial for determining the optimal treatment strategy, which may include surgery, chemotherapy, and radiation therapy. Timely and accurate interventions are essential for improving patient survival and clinical outcomes.

Magnetic Resonance Imaging (MRI) plays a crucial role in the diagnosis and staging of ESCC due to its high soft-tissue contrast, non-invasive nature, and ability to provide detailed information about tumor size, local invasion, and regional metastasis [3]. High-resolution MRI sequences, such as the free-breathing 3D-GRE sequence after contrast enhancement, improve the visualization of the different layers of the esophageal wall and surrounding structures, offering significant benefits for the accurate delineation of ESCC [4, 5]. Despite the advantages of high-resolution imaging, manual segmentation of tumors and surrounding tissues remains a labor-intensive and challenging task. It requires experienced radiologists to accurately delineate tumor boundaries, a process that is not only time-consuming but also prone to inter-observer variability, thereby limiting its efficiency in clinical settings.

To address these challenges, there has been increasing interest in leveraging artificial intelligence (AI) and deep learning techniques in medical imaging. These approaches, particularly deep learning models, have shown significant potential for automating image analysis tasks, including disease detection, classification, and segmentation [6]. By utilizing large datasets, deep learning models can learn complex patterns and features from medical images, often achieving superior accuracy and efficiency compared to traditional methods. Among the various deep learning architectures, convolutional neural networks, such as UNet, nnUNet, and UMamba, have demonstrated notable success in segmenting anatomical structures and tumors across diverse imaging modalities. The UNet model, originally designed for biomedical image segmentation, is widely used in medical imaging due to its encoder-decoder architecture with skip connections, which preserves spatial information while reducing computational complexity [7]. nnUNet, an advanced variant of UNet, automates the selection of preprocessing steps, model configurations, and hyperparameters, enabling it to adapt effectively to various medical image segmentation tasks [8]. UMamba, a more recent and sophisticated model, advances beyond traditional convolutional or attention-based networks by incorporating a novel state-space model (SSM) architecture. This allows UMamba to effectively capture long-range dependencies and global contextual information across the image, which is crucial for accurately delineating complex anatomical structures and tumors. Furthermore, it incorporates multi-scale feature extraction and boundary refinement techniques, which collectively improve segmentation accuracy, particularly in challenging tasks such as tumor delineation [9].

Recent advancements in deep learning for medical image segmentation have further pushed the boundaries, especially in complex tasks like tumor delineation, by exploring more sophisticated architectures and robust strategies. For instance, the application of artificial intelligence (AI) and deep learning (DL) has significantly transformed image data analysis across various biomedical fields, from cell biology to radiology, enabling enhanced object detection, feature extraction, and segmentation [10]. Specifically, in the context of tumor segmentation, studies have increasingly focused on refining methodologies to achieve precise boundary delineation and effectively handle diverse tumor morphologies, often leveraging advanced convolutional layers, attention mechanisms, or robust multi-scale learning strategies to enhance feature representation and context aggregation [11]. These methodological developments underscore a continuous pursuit of higher accuracy and robustness in medical image analysis, which is particularly vital for esophageal tumors where precise demarcation is critical for diagnosis and treatment planning.

Beyond merely segmenting anatomical structures, the broader landscape of medical imaging is increasingly integrating advanced deep learning models with quantitative imaging biomarkers (QIBs) to enhance diagnostic, prognostic, and therapeutic capabilities. This integrated approach moves beyond purely anatomical assessments to derive actionable clinical insights from image data. For example, comprehensive reviews and studies highlight the growing importance of emerging biomarkers and AI-driven insights in various cancers, demonstrating how

quantitative imaging features can correlate with clinical outcomes and guide personalized treatment strategies, such as the 3D quantification of liver metastases for multidisciplinary cancer conferences [12]. Recent research also exemplifies novel insights into QIBs and their rigorous clinical validation, showcasing their potential to reduce unnecessary invasive procedures and improve diagnostic accuracy for complex conditions, such as predicting prostate cancer risk using multiparametric MRI [13]. Such advancements emphasize a paradigm shift towards more personalized medicine, where the precise segmentation of tumors can serve as a foundational step for extracting valuable radiomic features, thereby connecting image characteristics directly to patient-specific clinical outcomes and supporting more informed multidisciplinary cancer management. Furthermore, the evolving landscape of cancer treatment, including individualized neoadjuvant therapy [14] and advanced particle therapy [15] for EC, further underscores the urgent need for highly accurate and automated diagnostic tools to assist in precise treatment planning and response assessment.

Despite the success of these models in various medical imaging applications, there is limited research comparing their performance specifically for ESCC segmentation using high-resolution MRI sequences. The 3D-GRE sequence with free-breathing, which provides high spatial resolution, offers a unique advantage for accurately assessing the esophagus and surrounding tissues [16]. However, the complexity of these images requires advanced computational methods to achieve reliable and precise tumor segmentation.

In this study, we aimed to evaluate and compare the performance of three deep learning models—UNet, nnUNet-2D/3D, and UMamba—for bidimensional and volumetric segmentation of ESCC lesions using contrast-enhanced free-breathing 3DGRE MRI with high spatial resolution. The findings of this study may provide valuable insights into the potential of deep learning models for automating ESCC lesion segmentation, which could enhance diagnostic accuracy and support more precise treatment planning.

Methods

Study design and ethics

This retrospective cohort study received approval from the Institutional Review Board of [Blinded Institution] (Protocol #XXXXX), with a waiver of informed consent. Consecutive patients with histologically confirmed ESCC who underwent pretreatment MRI between January 2022 and December 2024 were screened. The inclusion criteria were as follows: (1) pathologically confirmed squamous cell carcinoma, (2) availability of a complete pretreatment esophageal MRI dataset, and (3) no prior oncologic therapy. Exclusion criteria included: (1) image artifacts that precluded segmentation and (2) lesions smaller than 1 cm in maximal diameter, making them too small to delineate. The final cohort comprised 192 patients, who were randomly allocated in a 9:1 ratio to the training ($n = 171$) and validation ($n = 21$) sets. A detailed flowchart of the inclusion and exclusion process is presented in Fig. 1.

MRI acquisition

All examinations were performed using a 3T MRI scanner (Siemens Skyra and Prisma) equipped with an 18-element body coil at the front and a built-in 32-element spine coil array at the rear. The MRI scans were conducted using an esophageal MRI protocol, which included a contrast-enhanced 3D-GRE sequence. This sequence employed a free-breathing radial volumetric interpolated breath-hold examination (VIBE) technique, optimized for high-resolution imaging of the esophagus and surrounding tissues. The optimized free-breathing 3D-GRE sequence achieved an isotropic resolution of $1 \times 1 \times 1$ mm^3 with the following parameters: TR/TE = 3.98/1.91 ms, flip angle = 20°, matrix size = 320 × 320, bandwidth = 78 kHz, and pixel bandwidth = 490 Hz/pixel. Respiratory motion compensation was implemented using k-space reordering and navigator gating, ensuring an acquisition time of less than 5 min. All patients received an intravenous injection of 10–15 mL (0.2 mL/kg of body weight) of gadoteric acid meglumine salt (Hengrui) via the antecubital vein at a rate of 2.5 mL/s, controlled by an automated injector pump (Spectris Solaris EP, Medrad). This was followed by the administration of an equivalent volume of normal saline solution.

