# RadioFormer: Integrating Radiologist Inductive Bias for Tumor Classification on Multi-Sequence MR Images

Xiaoyu Bai[1,2] and Yong Xia[1,2,3(✉)]

[1] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China
[2] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[3] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China
yxia@nwpu.edu.cn

**Abstract.** Multi-sequence magnetic resonance imaging (MRI) plays a critical role in tumor diagnosis but relies heavily on manual interpretation, which is both labor-intensive and dependent on expert knowledge. While deep learning-based diagnostic methods show significant potential, they typically require large datasets for effective training. However, the high cost of data collection and annotation often limits the available dataset size. This highlights the need for models that can effectively train on small datasets, mitigate overfitting, and achieve reliable performance. To address these challenges, we propose RadioFormer, a novel model that incorporates radiologist inductive bias to facilitate efficient learning on small MRI datasets. Unlike traditional 2D or 3D architectures, RadioFormer emulates the radiologist's diagnostic process by explicitly parsing MRI data into three hierarchical levels: (1) single-sequence slice feature extraction, (2) multi-sequence slice information aggregation, and (3) inter-slice information (volume) aggregation. Each level builds upon the previous one, ensuring smooth information flow and a hierarchical understanding of lesion characteristics. By integrating expert knowledge into its design, RadioFormer effectively leverages inductive bias to enhance model generalization on small datasets. We evaluated RadioFormer on three public datasets for brain, breast, and liver tumor classification, where it achieved state-of-the-art performance across all tasks. The code and pre-processed data for RadioFormer are available at https://github.com/aa1234241/RadioFormer/tree/master.

**Keywords:** Tumor classification · Multi-sequence MRI · Inductive bias.

## 1 Introduction

Cancer remains one of the leading causes of mortality worldwide, accounting for millions of deaths each year [6]. Magnetic resonance imaging (MRI) plays an

* Corresponding author: Y. Xia.

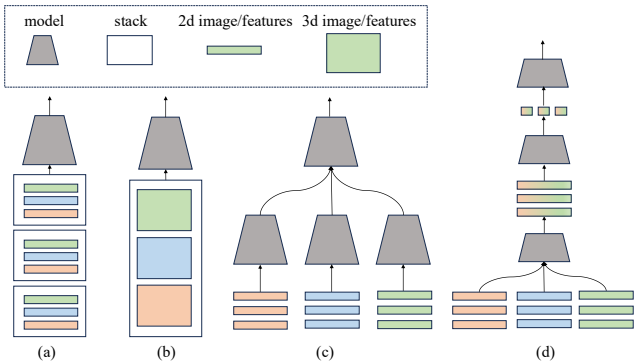


Fig. 1: Illustration of common tumor classification models: (a) The 2D model, which stacks slices from each MRI sequence as input; (b) The 3D model, which stacks all sequences together as input; (c) The mid-fusion model, which first extracts features from each sequence and then fuses them in the middle stage of the model; (d) Our RadioFormer, which divides feature extraction and fusion into three stages, inspired by the diagnostic practice of radiologists.

important role in tumor detection, characterization, and treatment planning. In particular, multi-sequence MRI provides a more detailed view of tissue properties by capturing images at different time points after contrast agent injection or across different imaging protocols. However, accurately interpreting multi-sequence MRI remains challenging, requiring extensive expertise and experience from radiologists. Additionally, manual diagnosis is time-consuming and subject to inter-observer variability, highlighting the need for automated diagnostic models to assist clinicians in tumor classification.

The advent of deep learning has catalyzed significant progress in image-based tumor diagnosis. While current methods primarily focus on specific organs, such as the liver [28,32,29,25,27], lung [1,19], kidney [26,8], breast [30,31], and brain [3,22], they generally follow similar design strategies. These methods typically process input 3D images in one of three ways: (a) as 2D slices, using 2D models for classification [18,11,20]; (b) by leveraging 3D models for direct classification [17,23]; or (c) by first extracting features from each sequence and fusing them at an intermediate stage of the model [16,27,25], as shown in Fig. 1(a)(b)(c). The 2D approach often utilizes pre-trained models from ImageNet to enhance performance. For instance, Swati et al. used the pre-trained VGG-16 model for brain tumor classification [20]. While this approach is straightforward and effective, it lacks a holistic interpretation of the 3D tumor volume, which can limit its ability to capture complex, inter-slice tumor characteristics. In contrast, the 3D approach classifies the entire 3D volume directly but suffers from the challenge of limited dataset size. Tumor datasets typically consist of only hundreds to a few thousand cases, making them prone to overfitting and resulting in suboptimal performance. Some researchers attempt to address this limitation by using

