PDF-Net: Prototype-Aware Dynamic Fusion Network for Nasopharyngeal Carcinoma T-staging Classification with Epstein-Barr Virus DNA

Wantong Lu[1,2,3], Xu Han[1], Yibo Wei[1,2,3], Zanting Ye[1,2,3], and Lijun Lu[1,2,3,4](✉)

[1]School of Biomedical Engineering, Southern Medical University, Guangzhou, China
ljlubme@gmail.com
[2]Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, China
[3]Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou, China
[4]Pazhou Lab, Guangzhou, China

Abstract. Accurate T-staging classification of nasopharyngeal carcinoma (NPC) is crucial for guiding individualized treatment strategies and predicting patient prognosis. However, this task remains challenging due to the limitations of unimodal approaches, which often fail to capture the full complexity of NPC progression, and the severe class imbalance in clinical datasets, where early-stage cases (T1 / T2 stage) are significantly underrepresented. In this paper, we propose a Prototype-Aware Dynamic Fusion Network (PDF-Net), a novel multimodal framework that integrates MR images with Epstein-Barr virus (EBV) DNA tabular data to improve NPC T-staging classification. Our framework introduces two key components: (1) the Dynamic Multi-Modal Alignment (DMMA) module, which aligns MR imaging features with EBV DNA data to capture complementary information across modalities, and (2) the Optimal Prototype-Aware Transport (OPAT) module, which incorporates a Prototypical Constraint to enhance the representation of T2-staging features and mitigate class imbalance. To the best of our knowledge, PDF-Net is the first framework to leverage EBV DNA data as an auxiliary tool for T-staging classification, significantly improving accuracy and robustness. Experimental results in a real clinical dataset demonstrate that our approach outperforms state-of-the-art methods, achieving an accuracy of 0.8006 ± 0.0488 and an AUC of 0.8191 ± 0.0551 for T1C images, highlighting its potential to advance NPC diagnosis and personalized treatment strategies.

Keywords: Nasopharyngeal Carcinoma · T-staging · Epstein-Barr Virus · Multimodal Fusion · Few-shot Learning · Optimal Transport.

1 Introduction

Nasopharyngeal carcinoma (NPC) is a malignancy originating from the epithelial lining of the nasopharynx, with magnetic resonance imaging (MRI) serving as the primary modality for tumor assessment and staging due to its superior soft tissue contrast

W. Lu, X. Han—Co-first authors.

and noninvasive characteristics[1, 2]. In the current, the clinical staging of NPC predominantly relies on the UICC/AJCC tumor-node-metastasis (TNM) staging system, in which the T-staging (T1 – T4), reflecting the size, location, and extent of the primary tumor invasion, plays a pivotal role in treatment planning[3, 4]. Since the treatment approach for NPC varies significantly across different stages, accurate T-staging is critical for guiding individualized precision treatment and predicting prognosis.

In clinical practice, T-staging of NPC primarily depends on manual slice-by-slice inspection of MR images, which is time-consuming, labor-intensive, and highly dependent on the expertise of radiologists[1]. These challenges make it difficult to obtain large-scale annotated MR image datasets, resulting in a scarcity of labeled data. Consequently, this manual approach is neither efficient nor reliable enough to be considered the gold standard for staging. These limitations underscore the urgent need to develop effective computer-aided diagnosis (CAD) systems to improve the accuracy and efficiency of NPC staging, ultimately supporting individualized precision treatment.

The clinical manifestations of NPC are highly diverse, with early-stage symptoms often being subtle and nonspecific. As a result, the majority of patients are diagnosed at locally advanced stages (T3 / T4 stage), while early-stage diagnoses (T1 / T2 stage) remain relatively rare. This disparity creates a significant class imbalance in clinical datasets, posing a major challenge for accurate staging. Models trained on such imbalanced data tend to exhibit a bias toward predicting advanced stages, resulting in poor recognition of early-stage cases. Addressing this class imbalance is critical for improving the accuracy of early-stage diagnosis and enabling timely, personalized treatment strategies. In recent years, few-shot learning (FSL) has demonstrated considerable potential in medical image analysis, particularly in scenarios with limited sample sizes, by enabling efficient feature extraction and classification with minimal data[5, 6]. Consequently, FSL holds significant promise for enhancing early diagnosis in NPC, aligning with our goal of early detection, timely treatment, and the implementation of individualized precision medicine.