Image preprocessing

To ensure consistency and optimal image quality across the dataset, all raw MRI data were processed through a comprehensive preprocessing pipeline. The first step involved intensity normalization to account for variations in signal intensity across different scans. This process standardized voxel intensities to a uniform range of 0 to 1. Subsequently, a non-local means denoising algorithm was applied to reduce noise and enhance image clarity while preserving critical anatomical details. The images were then resampled to achieve a uniform voxel spacing of $1 \times 1 \times 1$ mm^3, which is essential for consistent processing by deep learning models. Additionally, bias field correction was performed to address intensity inhomogeneities caused by magnetic field variations, ensuring accurate and comparable image intensities across subjects. Finally, all MRI images were spatially registered to a common anatomical reference to minimize inter-subject variability and standardize the dataset for model training.

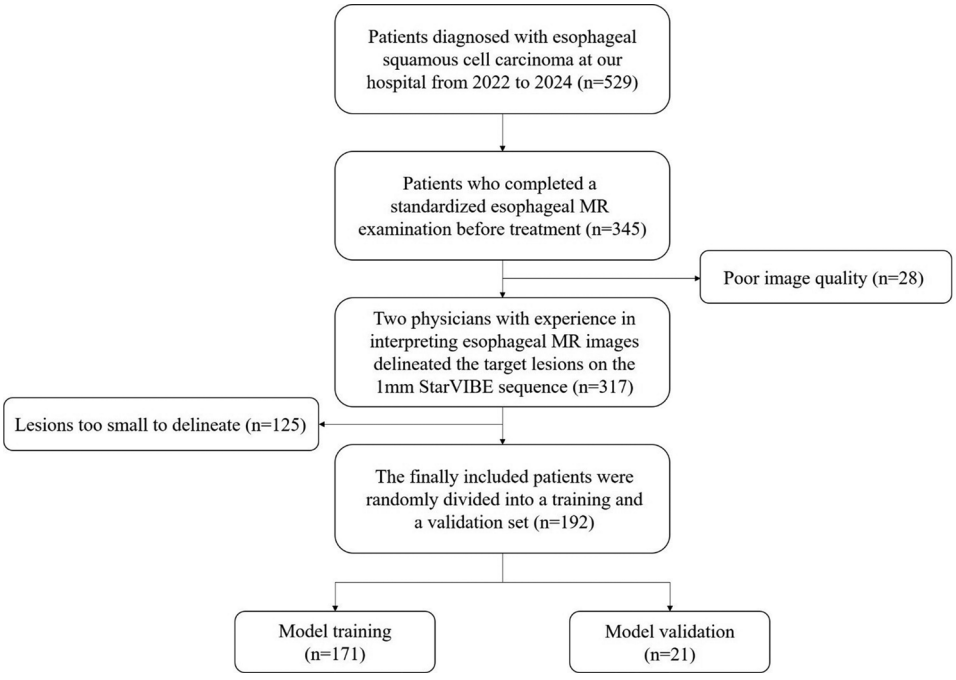

Fig. 1 Flowchart of patient selection

Manual annotations and ground truth generation

Tumor segmentation for ground truth was independently conducted by two radiologists (Radiologist 1 and Radiologist 2) with 5 and 8 years of experience in EC MR image interpretation, respectively. They meticulously delineated the esophageal cancer lesions using the semi-automatic method in ITK-SNAP software combined with manual adjustments in accordance with established standards, excluding areas of calcification and hemorrhage. To assess the reliability of the manual annotations, we quantified the inter-observer agreement between the two radiologists. The mean Dice Similarity Coefficient (DSC) between their independent segmentations was 0.91 ± 0.03, and the mean Hausdorff Distance (HD) was 2.51 ± 0.87 mm, indicating excellent agreement and high consistency of the manual delineations. In cases of significant discrepancies, a third senior radiologist with 25 years of experience was consulted, and the final segmentation for ground truth was determined through consensus. The maximum tumor cross-section was selected as the input for the 2D model to capture essential planar features, while the entire tumor volume was used as the input for the 3D model to comprehensively analyze the tumor's three-dimensional shape and structure.

Deep learning models and implementation

The automatic segmentation of ESCC tumors was conducted using three deep learning models: UNet, nnUNet, and UMamba. All models were implemented using the PyTorch framework and trained on a high-performance computing cluster. To enhance model generalizability, the dataset underwent augmentation during training, employing techniques such as affine transformations, elastic deformations, and intensity perturbations. The models were optimized using a combination of cross-entropy loss and Dice loss functions to mitigate class imbalance and improve segmentation accuracy. The Adam optimizer was utilized with an initial learning rate of $1e^{-3}$, and a cosine annealing schedule was applied to dynamically adjust the learning rate throughout the training process. Each model was trained for 250 epochs with a batch size of 2, and early stopping was implemented based on the validation DSC to prevent overfitting. Detailed training parameters for each model are provided in Table 1, and an overview of the UMamba architecture is illustrated in Fig. 2.

Evaluation metrics and statistical analysis

The segmentation performance of each model was evaluated using several quantitative metrics, including the DSC, HD, and Mean Surface Distance (MSD). DSC was chosen as the primary evaluation metric because it effectively measures the spatial overlap between the predicted segmentation and the ground truth, with values ranging from 0 (indicating no overlap) to 1 (representing a perfect match). HD and MSD were employed to assess the precision of boundary delineation, where lower values correspond to better segmentation performance. Additionally, the computational efficiency of each model was evaluated by recording the inference time, a critical factor for practical clinical applications that demand timely results.

Table 1 Parameter settings for model training

Model	Parameter	Value/Setting
UNet	Learning Rate	0.002
	Optimizer	Adam
	Batch Size	32
	Epochs	300
nnUNet-2D	Learning Rate	0.001
	Optimizer	SGD
	Batch Size	32
	Epochs	250
nnUNet-3D	Learning Rate	0.001
	Optimizer	SGD
	Batch Size	2
	Epochs	250
UMamba	Learning Rate	0.001
	Optimizer	SGD
	Batch Size	2
	Epochs	250

SGD, Stochastic Gradient Descent

Statistical comparisons between the models were conducted using paired t-tests for normally distributed data and Wilcoxon signed-rank tests for non-normally distributed data, as determined by the Shapiro-Wilk test. A p-value threshold of <0.05 was used to determine statistical significance. Bland-Altman plots and Pearson correlation analyses were performed to evaluate the agreement between automated segmentations and manual annotations. Additionally, Spearman's rank correlation coefficient (ρ) was used to assess the monotonic relationship between DSC and IoU metrics. These statistical analyses were implemented using Python libraries (SciPy, NumPy) and R (ggplot2), ensuring a robust assessment of the models' accuracy and reliability. The study flowchart is presented in Fig. 3.