pre-trained 3D models from video data [16,14], but the significant domain and semantic gap between medical imaging and video data often lead to unsatisfactory results. Other approaches utilize customized models to better extract and fuse features from different sequences. For example, Wang et al. proposed a method called TransLiver [25], which first uses independent Pyramid Vision Transformers (PVTs) [24] for each sequence image to extract sequence-specific features and then inputs them into a fusion module to obtain the final feature representation. While this approach improves feature fusion, it still faces the challenge of limited data size.

In this study, we introduce RadioFormer, a radiologist-inspired hierarchical Transformer model designed to effectively learn from modestly sized tumor datasets. RadioFormer departs from traditional 2D or 3D model designs by incorporating a hybrid 2D-3D data flow that mirrors the inductive bias of radiologists. Specifically, the architecture of RadioFormer is informed by the diagnostic practice of radiologists, who typically begin by inspecting each sequence image slice-by-slice, identifying key slices containing valuable diagnostic clues. These key slices are then collectively reviewed across sequences to reach a comprehensive diagnosis. RadioFormer mirrors this process through tripartite levels: the single-sequence slice information extraction level, the multi-sequence slice information aggregation level, and the inter-slice (volume) information aggregation level. At each level, RadioFormer employs pure vision Transformer blocks to encode information. The output of each level serves as the input tokens for the subsequent level, enabling a seamless flow of information and a hierarchical understanding of the lesion characteristics.

Our RadioFormer is examined on three public datasets: the LLD-MMRI liver tumor dataset [16] for classifying seven types of liver tumors, the Advanced-MRI-Breast-Lesions dataset [9] for benign and malignant breast tumor classification, and the ReMIND dataset [13] for classifying three types of brain tumors. The experimental results demonstrate that our RadioFormer achieves the best performance across all three datasets, highlighting its superior capability. The code and pre-processed data for our method are made publicly available.

## 2 Method

Figure 2 illustrates the image pre-processing step and overall architecture of the proposed RadioFormer model. In the data pre-processing step, given the original, unregistered multi-sequence MRI images, we employ the open-source UAE-M method [4] to register all sequence images to one sequence. Subsequently, the aligned multi-sequence tumor volumes are cropped from the registered images. These cropped tumor volumes then serve as the input to our RadioFormer model.

RadioFormer is a tripartite-level Transformer-based model designed to fully harness the diagnostic cues from volumetric data. The first level focuses on feature extraction from individual sequence slices. The second level aggregates information from aligned single-sequence slices, processing and consolidating it into a series of multi-sequence slice tokens. Finally, the third level fuses the multi-

sequence slice tokens to produce the final classification result. In the following subsections, we will delve into each level in greater detail.

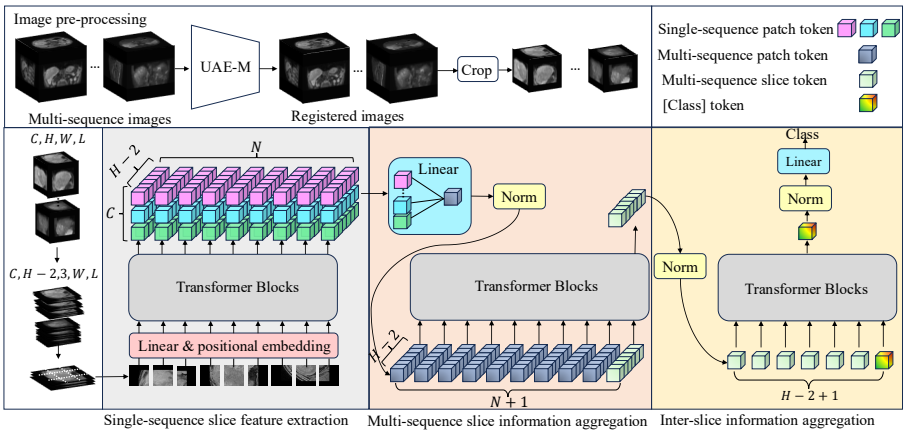


Fig. 2: The overall architecture of the proposed RadioFormer model.