Although deep learning has made significant progress in NPC staging, most existing studies focus solely on unimodal imaging data, overlooking the rich information contained in clinical tabular data[7, 8]. Vision-only approaches often misclassify T1/T2 cases due to their subtle features or fail to differentiate between tumor invasion, inflammation, and edema, leading to staging inaccuracies. Alongside imaging, plasma circulating cell-free EBV DNA has emerged as a sensitive and specific biomarker for EBV-associated NPC, which consists of short DNA fragments released by NPC cells, can be detected through highly sensitive polymerase chain reaction (PCR)[9, 10]. However, the measurement of EBV DNA is challenging due to substantial interlaboratory variability. Despite this, dynamic changes in EBV DNA levels during treatment have been linked to therapeutic efficacy and prognosis, with complete biological response (cBR) after multiple chemotherapy cycles serving as a strong predictor of long-term outcomes[11]. Given the close association between cfEBV DNA levels and NPC prognosis, integrating this biomarker with imaging data can significantly enhance the accuracy of T-stage prediction, providing a more comprehensive assessment of disease progression and treatment response.

Motivated by the above discussion, we propose a **Prototype-Aware Dynamic Fusion Network (PDF-Net)**, a novel multimodal framework that integrates MR images with EBV DNA tabular data for the classification of NPC T-staging. Our main contributions are as follows:

- We propose the **Dynamic Multi-Modal Alignment (DMMA)** module, which enables structured fusion of MR images and EBV tabular data, allowing the model to learn clinically relevant progression patterns beyond vision-only staging.
- We introduce the **Optimal Prototype-Aware Transport (OPAT)** module, which incorporates Prototypical Constraint into the Optimal Transport framework to enhance T2 feature representation and mitigate class imbalance.
- To the best of our knowledge, PDF-Net is the first framework to incorporate EBV DNA data as an auxiliary tool for T-staging classification, effectively leveraging complementary information across modalities to significantly improve classification accuracy.

2 Method

Figure 1 presents the overall framework of our proposed method for integrating MR images and EBV DNA tabular data. The framework consists of three key components: the PDF-Net, the DMMA module, and the OPAT module. These components work collaboratively to enhance T-staging classification by leveraging multi-modal data fusion and representation learning.

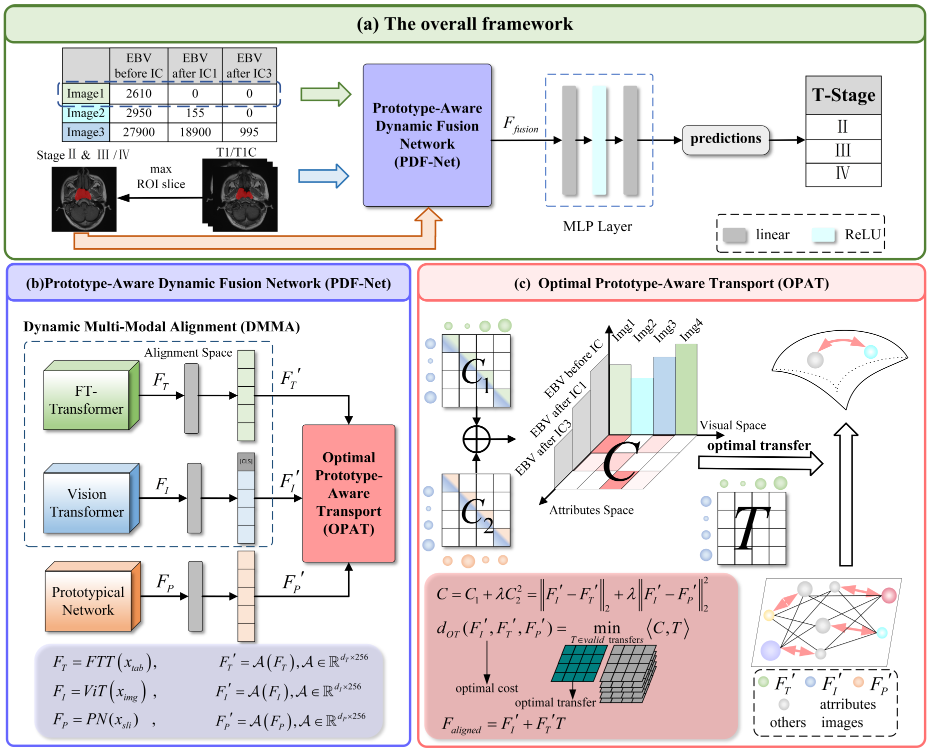
2.1 Dynamic Multi-Modal Alignment (DMMA) module