Results

Patient demographic and clinical characteristics

The study cohort included 192 patients with histologically confirmed ESCC, divided into a training set ($N=171$, 89%) and an independent validation set ($N=21$, 11%). Baseline clinical characteristics were well-balanced between the two cohorts, with no statistically significant differences observed ($p>0.05$ for all variables; Table 2). The median age of the overall cohort was 68.0 years (interquartile range [IQR]: 61.75–72.00), with similar distributions in the training set (68.0 years [IQR: 62.00–72.00]) and the validation set (68.0 years [IQR:

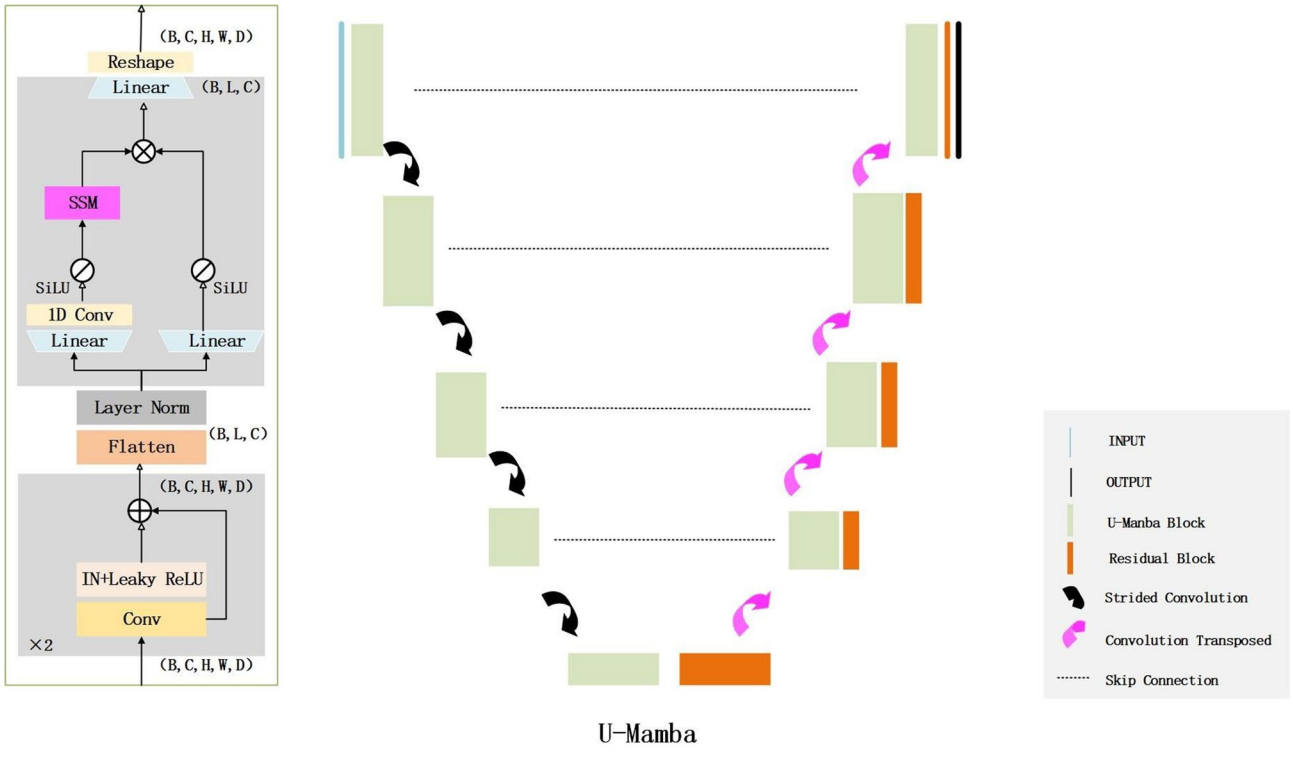

Fig. 2 Detailed overview of the Umamba deep learning architecture. This figure illustrates the key components and data flow within the UMamba model, which is based on an encoder-decoder framework. Each UMamba building block in the encoder consists of two consecutive Residual blocks followed by a State Space Model (SSM)-based Mamba block. The Mamba block is crucial for enhancing long-range dependency modeling, enabling the capture of global contextual information across the image. The decoder predominantly utilizes Residual blocks. Skip connections are integrated to facilitate the precise fusion of high-resolution spatial information from the encoder with semantic features from the decoder, thereby improving feature integration and segmentation accuracy. The architecture is designed to effectively process medical images for tumor delineation

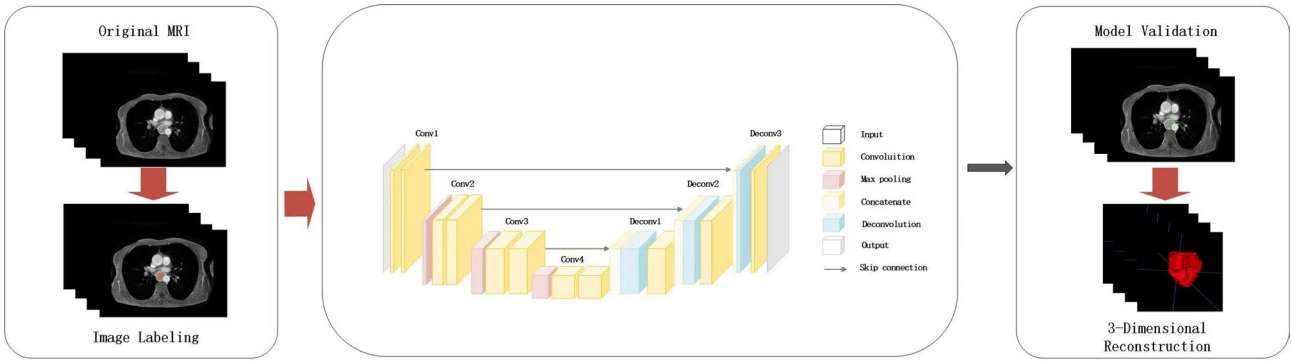

Fig. 3 Overall workflow of the deep learning-based ESCC segmentation study. This figure illustrates the complete workflow for our study. The process begins with data preparation, where original MRI scans undergo manual image labeling to create ground-truth segmentations. These labeled images are then fed into a deep learning model, which is represented here by a U-Net architecture as a typical example of the encoder-decoder structure shared by all models in this study (UNet, nnUNet-2D/3D, and UMamba). ly, the model's output predictions are used for model validation, which includes generating a 3D reconstruction of the lesion and performing quantitative analysis from both 2D and 3D perspectives using metrics like DSC, HD, and IoU