### 2.1 Single-sequence Slice Feature Extraction

Radiologists typically interpret MR images by first examining them slice by slice, looking for important diagnostic clues for further investigation. Our RadioFormer model emulates this approach by first extracting features from each sequence slice independently. Given a multi-sequence tumor volume with dimensions $(C, H, W, L)$, where $C$ denotes the number of sequences, and $H, W, L$ represents the height, width, and length of the tumor volume, respectively. We first reorganize the volume into an array of 2D transverse images. Each transverse image contains three consecutive slices and has a shape of $(3, W, L)$.This reorganization results in an array of images with dimensions $(C, H-2, 3, W, L)$.

Subsequently, we divide each transverse image into patches of size $(P, P)$ and flatten them, resulting in a sequence of vectors $[\mathbf{x}_{c,h}^1; \mathbf{x}_{c,h}^2; \cdots; \mathbf{x}_{c,h}^N]$, where the length of this sequence is denoted as $N = \frac{W}{P} \times \frac{L}{P}$. These vectors are then linearly embedded, and positional embeddings are added, as shown in the following equation:

$$\mathbf{z}_{c,h}^{(0)} = \left[\mathbf{x}_{c,h}^1 \mathbf{E}_\mathbf{s}; \mathbf{x}_{c,h}^2 \mathbf{E}_\mathbf{s}; \cdots; \mathbf{x}_{c,h}^N \mathbf{E}_\mathbf{s}\right] + \mathbf{E}_{\text{pos}}, \tag{1}$$

where $c \in [1, C]$, $h \in [1, H-2]$, $D$ denotes the feature length, $\mathbf{E}_\text{s} \in \mathbb{R}^{\left(P^2 \cdot 3\right) \times D}$ is the linear embedding layer and $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{N \times D}$ is the positional embedding. Each $\mathbf{z}_{c,h}^{(0)}$ is then input into a series of ViT [10] Transformer encoder blocks. These blocks consist of alternating layers of multiheaded self-attention (MSA) and

multi-layer perception (MLP) modules, with Layer Norm (LN) applied before each module. The mathematical representation is as follows:

$$\begin{aligned} \mathbf{z}'_{c,h} &= \mathrm{MSA}\left(\mathrm{LN}\left(\mathbf{z}_{c,h}^{(\ell-1)}\right)\right) + \mathbf{z}_{c,h}^{(\ell-1)}, \\ \mathbf{z}_{c,h}^{(\ell)} &= \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{z}'_{c,h}\right)\right) + \mathbf{z}'_{c,h}, \end{aligned} \tag{2}$$

where $\ell \in [1, L_{\mathrm{s}}]$, $L_{\mathrm{s}}$ is the number of the Transformer blocks in this level. The output of the final Transformer block $\mathbf{z}_{c,h}^{(L_{\mathrm{s}})} = [\mathbf{y}_{c,h}^1; \mathbf{y}_{c,h}^2; \cdots; \mathbf{y}_{c,h}^N]$ are regarded as single-phase patch tokens. These tokens will serve as the input for the subsequent multi-phase slice information aggregation level.

### 2.2 Multi-sequence Slice Information Aggregation

The objective of this level is to integrate the features extracted from single-phase slices into a higher-level representation that captures multi-sequence information. We begin with the spatially corresponding single-sequence patch tokens $[\mathbf{y}_{1,h}^i; \mathbf{y}_{2,h}^i; \cdots; \mathbf{y}_{C,h}^i]$. A linear layer $\mathbf{E}_{\mathrm{m}} \in \mathbb{R}^{CD \times D}$ is utilized to project the concatenated tokens into a lower-dimensional multi-sequence patch token, as defined by:

$$\mathbf{m}_h^i = \mathrm{LN}(\mathrm{Concat}[\mathbf{y}_{1,h}^i; \mathbf{y}_{2,h}^i; \cdots; \mathbf{y}_{C,h}^i]\mathbf{E}_{\mathrm{m}}). \tag{3}$$