Muti-modal Feature Extraction. We extract features separately from two modalities: two sequences of MR images and EBV Dynamics tabular data. For images, we utilize a Vision Transformer (ViT) to extract image features $F_I \in \mathbb{R}^{d_I}$ from each MRI sequence[12]. The extracted image features capture spatial and structural characteristics of the tumor region, which are essential for accurate T-staging classification. For tabular data, we employ the FT-Transformer to process the data and extract tabular features $F_T \in \mathbb{R}^{d_T}$ that represent the temporal dynamics and numerical relationships in the EBV values across the different chemotherapy stages[13].

Since these extracted features from different modalities and have different dimensions, we project them into a unified 256-dimensional alignment space $\mathcal{A}$ using a linear transformation layer before further processing. After projection, both sets of features are transformed into $F_I^{'} \in \mathbb{R}^{1\times256}$ and $F_T^{'} \in \mathbb{R}^{1\times256}$. The alignment is achieved through the following linear transformations:

$$F_I^{'} = \mathcal{A}(F_I) = linear(F_I), \mathcal{A} \in \mathbb{R}^{d_I\times256}, \tag{1}$$

$$F_T^{'} = \mathcal{A}(F_T) = linear(F_T), \mathcal{A} \in \mathbb{R}^{d_T\times256}. \tag{2}$$



	EBV before IC	EBV after IC1	EBV after IC3
Image1	2610	0	0
Image2	2950	155	0
Image3	27900	18900	995

(a) The overall framework
Stage II & III / IV
max
ROI slice
T1/T1C
Prototype-Aware Dynamic Fusion Network (PDF-Net)
F_{fusion}
MLP Layer
predictions
T-Stage
II
III
IV
linear
ReLU
(b)Prototype-Aware Dynamic Fusion Network (PDF-Net)
Dynamic Multi-Modal Alignment (DMMA)
Alignment Space
FT-Transformer
Vision Transformer
Prototypical Network
F_T
F_T'
F_I
F_I'
F_P
F_P'
[CLS]
Optimal Prototype-Aware Transport (OPAT)
$F_T = FTT(x_{tab})$,
$F_T' = \mathcal{A}(F_T), \mathcal{A} \in \mathbb{R}^{d_T \times 256}$
$F_I = ViT(x_{img})$,
$F_I' = \mathcal{A}(F_I), \mathcal{A} \in \mathbb{R}^{d_I \times 256}$
$F_P = PN(x_{sli})$,
$F_P' = \mathcal{A}(F_P), \mathcal{A} \in \mathbb{R}^{d_P \times 256}$
(c) Optimal Prototype-Aware Transport (OPAT)
C_1
C_2
C
T
Img1
Img2
Img3
Img4
EBV before IC
EBV after IC1
EBV after IC3
Visual Space
Attributes Space
optimal transfer
$C = C_1 + \lambda C_2^2 = \left\|F_I' - F_T'\right\|_2 + \lambda\left\|F_I' - F_P'\right\|_2^2$
$d_{OT}(F_I', F_T', F_P') = \min_{T \in valid\ transfers} \langle C, T \rangle$
optimal cost
optimal transfer
$F_{aligned} = F_I' + F_T'T$
F_T'
F_I'
F_P'
others
atrributes
images