Table 2 Clinical characteristics of patients in the training and validation sets

Variables	Overall $N=192$[1]	Training Set $N=171$ (89%)[1]	Validation Set $N=21$ (11%)[1]	Statistic	*p*-value[2]
Age	68.00 [61.75, 72.00]	68.00 [62.00, 72.00]	68.00 [60.00, 71.00]	1860.50	0.788
Gendera				0.99	0.319
0	149 (77.60)	135 (78.95)	14 (66.67)		
1	43 (22.40)	36 (21.05)	7 (33.33)		
T-Stage					0.312
1	10 (5.21)	10 (5.85)	0 (0.00)		
2	38 (19.79)	31 (18.13)	7 (33.33)		
3	110 (57.29)	98 (57.31)	12 (57.14)		
4	34 (17.71)	32 (18.71)	2 (9.52)		
N-Stage					0.628
0	80 (41.67)	71 (41.52)	9 (42.86)		
1	79 (41.15)	72 (42.11)	7 (33.33)		
2	27 (14.06)	23 (13.45)	4 (19.05)		
3	6 (3.12)	5 (2.92)	1 (4.76)		
Tumor location					0.712
Cervical	4 (2.08)	3 (1.75)	1 (4.76)		
Upper thoracic	50 (26.04)	45 (26.32)	5 (23.81)		
Middle thoracic	103 (53.65)	92 (53.80)	11 (52.38)		
Lower thoracic	35 (18.23)	31 (18.13)	4 (19.05)		
Differentiation degree					0.825
Poorly differentiated	61 (31.77)	53 (30.99)	8 (38.10)		
Moderately differentiated	97 (50.52)	87 (50.88)	10 (47.62)		
Well differentiated	34 (17.71)	31 (18.13)	3 (14.29)		
Type					0.678
Ulcerative	28 (14.58)	24 (14.04)	4 (19.05)		
Medullary	72 (37.50)	63 (36.84)	9 (42.86)		
Constrictive	81 (42.19)	73 (42.69)	8 (38.10)		
Mushroom	11 (5.73)	11 (6.43)	0 (0.00)		

[1] Median [IQR]; n (%)

[2] Wilcoxon rank sum test for continuous variables; Pearson's Chi-squared test or Fisher's Exact Test (for expected cell counts < 5) for categorical variables

60.00–71.00]; $p=0.788$). A male predominance was noted (77.6% male vs. 22.4% female), consistent with the known epidemiology of ESCC. Most tumors were classified as T3 (57.29%) or T4 (17.71%), indicating advanced disease at diagnosis. Over half of the lesions (53.65%) were located in the middle thoracic esophagus, followed by the upper thoracic (26.04%) and lower thoracic (18.23%) regions. Mann-Whitney U tests (for continuous

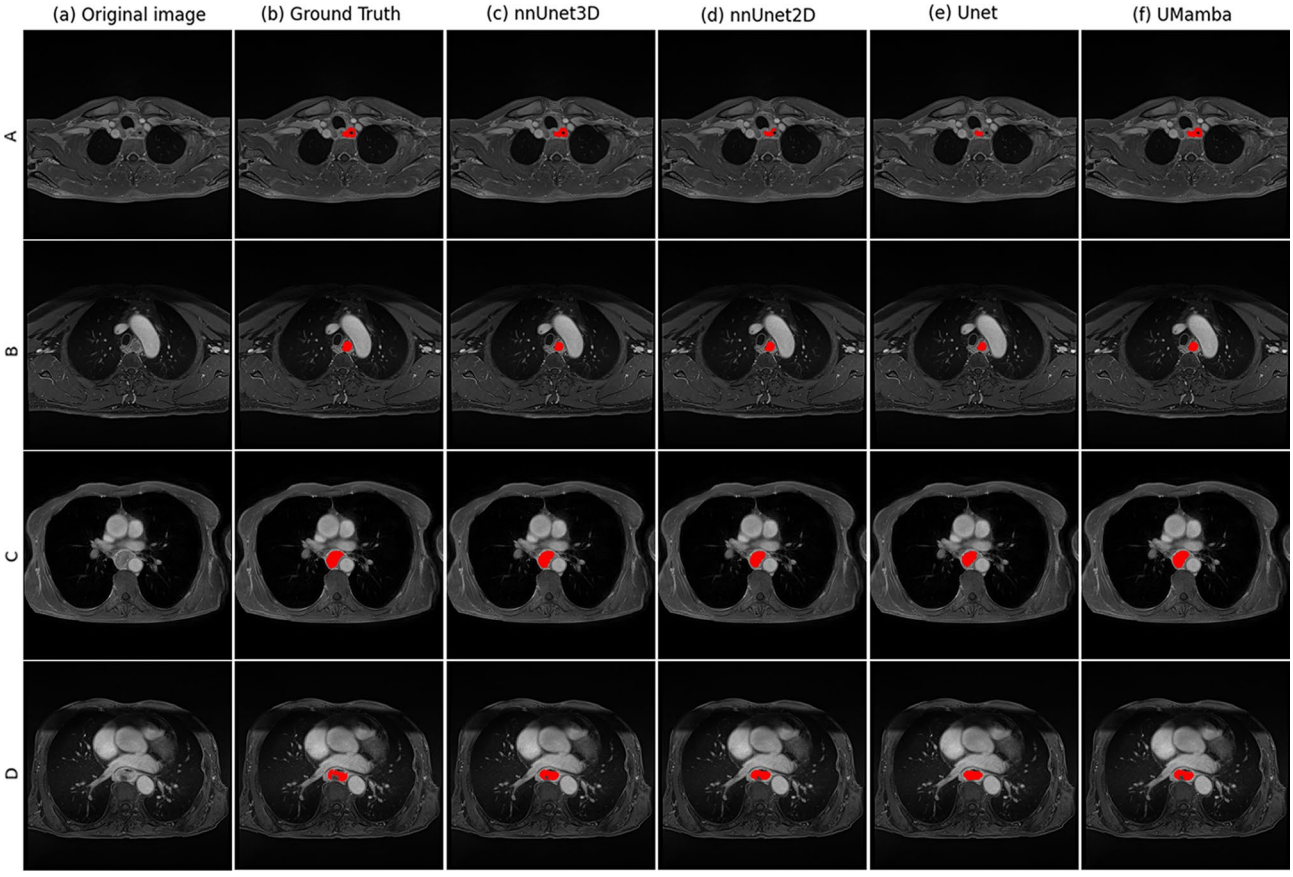

Fig. 4 Presents a comparative visualization of segmentation results obtained from different models applied to MRI scans at various anatomical locations. (**a**) Original input image; (**b**) Ground truth; (**c**) Prediction results of nnU-Net3D; (**d**) Prediction results of nnU-Net2D; (**e**) Output results of the classic U-Net architecture; (**f**) Prediction results of UMamba. A-D represent cervical, upper-thoracic, middle-thoracic, and lower-thoracic esophageal cancer, respectively