Each multi-sequence slice is then represented by a series of these multi-sequence patch tokens $[\mathbf{m}_h^1; \mathbf{m}_h^2; \cdots; \mathbf{m}_h^N]$. Building upon the previous level, we process these multi-sequence patch tokens using a stack of Transformer encoder blocks. To aggregate patch-wise information into a slice-wise representation, we introduce a learnable multi-sequence slice token $\mathbf{s}$ to the patch tokens $[\mathbf{m}_h^1; \cdots; \mathbf{m}_h^N; \mathbf{s}]$. The fusion of multi-phase slice information is encapsulated in the following equations:

$$\begin{aligned} \mathbf{v}_h^{(0)} &= \left[\mathbf{m}_h^1; \mathbf{m}_h^2; \cdots; \mathbf{m}_h^N; \mathbf{s}\right], \\ \mathbf{v}'_h &= \mathrm{MSA}\left(\mathrm{LN}\left(\mathbf{v}_h^{(\jmath-1)}\right)\right) + \mathbf{v}_h^{(\jmath-1)}, \\ \mathbf{v}_h^{(\jmath)} &= \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{v}'_h\right)\right) + \mathbf{v}'_h, \end{aligned} \tag{4}$$

where $\jmath \in [1, L_{\mathrm{m}}]$, and $L_{\mathrm{m}}$ denotes the number of Transformer blocks in this level. The output from the final Transformer block retains only the multi-sequence slice token $\mathbf{s}_h$, corresponding to the initial input token $\mathbf{s}$, as it encapsulates the integrated information from all multi-sequence patches within this slice.

### 2.3 Inter-slice Information Aggregation

Having obtained the multi-sequence slice tokens $\mathbf{s}_h$ from the previous level, we now aggregate information across the entire series of multi-sequence slices $[\mathbf{s}_1, \cdots, \mathbf{s}_{H-2}]$ to make the final classification. We begin by applying layer normalization to this sequence and, following the Vision Transformer (ViT) paradigm, we append a [class] token to the series of multi-sequence slice tokens. As with the

previous levels, this processed sequence is then input into a series of Transformer encoder blocks.

$$\begin{aligned} \mathbf{o}^{(0)} &= [\mathrm{LN}([\mathbf{s}_1; \mathbf{s}_2; \cdots; \mathbf{s}_{H-2}]); \mathbf{c}], \\ \mathbf{o}' &= \mathrm{MSA}\left(\mathrm{LN}\left(\mathbf{o}^{(\imath-1)}\right)\right) + \mathbf{o}^{(\imath-1)}, \\ \mathbf{o}^{(\imath)} &= \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{o}'\right)\right) + \mathbf{o}', \end{aligned} \tag{5}$$

where $\imath \in [1, L_\mathrm{f}]$, and $L_\mathrm{f}$ represents the number of Transformer blocks in this final level. The output corresponding to the [class] token $\mathbf{c}$ is then subjected to an additional layer normalization followed by a linear projection layer to produce the classification logits.

## 3 Experiments

**Data**: Our RadioFormer model was trained and evaluated on three datasets: the LLD-MMRI liver tumor dataset [16], the Advanced-MRI-Breast-Lesions (ABL) dataset [9], and the ReMIND brain tumor dataset [13]. The LLD-MMRI dataset consists of seven types of liver tumors (hepatocellular carcinoma, intrahepatic cholangiocarcinoma, hepatic metastasis, hepatic cysts, hepatic hemangiomas, focal nodular hyperplasia, hepatic abscesses), with a total of 498 cases, each having 8 MRI sequences. The ABL dataset includes 94 cases, each with 6 MRI sequences for benign and malignant breast tumor classification. The ReMIND dataset contains 71 cases, each with 2 MRI sequences for classifying three types of brain tumors (Oligodendroglioma, Astrocytoma, Glioblastoma). For the LLD-MMRI dataset, we follow the LLD-MMRI2023 challenge protocol[4], performing 5-fold cross-validation using the official train-validation splits (316 cases for training and validation, and 78 cases for testing, Stage One rule) and also report results on the test set (316 cases for training, 78 for validation, and 104 for testing, Stage Two rule). For the ABL and ReMIND datasets, we perform 5-fold cross-validation using random train-validation-test splits with a ratio of 3:1:1.