Fig. 1. Overview of the proposed framework for T-staging classification. (a) The overall framework integrating MR images and EBV DNA tabular data; (b) The PDF-Net; (c) The OPAT module. In (c), color represents modality: blue for image, green for EBV tabular data, and orange for the T2 prototype. Red arrows indicate alignment flows via optimal transport. Ball size reflects different patients and is purely for visualization—not quantitative.

Tabular-to-Image Alignment. To align the image and tabular features, we employ Optimal Transport (OT) to compute a cost matrix based on Euclidean distance, which measures the pairwise similarity between the image features and the tabular features[14, 15]. Smaller distances indicate greater similarity in the shared feature space. The cost matrix C_1 is computed as follows:

$$C_1 = \|F_I' - F_T'\|_2. \tag{3}$$

2.2 Optimal Prototype-Aware Transport (OPAT) module

Prototype Extraction. To address the class imbalance, particularly the underrepresentation of T2-stage patients, we introduce a FSL approach[16, 17]. We use a Prototypical Network to extract prototype features for the T2 stage[18]. The input to the net is the axial slice with the largest Region of Interest (ROI) area from the 3D MR volume for each patient. This slice is selected as it contains the most significant pathological information relevant to the T-staging task.

Specifically, we fix the T2-stage samples and randomly sample T3- and T4-stage patients to construct a binary T2 vs. non-T2 classification task. We adopt few-shot episodic training with 5 support and 5 query samples per class in each episode, maintaining a 1:1 balance between T2 and T3/T4 cases. For each training epoch, the T2 prototype is computed by averaging the feature representations of all T2 samples in the support set, and the model is evaluated on the query set. We select the prototype from the epoch that achieves the best performance on the validation set as the final T2 prototype $F_P \in \mathbb{R}^{d_P}$, which captures key characteristics of the T2 stage and enhances representation for this minority class.

To ensure consistency across modalities, the extracted prototype features are projected into a 256-dimensional alignment space using the following transformation:

$$F_P^{'} = \mathcal{A}(F_P) = linear(F_P), \mathcal{A} \in \mathbb{R}^{d_P \times 256}. \tag{4}$$

The prototype extraction process is performed separately from the OPAT module and does not involve joint optimization. The resulting prototype is fixed and used during OPAT training as a prior for guiding alignment.

Prototype-to-Image Alignment. Similarly, we compute the distance between prototype and image features, reinforcing alignment for the underrepresented T2 stage. The cost matrix C_2 is computed as follows:

$$C_2 = ||F_I^{'} - F_P^{'}||_2. \tag{5}$$

Prototypical Constraint. The final cost matrix integrates T2 prototype constraints to mitigate bias towards majority classes (T3 & T4 stage), improving T2-stage recognition:

$$C = C_1 + \lambda C_2^2 = ||F_I^{'} - F_T^{'}||_2 + \lambda ||F_I^{'} - F_P^{'}||_2^2, \tag{6}$$

where λ is a weighting parameter that controls the influence of the T2 prototype constraint. The addition of this constraint helps to mitigate the bias towards the majority classes (T3 and T4), improving the model's ability to predict the T2 stage more accurately.