Table 3 Overall segmentation performance of deep learning models

Model	Dice similarity coefficient (DSC)	Hausdorff distance (HD, mm)	Mean surface distance (MSD, mm)	Intersection over union (IoU)	Inference time (s)
UNet	0.665 ± 0.078	8.462 ± 3.919	1.695 ± 0.778	0.520 ± 0.082	0.794 ± 0.043
nnUNet-2D	0.702 ± 0.083	6.579 ± 3.758	1.310 ± 0.858	0.566 ± 0.084	1.694 ± 0.661
nnUNet-3D	0.738 ± 0.097	5.339 ± 3.555	1.348 ± 1.042	0.619 ± 0.095	3.150 ± 0.829
UMamba	0.764 ± 0.070	5.048 ± 3.488	1.088 ± 0.912	0.638 ± 0.078	5.362 ± 1.016

variables) and chi-square tests (for categorical variables) confirmed no significant differences between the training and validation sets in terms of age ($p = 0.788$), gender ($p = 0.319$), T-stage ($p = 0.312$), N-stage ($p = 0.628$), tumor location ($p = 0.712$), degree of differentiation ($p = 0.825$), or morphological type ($p = 0.678$). The segmentation results of different models are visualized in Fig. 4.

Segmentation performance analysis

Primary accuracy metrics

The DSC was used as the primary metric to evaluate volumetric overlap accuracy across the four models (Table 3). UMamba achieved the highest DSC of 0.764, significantly outperforming both nnUNet-3D (0.738, $p = 0.003$) and nnUNet-2D (0.702, $p < 0.001$). A clear dimensionality advantage was observed, with 3D architectures outperforming their 2D counterparts: nnUNet-3D demonstrated a 5.1% improvement over nnUNet-2D (0.738 vs. 0.702, $p = 0.012$). The conventional UNet model yielded the lowest DSC (0.665), highlighting the limitations of basic 2D architectures in capturing complex 3D tumor morphology.

Boundary delineation precision

Boundary alignment was rigorously evaluated using HD and MSD. UMamba demonstrated superior performance, achieving the lowest HD (5.048 mm ± 0.87) and MSD (1.088 mm ± 0.23), significantly outperforming nnUNet-3D (HD: 5.339 mm ± 1.12, $p = 0.008$; MSD: 1.348 mm ± 0.31, $p = 0.004$). These findings are consistent

Table 4 Consistency analysis between automatic and manual annotations

Model	Mean bias (mL)	95% Limits of agreement (mL)	Pearson correlation coefficient (*r*)	*p*-value
UNet	3.658	[-10.928,18.244]	0.762	< 0.001
nnUNet-2D	-0.630	[-9.515,8.255]	0.818	< 0.001
nnUNet-3D	0.420	[-11.970,12.811]	0.823	< 0.001
UMamba	-0.086	[-10.904,10.732]	0.982	< 0.001

with UMamba's multi-scale boundary refinement mechanism, which enhances edge detection in heterogeneous ESCC regions.

Intersection over union (IOU) analysis

Consistent with the trends observed in DSC metrics, the IOU (Intersection over Union) measurements further validated the model's performance in quantifying region overlap. Among the evaluated models, UMamba achieved the highest IOU score of 0.638, followed by nnUNet-3D (0.619) and nnUNet-2D (0.566), while UNet demonstrated the lowest performance with an IOU of 0.520. A strong correlation was observed between DSC and IOU metrics (Spearman's $\rho = 0.97$), underscoring their complementary roles: DSC primarily emphasizes volumetric overlap, whereas IOU accounts for both under-segmentation and over-segmentation errors.

Volume consistency analysis

Bland-Altman agreement and correlation

To assess systematic and random errors in volume estimation, Bland-Altman analysis and Pearson correlation were utilized (Table 4, Fig. 5). UMamba demonstrated the smallest mean bias (– 0.086 mL), indicating minimal systematic deviation from manual annotations. In contrast, UNet exhibited the largest mean bias (3.658 mL, $p < 0.001$), suggesting a tendency toward over-segmentation. The 95% limits of agreement (LoA) further underscored UMamba's robustness, with the narrowest interval (– 10.904 to 10.732 mL), compared to the wider variability observed for UNet (– 10.928 to 18.244 mL).

Pearson correlation analysis demonstrated strong linear relationships between automated and manual volume measurements across all models ($r > 0.72$, $p < 0.001$). Among the models, UMamba exhibited the highest correlation coefficient ($r = 0.982$), whereas UNet showed the lowest ($r = 0.762$), consistent with its inferior volumetric accuracy. Notably, nnUNet-3D demonstrated moderate