**Implementation Details**: The RadioFormer model was developed using PyTorch and all experiments were conducted on an RTX 4090 GPU. The architecture comprises three levels, each with a specified number of Transformer blocks: $L_\mathrm{s} = 12$ for the single-sequence slice feature extraction level, $L_\mathrm{m} = 2$ for the multi-sequence slice information aggregation level, and $L_\mathrm{f} = 4$ for the inter-slice information aggregation level. To initialize the first level, we loaded ImageNet-pretrained ViT model weights. For training, we employed the AdamW optimizer with an initial learning rate of 1e-4, which was adjusted using a cosine learning rate schedule. The minimum learning rate was set to 1e-5, and the training spanned 300 epochs. The first 5 epochs served as a warm-up phase, during which the learning rate was progressively increased. The standard cross-entropy was used as the loss function. The batch size was fixed at 4, and each lesion volume was resized to $16 \times 128 \times 128$. Our data augmentation strategy included random rotations and flips across various anatomical axes. During the training

[4] https://github.com/LMMMEng/LLD-MMRI2023

Table 1: Comparative performance for tumor classification on LLD-MMRI Stage One test set. The results are formulated as F1-score / Cohen's Kappa.

| | LLD-MMRI2023 Stage One train-val splits | | | | |
|---|---|---|---|---|---|
| Methods | Fold1 | Fold2 | Fold3 | Fold4 | Fold5 |
| Swin3D [7] | 0.662/0.610 | 0.643/0.606 | 0.679/0.634 | 0.644/0.612 | 0.675/0.644 |
| Transliver [25] | 0.714/0.695 | 0.705/0.667 | 0.687/0.681 | 0.650/0.624 | 0.689/0.679 |
| Swin-S [15] | 0.755/0.698 | **0.735**/0.694 | 0.721/0.662 | 0.734/0.729 | 0.747/0.689 |
| Resnet18 [12] (Rank 2) | 0.740/0.699 | 0.702/0.699 | 0.708/0.692 | 0.729/0.713 | 0.721/0.710 |
| Uniformer [14] (Rank 1) | 0.746/0.728 | 0.716/**0.724** | 0.721/0.693 | 0.732/0.717 | 0.693/0.674 |
| ViViT [2] | 0.663/0.619 | 0.692/0.671 | 0.733/0.691 | 0.691/0.667 | 0.697/0.684 |
| VideoMAE [21] | 0.719/0.701 | 0.676/0.645 | 0.722/0.706 | 0.712/0.672 | 0.693/0.674 |
| TimeSformer [5] | 0.641/0.609 | 0.689/0.681 | 0.714/0.666 | 0.643/0.607 | 0.641/0.622 |
| RadioFormer | **0.796**/**0.777** | 0.710/0.713 | **0.760**/**0.747** | **0.771**/**0.761** | **0.796**/**0.779** |

phase, lesion volumes were randomly cropped to 14 × 112 × 112. For evaluation, a central crop of the same dimensions was extracted. In line with the LLD-MMRI2023 challenge, the model's performance was evaluated using the F1-score and Cohen's Kappa metrics. On LLD-MMRI dataset, the training time of each model is about 2 hours, and the inference time is 1s per case.

**Results**: In Table 1, we present the detailed performance results for each fold under the LLD-MMRI Stage One rule. To ensure a comprehensive comparison, we evaluate recent classification methods, including Transliver [25], a Transformer-based model specifically designed for liver tumor classification; Swin-S [15]; Swin3D [7]; and video-based models such as ViViT [2], VideoMAE [21], and TimeSformer [5]. Additionally, we compare our method with the top two approaches from the LLD-MMRI2023 challenge[5], which were based on a modified Uniformer [14] (Rank 1) and a ResNet18 model [12] (Rank 2). Our RadioFormer consistently outperforms all compared methods in terms of both F1-score and Cohen's Kappa across four out of five folds, achieving competitive results on the remaining fold. This strong and consistent performance across different train-validation splits underscores the effectiveness of our approach in liver tumor classification. We further report the average five-fold cross-validation results for the ABL and ReMIND datasets, along with the results on the LLD-MMRI Stage Two test set in Table 2. In all cases, RadioFormer achieves the best overall performance, surpassing the recent SDR-Former [16], the solution provided by the official LLD-MMRI dataset team. Additionally, we report the computational complexity of the comparative methods. As shown, RadioFormer has fewer training parameters than most competing methods while maintaining moderate GFLOPS, demonstrating a favorable balance between efficiency and accuracy.

**Ablation Study**: To evaluate the contributions of the different levels in our RadioFormer model, we first conducted an ablation study focusing on the multi-sequence slice information aggregation (second) level and the inter-slice information aggregation (third) level. We designed two variant models for this purpose. Model1: In this variant, we replaced the second level of our RadioFormer with a global average pooling layer. This modification eliminates the explicit feature processing that occurs at the second level. Model2: we replaced the third level of

[5] The codes of the top five teams are open-sourced.