Optimal Prototype-Aware Transport. After computing the cost matrix,the Sinkhorn Algorithm is applied to compute the optimal transport plan that minimizes the cost of aligning tabular features to image features[19]. The optimization problem is formulated as:

$$d_{OT}(F_I^{'}, F_T^{'}, F_P^{'}) = \min_{T \in \text{valid plans}} \langle C, T \rangle, \tag{7}$$

where C is the total cost matrix, T is the transport matrix. The objective of the problem is to minimize the cost associated with transporting the tabular features to the image features. The result is the optimal transport matrix T , which indicates the best alignment between the features from both modalities, taking into account both the spatial and temporal aspects of the data. This optimal transport framework ensures that the multimodal features are aligned in a way that minimizes the overall cost, leading to improved feature fusion and more accurate T-stage classification.

Muti-modal Feature Fusion. After obtaining the transport matrix, the aligned features are computed by combining the image features $F_I^{'}$ and the tabular features $F_T^{'}$ according to the transport matrix $T^{'}$:

$$F_{aligned} = F_I^{'} + F_T^{'} T. \tag{8}$$

Finally, the aligned features are passed through a multi-layer perceptron (MLP) classifier for the final T-stage classification. The addition of prototype constraints ensures that T2 patients are better represented in the feature space, thereby improving classification performance for this underrepresented group.

3 Experiment

3.1 Experimental Settings

Dataset. To validate the effectiveness of our method, we conducted a study using data from a single medical center, comprising a total of 1,634 NPC patients. For T-staging classification, the dataset includes 113 T2-stage cases, 893 T3-stage cases, and 628 T4-stage cases. We collected T1-weighted imaging (T1w) and contrast-enhanced T1-weighted imaging (T1Cw) sequence, along with the corresponding primary tumor mask images. To prepare the input data, we multiplied the T1w and T1Cw images with their respective mask images, effectively highlighting the ROIs corresponding to the primary tumor. These masked images were then used as inputs to our model, ensuring that the focus remained on the clinically relevant areas for accurate T-staging classification. Additionally, we utilized clinical information in the form of EBV DNA levels, measured at three time points: before chemotherapy, after one cycle of chemotherapy, and after three cycles of chemotherapy. Among the 1,634 patients, only 802 had complete EBV DNA measurements across all three time points.

Implementation Details. The MR images were resampled to a resolution of (2, 2, 2) and resized to a uniform dimension of (112, 112, 112). All images were normalized to ensure consistency. The dataset was split into training, validation, and testing sets in an 7:1.5:1.5 ratio, ensuring that the division was consistent for both MR images and EBV DNA data.

The training process was divided into three independent stages. First, we trained modality-specific encoders: a ViT for MR images and an FT-Transformer for EBV DNA data. Features were extracted before the final classification layers and fixed for downstream use. Second, Prototypical Network was also trained independently to extract a fixed T2 prototype, which was used in the OPAT module to guide the alignment process. Finally, the OPAT module was trained using the extracted image and tabular features along with the fixed prototype. During this stage, only the OPAT parameters were updated.

Models were trained using the Adam optimizer with a learning rate schedule: 1e-4 initially, decayed to 5e-5 at epoch 50 and 2e-5 at epoch 100, for a total of 150 epochs. The best checkpoint was selected based on validation ACC. All code was implemented in PyTorch and executed on a NVIDIA RTX 4090 GPU.

Table 1. Comparisons between our proposed method and other SOTA appoaches.

Method	Images	EBV	T1		TIC	
			ACC	AUC	ACC	AUC
ResNet18[20]	✓		0.6497±0.0279	0.7453±0.0165	0.6850±0.0346	0.7558±0.0382
ResNet34[20]	✓		0.6791±0.0485	0.7454±0.0246	0.7162±0.0.371	0.7524±0.0138
DenseNet121[21]	✓		0.6871±0.0340	0.7768±0.0285	0.7122±0.0336	**0.7536±0.0357**
ViT[12]	✓		**0.7170±0.0368**	**0.7870±0.2222**	**0.7243±0.0320**	0.7424±0.0362
BERT[22]		✓	0.5949±0.0346	0.5780±0.3250		
FT-Transformer[13]		✓	**0.6203±0.0295**	**0.5762±0.0271**		
CA[23]	✓	✓	0.6703±0.0314	0.6900±0.0321	0.6209±0.0345	0.6626±0.0335
Concat	✓	✓	0.7403±0.0580	0.7443±0.0389	0.7431±0.0442	0.7485±0.0770
OT[14]	✓	✓	0.7659±0.0431	0.7802±0.0495	0.7736±0.0524	0.7522±0.0551
GMU[24]	✓	✓	0.7535±0.0313	0.778±0.0290	0.7592±0.0368	0.7645±0.0200
3MT[25]	✓	✓	0.7614±0.0385	0.7805±0.0453	0.7778±0.0669	0.7876±0.0684
AGGN[26]	✓	✓	0.7432±0.0320	0.7495±0.0218	0.7615±0.0403	0.7868±0.0186
Ours	✓	✓	**0.7762±0.0317**	**0.7924±0.0183**	**0.8006±0.0488**	**0.8191±0.0551**