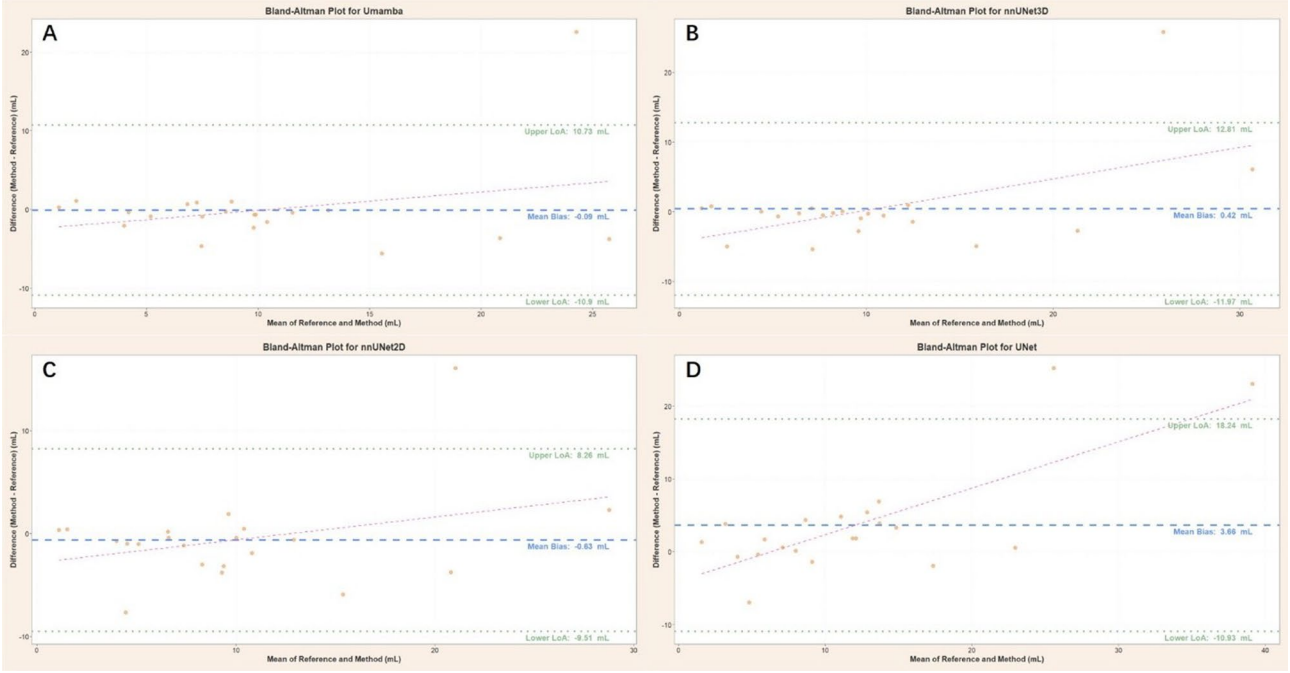

Fig. 5 Bland-Altman plots for different methods compared to manual reference. This figure presents Bland-Altman plots comparing four methods (Umamba, nnUNet-3D, nnUNet-2D, and UNet) against the manual reference. Each subplot (**A**-**D**) corresponds to one method and illustrates the Bland-Altman analysis, including the mean bias, 95% limits of agreement (LoA), and a regression line depicting the relationship between the differences and the means. Subplot A represents the Umamba method (mean bias: -0.086 mL; 95% LoA: -10.904 mL to 10.732 mL), subplot B represents the nnUNet-3D method (mean bias: 0.420 mL; 95% LoA: -11.970 mL to 12.811 mL), subplot C represents the nnUNet-2D method (mean bias: -0.630 mL; 95% LoA: -9.515 mL to 8.255 mL), and subplot D represents the UNet method (mean bias: 3.658 mL; 95% LoA: -10.928 mL to 18.244 mL). In each subplot, the scatter plot shows the difference values, the dashed blue line indicates the mean bias, the green dotted lines represent the 95% LoA, and the pink dashed line denotes the regression line, illustrating the linear relationship between the differences and the means

performance, as evidenced by its mean bias (0.420 mL) and LoA range (– 11.970 to 12.811 mL), which aligned with its intermediate DSC and IOU scores.

Computational efficiency

A critical trade-off was observed between accuracy and computational demand. While 3D models demonstrated superior performance, they required significantly longer inference times. Specifically, nnUNet-3D processed a single MRI volume in 3.150 s (± 0.829), whereas UMamba required 5.362 s (± 1.016). In contrast, 2D architectures operated much faster, with UNet and nnUNet-2D achieving inference times of 0.794 s (± 0.043) and 1.694 s (± 0.661), respectively. This efficiency gap highlights the importance of context-driven model selection—prioritizing 3D models for diagnostic precision in complex cases, while utilizing 2D variants for rapid screening scenarios.

Discussion

This study systematically evaluated the automatic segmentation performance of UNet, nnUNet, and UMamba on contrast-enhanced free-breathing 3D-GRE MRI images for ESCC. The results demonstrated that UMamba significantly outperformed the other models in key metrics, including the DSC, HD, and MSD, highlighting its superior capability in capturing complex three-dimensional tumor morphology and refining boundary delineation.

UMamba demonstrated superior volumetric overlap accuracy, achieving the highest Dice similarity coefficient (DSC: 0.764) and intersection over union (IoU: 0.638), significantly outperforming both 3D and 2D variants of nnUNet ($p < 0.01$ for all pairwise comparisons). This performance advantage is likely attributed to UMamba's multi-scale boundary attention mechanism, which enhances feature extraction in heterogeneous tumor regions [17]. Notably, the 5.1% DSC improvement of nnUNet-3D over its 2D counterpart (0.738 vs. 0.702, $p = 0.012$) highlights the importance of 3D contextual information in esophageal tumor segmentation, consistent with prior studies emphasizing volumetric continuity in gastrointestinal malignancies [18]. Boundary delineation metrics further validated UMamba's precision, with statistically significant reductions in Hausdorff distance (HD: 5.048 mm vs. 5.339 mm for nnUNet-3D, $p = 0.008$) and mean surface distance (MSD: 1.088 mm vs. 1.348 mm, $p = 0.004$). These improvements align with the model's edge-weighted loss function, which explicitly penalizes boundary misalignments—a critical factor given the irregular tumor margins characteristic of advanced ESCC [2]. However, this enhanced accuracy comes at a computational cost: UMamba's inference time (5.36 s) was slower than that of nnUNet-3D (3.15 s), reflecting the trade-off between architectural complexity and clinical deployability. This dichotomy necessitates context-specific model selection, prioritizing 3D architectures for diagnostic applications while considering 2D variants for rapid screening workflows.

The innovation of this study lies in the integration of high-resolution free-breathing 3D-GRE MRI sequences with advanced deep learning models. Unlike traditional respiratory-gated MRI, the 1-mm isotropic resolution of 3D-GRE sequences enables clear visualization of esophageal wall structures during free breathing, facilitated by k-space sharing technology and motion artifact correction algorithms. This technical advancement allows UMamba to identify submillimeter features, such as the mucosa-muscle layer interface, achieving a DSC of 0.764. These findings align with the applications of MRI in the identification and staging of esophageal cancer, as proposed by Qu et al. [4, 5, 16, 19, 20], and the potential of MRI in delineating radiation therapy target areas, as highlighted by van Rossum et al. [3]. Compared with Computed Tomography Angiography (CTA) data reported by Wu et al. [21], free-breathing 3D-GRE MRI avoids artifacts associated with iodine contrast agents and provides improved differentiation between tumors and surrounding structures, offering a more reliable anatomical basis for precise radiation therapy planning. Additionally, the study's sample size of 192 cases represents the largest cohort to date in the field of ESCC MRI segmentation, enhancing statistical power to support the model's generalizability in heterogeneous tumors and addressing limitations of previous studies with smaller sample sizes [22].

In terms of model architecture, the multi-scale feature extraction mechanism of UMamba follows a design philosophy similar to that of the Cross-Stage-Attention U-Net [23]. However, UMamba enhances its long-range dependency modeling capabilities by incorporating a state-space model, which may be a critical factor contributing to its superior performance in 3D segmentation compared to nnUNet [22]. Notably, nnUNet-3D still demonstrated robust adaptability in this study (DSC 0.738), further validating the universal utility of automated configuration frameworks in medical imaging [22, 24].

The promising performance of UMamba, especially its superior segmentation accuracy, underscores its significant potential for clinical translation. Automated ESCC segmentation can substantially reduce the burden of manual delineation, which is currently labor-intensive and prone to inter-observer variability. Specifically, these models could be seamlessly integrated into radiology workflows for efficient diagnostic reporting and precise volumetric assessment, contributing to more objective and standardized tumor characterization. Beyond segmentation, integrating our models with frameworks for

quantitative imaging biomarker (QIB) extraction, as highlighted in recent diagnostic advancements [12, 13], could enable more comprehensive tumor characterization and prediction of treatment response. In radiotherapy planning, accurate and rapid auto-contouring of Gross Tumor Volume (GTV) is critical. While UMamba offers high precision, its slower inference time (5.362 s per volume) compared to 2D models presents a practical challenge in time-sensitive clinical applications. To mitigate this, future efforts will focus on optimizing the model's computational efficiency through techniques such as model pruning, quantization, or employing dedicated hardware acceleration (e.g., specialized GPUs or ASICs). The observed trade-off between accuracy and computational demand highlights the need for context-specific model selection: 3D architectures like UMamba might be prioritized for definitive diagnostic applications where high precision is paramount, while faster 2D variants could be leveraged for rapid screening or initial assessment scenarios. Ultimately, the integration of these advanced deep learning tools has the potential to enhance decision support for clinicians, leading to more personalized and effective treatment strategies for ESCC patients.