Table 2: Comparative performance for tumor classification on the LLD-MMRI Stage Two test set and the average 5-fold cross-validation results on the ABL and ReMIND datasets. The results are reported as F1-score / Cohen's Kappa.

| | Datasets | | | Complexity |
|---|---|---|---|---|
| Methods | ABL | ReMIND | LLD-MMRI Test | GFLOPS/Training Params (M) |
| Swin3D [7] | 0.611/0.214 | 0.407/0.182 | 0.688/0.651 | 21.7/31.6 |
| Transliver [25] | 0.591/0.178 | 0.422/0.179 | 0.716/0.654 | 572.8/154.6 |
| Resnet18 [12] | 0.646/0.311 | 0.489/0.249 | 0.711/0.665 | 384.0/15.1 |
| Uniformer [14] | 0.542/0.103 | 0.491/0.258 | 0.712/0.666 | 112.6/49.5 |
| ViViT [2] | 0.621/0.303 | 0.364/0.102 | 0.733/0.691 | 67.1/88.5 |
| VideoMAE [21] | 0.630/0.290 | 0.417/0.173 | 0.722/0.706 | 64.4/88.0 |
| TimeSformer [5] | 0.589/0.181 | 0.433/0.178 | 0.714/0.666 | 84.2/ 122.9 |
| H2Former[16] | - | - | 0.774/0.726 | - |
| SDR-Former[16] | - | - | 0.791/0.747 | - |
| RadioFormer | **0.695**/**0.362** | **0.568**/**0.357** | **0.806**/**0.745** | 209.8/32.1 |

Table 3: Ablations on the effects of second and third levels.

| model | F1-score/Cohen's Kappa |
|---|---|
| Model1 | 0.788/0.762 |
| Model2 | 0.815/0.807 |
| RadioFormer | 0.838/0.812 |

our RadioFormer with a global average pooling layer. We then compared their average F1-score and Cohen's Kappa of the last 100 epochs on the fold1 validation subset. The results are shown in Table 3. The comparison reveals that both the second and third levels of RadioFormer contribute to the model's overall performance.

Our RadioFormer explicitly processes the input multi-sequence data in three levels, following a C-N-H order. First, it fuses the slices along the sequence dimension (C), reducing the number of channels (C) to 1. Next, it consolidates the N tokens within each slice into a single token. Finally, it aggregates the H slice tokens to produce the final classification result. This approach mirrors the clinical practice of radiologists. Additionally, we tested other parsing orders on the LLD-MMRI Stage One Fold1, with results presented in Table 4. As shown, our C-N-H parsing order achieves the best performance. This further demonstrates the importance of incorporating radiologists' inductive bias into model design.

## 4 Conclusion

In this work, we introduce RadioFormer, a radiologist-inspired model designed to effectively learn from modestly sized multi-sequence MRI tumor datasets. Unlike traditional 2D or 3D model architectures, RadioFormer incorporates a hybrid 2D-3D data processing flow that mirrors the way radiologists interpret MRI images. Our extensive testing on three different datasets demonstrates that RadioFormer consistently delivers stable and high-performance results across all datasets.

Table 4: Ablations on the parsing order.

| order | F1-score/Cohen's Kappa |
|---|---|
| N-C-H | 0.656/0.619 |
| N-H-C | 0.683/0.637 |
| H-N-C | 0.631/0.598 |
| H-C-N | 0.729/0.726 |
| C-H-N | 0.754/0.741 |
| C-N-H | 0.796/0.777 |

**Acknowledgments.** This work was supported in part by the Shenzhen Science and Technology Program under Grants JCYJ20220530161616036, in part by the National Natural Science Foundation of China under Grants 62171377 and 92470101, in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang, China, under Grant 2025C01201(SD2), and in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYKFZD06).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Ardila, D., et al.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine **25**(6), 954–961 (2019)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
3. Ayadi, W., Elhamzi, W., Charfi, I., Atri, M.: Deep cnn for brain tumor classification. Neural processing letters **53**, 671–700 (2021)
4. Bai, X., et al.: UAE: Universal anatomical embedding on multi-modality medical images. arXiv preprint arXiv:2311.15111 (2024)
5. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
6. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **68**(6), 394–424 (2018)
7. Cardoso, M.J., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
8. Dai, C., et al.: Deep learning assessment of small renal masses at contrast-enhanced multiphase ct. Radiology **311**(2), e232178 (2024)
9. Daniels, D., et al.: Standard and delayed contrast-enhanced MRI of malignant and benign breast lesions with histological and clinical supporting data (advanced-mri-breast-lesions)(version 2). `https://www.cancerimagingarchive.net/collection/advanced-mri-breast-lesions/`
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

11. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 289–293. IEEE (2018)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
13. Juvekar, P., et al.: The brain resection multimodal imaging database (ReMIND). https://www.cancerimagingarchive.net/collection/remind/
14. Li, K., Wang, Y., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unified transformer for efficient spatiotemporal representation learning. arXiv preprint arXiv:2201.04676 (2022)
15. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
16. Lou, M., Ying, H., Liu, X., Zhou, H.Y., Zhang, Y., Yu, Y.: Sdr-former: A siamese dual-resolution transformer for liver lesion classification using 3d multi-phase imaging. Neural Networks p. 107228 (2025)
17. Qu, R., Xiao, Z.: An attentive multi-modal cnn for brain tumor radiogenomic classification. Information **13**(3), 124 (2022)
18. Romero, F.P., et al.: End-to-end discriminative deep network for liver lesion classification. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 1243–1246. IEEE (2019)
19. Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W.: Explainable hierarchical semantic convolutional neural network for lung cancer diagnosis. In: CVPR workshops. pp. 63–66 (2019)
20. Swati, Z.N.K., Zhao, Q., Kabir, M., Ali, F., Ali, Z., Ahmed, S., Lu, J.: Brain tumor classification for mr images using transfer learning and fine-tuning. Computerized Medical Imaging and Graphics **75**, 34–46 (2019)
21. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)
22. Tummala, S., Kadry, S., Bukhari, S.A.C., Rauf, H.T.: Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. Current Oncology **29**(10), 7498–7511 (2022)
23. Wang, R., Shi, X., Pang, S., Chen, Y., Zhu, X., Wang, W., Cai, J., Song, D., Li, K.: Cross-attention guided loss-based deep dual-branch fusion network for liver tumor classification. Information Fusion **114**, 102713 (2025)
24. Wang, W., et al.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
25. Wang, X., Ying, H., Xu, X., Cai, X., Zhang, M.: Transliver: A hybrid transformer model for multi-phase liver lesion classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 329–338. Springer (2023)
26. Xi, I.L., Zhao, Y., Wang, R., Chang, M., Purkayastha, S., Chang, K., Huang, R.Y., Silva, A.C., Vallieres, M., Habibollahi, P., et al.: Deep learning to distinguish benign from malignant renal lesions based on routine mr imaging. Clinical Cancer Research **26**(8), 1944–1952 (2020)
27. Xu, X., Zhu, Q., Ying, H., Li, J., Cai, X., Li, S., Liu, X., Yu, Y.: A knowledge-guided framework for fine-grained classification of liver lesions based on multi-phase

ct images. IEEE Journal of Biomedical and Health Informatics **27**(1), 386–396 (2022)
28. Yasaka, K., Akai, H., Abe, O., Kiryu, S.: Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced ct: a preliminary study. Radiology **286**(3), 887–896 (2018)
29. Ying, H., Liu, X., Zhang, M., Ren, Y., Zhen, S., Wang, X., Liu, B., Hu, P., Duan, L., Cai, M., et al.: A multicenter clinical ai system study for detection and diagnosis of focal liver lesions. Nature Communications **15**(1), 1131 (2024)
30. Zhang, Y., Chen, J.H., Lin, Y., Chan, S., Zhou, J., Chow, D., Chang, P., Kwong, T., Yeh, D.C., Wang, X., et al.: Prediction of breast cancer molecular subtypes on dce-mri using convolutional neural network with transfer learning between two centers. European radiology **31**, 2559–2567 (2021)
31. Zhang, Y., Liu, Y.L., Nie, K., Zhou, J., Chen, Z., Chen, J.H., Wang, X., Kim, B., Parajuli, R., Mehta, R.S., et al.: Deep learning-based automatic diagnosis of breast cancer on mri using mask r-cnn for detection followed by resnet50 for classification. Academic radiology **30**, S161–S171 (2023)
32. Zhou, J., Wang, W., Lei, B., Ge, W., Huang, Y., Zhang, L., Yan, Y., Zhou, D., Ding, Y., Wu, J., et al.: Automatic detection and classification of focal liver lesions based on deep convolutional neural networks: a preliminary study. Frontiers in oncology **10**, 581210 (2021)