Evaluation Metrics. We conducted comparative experiments and ablation studies to demonstrate the advantages of our method. The performance was evaluated using several metrics, including accuracy (ACC), area under the curve (AUC). These metrics were chosen to comprehensively assess the model's ability to classify the T stage of NPC accurately and reliably.

3.2 Result and Analysis

We performed comparative experiments and ablation studies to demonstrate the advantages of our proposed framework.

Comparison to Other Methods. For single image modality validation, we compared our method with classical classification networks, including ResNet18, ResNet34[20], DenseNet121[21], and ViT[12]. For clinical tabular data, we compared FT-Transformer[13] with BERT[22] to assess the performance of structured tabular data versus textual data representation. To validate the effectiveness of multimodal fusion, we compared our method with several state-of-the-art approaches, including Cross Attention (CA)[23], Concatenation, OT[14, 15], GMU[24], 3MT[25], and AGGN[26]. All multimodal baselines use the same ViT and FT-Transformer encoders as our method to ensure a fair comparison.

As shown in Table 1, our model achieved the best classification performance. For T1 images, the ACC and AUC were 0.7762 ± 0.0371 and 0.7924 ± 0.0138, respectively. For T1C images, the performance further improved, with an ACC of 0.8006 ± 0.0488 and an AUC of 0.8191 ± 0.0551. For single image modality validation, ViT achieved the best performance when using masked images as input, demonstrating its superior ability to capture subtle features in the ROIs. For clinical data, results demonstrated

Table 2. Ablation results of our method.

Method	T1		T1C	
	ACC	AUC	ACC	AUC
w / o OPAT, w / o DMMA	0.7170±0.0368	0.7870±0.0222	0.7243±0.0302	0.7424±0.0326
w / o DMMA	0.7432±0.0488	0.7458±0.0549	0.7436±0.0629	0.7501±0.0637
w / o OPAT	0.7403±0.0580	0.7443±0.0879	0.7431±0.0422	0.7485±0.0770
w / o PC	0.7569±0.0431	0.7802±0.0915	0.7736±0.0524	0.7522±0.0951
Ours	**0.7762±0.0371**	**0.7924±0.0138**	**0.8006±0.0488**	**0.8191±0.0551**

that methods based on structured tabular data outperformed text-based approaches, underscoring the advantages of structured numerical data over textual representations for NPC T-stage classification. However, single-modality approaches, whether based on masked images or EBV DNA tabular data, consistently underperformed compared to multimodal methods, underscoring the importance of integrating complementary information from multiple modalities. Additionally, other multimodal fusion methods either ignore the significant dimensional differences between modalities or are designed for prognostic tasks, which may not be directly applicable to the unique diagnostic scenarios of NPC. These limitations often result in suboptimal alignment between image and non-image feature spaces. In contrast, our PDF-Net effectively bridges the gap between MR images and EBV DNA tabular data, enabling the network to learn more discriminative features and significantly improving classification performance.