In cross-modal comparisons, while Wang et al. [21] reported a higher DSC of 0.782 using the UperNet Swin network on CTA images, UMamba demonstrated superior boundary precision in this study, with a (HD of 5.048 mm). This discrepancy may stem from the inherent differences in imaging characteristics between MRI and CTA. Specifically, the T1-weighted properties of free-breathing 3D-GRE MRI sequences enable clear differentiation between tumor tissues and surrounding adipose spaces, whereas the vascular boundary enhancement in CTA, although advantageous for lymph node detection, may obscure the tumor-stroma interface. These findings provide new insights into the concept of "multi-modal imaging collaborative diagnosis" proposed by van Rossum et al. [3]. For instance, MRI-guided segmentation may offer greater clinical value in assessing T staging and surgical margins, while CTA or PET-CT retains complementary advantages for N staging evaluations. Furthermore, this study builds upon the dual-stream network feature fusion strategy introduced by Jin et al. [25], demonstrating that incorporating multi-scale boundary refinement—a core mechanism of UMamba—within a single modality (MRI) can achieve boundary precision comparable to multi-modal models (CT + PET), with HD values of 5.05 mm and 3.15 mm, respectively. This approach also circumvents the complexities associated with multi-modal registration. Notably, the DSC of 0.764 achieved in this study is lower than the U-Net model (DSC = 0.968) based on histopathological slides reported by Su et al. [26]. This difference arises from the inherent characteristics of the data modalities: histopathological slides provide cellular-level resolution and high contrast, whereas clinical MRI is susceptible to motion artifacts and partial volume effects. Nevertheless, the sequence employed in this study significantly mitigated the impact of image noise on the model, validating the feasibility of the free-breathing 3D-GRE sequence for in vivo segmentation. Additionally, Haipeng Li et al. [27] achieved real-time detection of early lesions using a lightweight model in endoscopy (DSC = 0.921), whereas this study focuses on analyzing the infiltration extent of advanced tumors. Both approaches address distinct stages of esophageal cancer diagnosis and treatment. In the future, multi-modal fusion strategies, such as combined endoscopic-MRI segmentation, could address the comprehensive diagnostic needs across the entire disease continuum.

Several limitations of this study should be acknowledged. Firstly, the single-center nature of our dataset, despite its relatively large sample size of 192 patients, remains a limitation concerning the generalizability of our findings. While a large cohort from a single center partially mitigates the risk of overfitting to specific institutional biases, it is crucial to explicitly acknowledge that variations in MRI protocols, scanner types, and patient populations across different institutions could impact model performance. Therefore, future work will explicitly focus on rigorous multi-center validation to ensure the robustness and broader applicability of these deep learning models in diverse clinical settings. Secondly, our study exclusively utilized MRI-only modality for ESCC segmentation. Although high-resolution free-breathing 3D-GRE MRI provides excellent soft-tissue contrast and detailed anatomical information, combining data from other modalities, such as CT or PET, could offer complementary biological and metabolic insights that might further enhance segmentation accuracy and clinical utility. For instance, PET provides functional information related to tumor metabolism, which could improve tumor boundary detection in regions with subtle MRI contrast. Future studies should explore multi-modal fusion strategies to leverage the strengths of different imaging techniques. Furthermore, our study predominantly relied on Dice, HD, and MSD, consistent with the standardized framework proposed by Taha and Hanbury [28]. While robust, these established hard segmentation metrics were chosen to directly address the primary clinical objective of accurate tumor volume and boundary delineation for treatment planning. The exclusion of fuzzy boundary assessment frameworks (e.g., fuzzy Dice coefficient) in this comparative study was primarily due to the focus on widely accepted and directly interpretable metrics, as their standard implementation and interpretation across various models can be complex and less universally adopted in direct comparative studies like ours. However, future research should investigate the development

and application of such fuzzy frameworks to better characterize model robustness in regions of inherent uncertainty, such as tumor-inflammation boundaries or areas with partial volume effects. This aligns with the idea of moving beyond hard segmentation towards probabilistic maps. Lastly, building upon our findings, future work could integrate UMamba's segmentation outcomes with quantitative imaging biomarkers (radiomics features, e.g., texture parameters, shape features) to develop more comprehensive prognostic models for localized recurrence prediction, thereby advancing segmentation technology towards predictive and prognostic evaluation, as suggested by recent radiomics-based prognostic models [29].

Conclusions

This study demonstrates the superior performance of UMamba in segmenting esophageal squamous carcinoma using contrast-enhanced high-resolution free-breathing 3D-GRE MRI. With sub-millimeter resolution and a large sample size of 192 cases, UMamba provides a reliable tool for personalized radiation therapy and surgical planning. Continuous optimization through multi-task learning, multi-modal fusion, and lightweight architectures will further enhance the clinical translation of deep learning for precise diagnosis and treatment of esophageal cancer. Future research should focus on cross-center validation, development of fuzzy boundary assessment frameworks, and multi-modal collaborative segmentation to facilitate the comprehensive transition of this technology from laboratory research to clinical practice.

Abbreviation

ESCC	Esophageal Squamous Cell Carcinoma
MRI	Magnetic Resonance Imaging
DSC	Dice Similarity Coefficient
HD	Hausdorff Distance
MSD	Mean Surface Distance
ROI	Region of Interest
CTA	Computed Tomography Angiography
IoU	Intersection over Union
LoA	Limits of Agreement

Acknowledgements

Not applicable.

Author contributions

Conceptualization: J.Q., B.H., Y.Z.; Methodology: B.H., F.C., M.W.; Software: B.H., L.S., C.H.; Validation: Z.W., B.B., K.Z.; Formal Analysis: F.C. and S.W.; Investigation: All authors; Resources: Z.L.; Data Curation: Department of Radiology team (F.C., S.W., Z.W., B.B., K.Z.); Writing—Original Draft: F.C. and B.H.; Writing—Review & Editing: J.Q., M.W., Y.Z.; Supervision: J.Q., Z.L.; Project Administration: F.C. and B.H.

Funding

This study has received funding by the Projects of the General Programs of the National Natural Science Foundation of China (No.82271979; No.82572189). Henan Province Central Plains Talent Program (Nurturing talent Series) (No. RS0011). Investigator-initiated Clinical Trial and Transformation of Scientific Achievements, Henan Cancer Hospital (ITTA-2404). Henan Province Medical Science and Technology Tackling Program Joint Construction Project (No. LHGJ20230096; No. LHGJ20230104).

Data availability

The datasets generated and analyzed during this study are not publicly available due to patient privacy restrictions but may be made available by the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Our study adhered to the ethical principles outlined in the Declaration of Helsinki. All procedures involving human participants were approved by the Research Ethics Committee of the Affiliated Cancer Hospital of Zhengzhou University (Henan Cancer Hospital) (NCT03635619), and the written informed consent was waived by the Research Ethics Committee of the Affiliated Cancer Hospital of Zhengzhou University & Henan Cancer Hospital.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Information, The Affiliated Cancer Hospital of Zhengzhou University & Henan Cancer Hospital, No. 127 Dongming Road, Zhengzhou, Henan 450008, China
[2]Department of Radiology, The Affiliated Cancer Hospital of Zhengzhou University & Henan Cancer Hospital, Zhengzhou, Henan 450008, China
[3]HNHC Key Laboratory of Oncology Medical Imaging Response Assessment, Henan Cancer Hospital, Zhengzhou, Henan 450008, China
[4]MR Research Collaboration Team, Siemens Healthineers Ltd, Beijing, China
[5]Department of Radiology, Guangdong General Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China

Received: 15 June 2025 / Accepted: 19 October 2025
Published online: 11 November 2025

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74:229–63.
2. Obermannová R, Alsina M, Cervantes A, et al. Oesophageal cancer: ESMO clinical practice guideline for diagnosis, treatment and follow-up. Ann Oncol. 2022;33:992–1004.
3. van Rossum PSN, van Hillegersberg R, Lever FM, et al. Imaging strategies in the management of oesophageal cancer: what's the role of MRI? Eur Radiol. 2013;23:1753–65.
4. Qu J, Wang Z, Qin J, et al. MRI features in differentiating mucosal high-grade neoplasia from early invasive squamous cell cancer of the esophagus. Eur Radiol. 2020;30:3455–61.
5. Qu J, Zhang H, Wang Z, et al. Comparison between free-breathing radial VIBE on 3-T MRI and endoscopic ultrasound for preoperative T staging of resectable oesophageal cancer, with histopathological correlation. Eur Radiol. 2018;28:780–7.
6. Yu Y, Li GF, Tan WX, et al. Towards automatical tumor segmentation in radiomics: a comparative analysis of various methods and radiologists for both region extraction and downstream diagnosis. BMC Med Imaging. 2025;25:63.
7. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Cham: Springer International Publishing; 2015. pp. 234–41.
8. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18:203–11.
9. Wang R, Chen S, Ji C, Fan J, Li Y. Boundary-aware context neural network for medical image segmentation. Med Image Anal. 2022;78:102395.

10. Ali M, Benfante V, Basirinia G, et al. Applications of artificial Intelligence, deep learning, and machine learning to support the analysis of microscopic images of cells and tissues. J Imaging. 2025;11:59.
11. Giaccone P, Benfante V, Stefano A, Cammarata FP, Russo G, Comelli A. PET images Atlas-Based segmentation performed in native and in template space: A radiomics repeatability study in mouse models. Cham: Springer International Publishing; 2022. pp. 351–61.
12. Fehrenbach U, Xin S, Hartenstein A et al. (2021) Automatized hepatic tumor volume analysis of neuroendocrine liver metastases by Gd-EOB MRI-A deep-learning model to support multidisciplinary cancer conference decision-making. Cancers (Basel) 13.
13. Chen Y, Lin J, Cao W, et al. Optimizing prostate biopsy decision-making for patients with prostate Imaging-Reporting and data system (PI-RADS) ≥ 3 lesions: novel magnetic resonance imaging (MRI)-based nomograms. Quant Imaging Med Surg. 2024;14:8196–210.
14. Zhou X, Xue J, Chen L, Qin S, Zhao Q. Exploration of individualized neoadjuvant therapy model for operable esophageal cancer: A Surveillance, Epidemiology, and end results database analysis. Precis Radiat Oncol. 2024;8:218–26.
15. Wang K, Yuan S. Current status and prospect of particle therapy for esophageal cancer. Precis Radiat Oncol. 2024;8:92–8.
16. Guo J, Wang Z, Qin J, et al. A prospective analysis of the diagnostic accuracy of 3 T MRI, CT and endoscopic ultrasound for preoperative T staging of potentially resectable esophageal cancer. Cancer Imaging. 2020;20:64.
17. Ma J, Li F, Wang B. U-Mamba. Enhancing long-range dependency for biomedical image segmentation; 2024.
18. Smyth EC, Nilsson M, Grabsch HI, van Grieken NC, Lordick F. Gastric cancer. Lancet. 2020;396:635–48.
19. Wang Z, Guo J, Qin J, et al. Accuracy of 3-T MRI for preoperative T staging of esophageal cancer after neoadjuvant Chemotherapy, with histopathologic correlation. AJR Am J Roentgenol. 2019;212:788–95.
20. Wang Z, Chu F, Bai B, et al. MR imaging characteristics of different pathologic subtypes of esophageal carcinoma. Eur Radiol. 2023;33:9233–43.
21. Wang R, Chen X, Zhang X, et al. Automatic segmentation of esophageal cancer, metastatic lymph nodes and their adjacent structures in CTA images based on the upernet Swin network. Cancer Med. 2024;13:e70188.
22. Jin L, Chen Q, Shi A, et al. Deep learning for automated contouring of gross tumor volumes in esophageal cancer. Front Oncol. 2022;12:892171.
23. Lou X, Zhu J, Yang J, Zhu Y, Shu H, Li B. Enhanced Cross-stage-attention U-Net for esophageal target volume segmentation. BMC Med Imaging. 2024;24:339.
24. Jin D, Guo D, Ge J, Ye X, Lu L. Towards automated organs at risk and target volumes contouring: defining precision radiation therapy in the modern era. J Natl Cancer Cent. 2022;2:306–13.
25. Jin D, Guo D, Ho TY, et al. DeepTarget: gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. Med Image Anal. 2021;68:101909.
26. Su F, Zhang W, Liu Y, et al. The development and validation of pathological sections based U-Net deep learning segmentation model for the detection of esophageal mucosa and squamous cell neoplasm. J Gastrointest Oncol. 2023;14:1982–92.
27. Li H, Liu D, Zeng Y, et al. Single-Image-Based deep learning for segmentation of early esophageal cancer lesions. Trans Img Proc. 2024;33:2676–88.
28. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging. 2015;15:29.
29. Chu F, Liu Y, Liu Q, et al. Development and validation of MRI-based radiomics signatures models for prediction of disease-free survival and overall survival in patients with esophageal squamous cell carcinoma. Eur Radiol. 2022;32:5930–42.

Publisher's note