Ablation Study. We evaluated the importance of key components in our framework through ablation experiments (Table 2). The results demonstrate that removing any module leads to a decline in both ACC and AUC, highlighting the critical role of each component in the overall performance of our model. Specifically, the removal of the OPAT module caused the most significant performance drop, particularly for early-stage (T2) cases, as it weakens the model's ability to handle class imbalance and enhance the representation of underrepresented T2-stage features. Similarly, removing the DMMA module led to a substantial decrease in performance, as it disrupts the alignment of MR image features with EBV DNA tabular data, limiting the model's ability to capture complementary information across modalities.

These results underscore the importance of both the OPAT and DMMA modules in improving the overall accuracy and robustness of NPC T-staging classification. The OPAT module is particularly critical for addressing class imbalance and enhancing early-stage detection, while the DMMA module enables effective multimodal fusion by aligning MR image features with EBV DNA tabular data. Additionally, the Prototypical Constraint (PC) contributes to refining feature representation, further supporting the model's performance.

4 Conclusion

We propose a PDF-Net for the multimodal T-staging classification of NPC for the first time. This method effectively integrates MR images and EBV DNA tabular data, achieving state-of-the-art classification performance. By leveraging the DMMA module, we align MR image features with EBV DNA data to capture complementary information across modalities. Additionally, the OPAT module enhances the representation of underrepresented T2-stage cases, mitigating class imbalance and improving early-stage NPC classification. Our framework pioneers a promising direction for multimodal diagnosis of NPC, offering significant potential for improving diagnostic accuracy and personalized treatment strategies.

Acknowledgments. This research was funded by the National Natural Science Foundation of China (62371221, 12326616 and 62201245), National High end Foreign Experts Recruitment Plan (G2023030025L), the Science and Technology Program of Guangdong Province 2022A0505050039 and the Open Project Program of Yunnan Key Laboratory of Intelligent Systems and Computing ISC23Y02.

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

References

1. Chen, Y. P., Chan, A. T., Le, Q. T., Blanchard, P., Sun, Y., Ma, J.: Nasopharyngeal carcinoma. The Lancet **394**(10192), 64–80 (2019)
2. King, A. D.: MR imaging of nasopharyngeal carcinoma. Magnetic Resonance Imaging Clinics of North America **30**(1), 19–33 (2022)
3. Pan, J. J., Mai, H. Q., Ng, W. T., Hu, C. S., Li, J. G., Chen, X. Z., . . . , Lee, A. W.: Ninth version of the AJCC and UICC nasopharyngeal cancer TNM staging classification. JAMA Oncology (2024)
4. Lydiatt, W. M., Patel, S. G., O'Sullivan, B., Brandwein, M. S., Ridge, J. A., Migliacci, J. C., . . . , Shah, J. P.: Head and neck cancers—major changes in the American Joint Committee on cancer eighth edition cancer staging manual. CA: A Cancer Journal for Clinicians **67**(2), 122–137 (2017)
5. Zheng, F., Cao, J., Yu, W., Chen, Z., Xiao, N., Lu, Y.: Exploring low-resource medical image classification with weakly supervised prompt learning. Pattern Recognition **149**, 110250 (2024)
6. Cheng, Z., Wang, S., Xin, T., Zhou, T., Zhang, H., Shao, L.: Few-shot medical image segmentation via generating multiple representative descriptors. IEEE Transactions on Medical Imaging **43**(6), 2202–2214 (2024)
7. Yang, Q., Guo, Y., Ou, X., Wang, J., Hu, C.: Automatic T staging using weakly supervised deep learning for nasopharyngeal carcinoma on MR images. Journal of Magnetic Resonance Imaging **52**(4), 1074–1082 (2020)
8. Feng, Q., Liang, J., Wang, L., Niu, J., Ge, X., Pang, P., Ding, Z.: Radiomics analysis and correlation with metabolic parameters in nasopharyngeal carcinoma based on PET/MR imaging. Frontiers in Oncology **10**, 1619 (2020)
9. Lo, Y. D., Chan, A. T., Chan, L. Y., Leung, S. F., Lam, C. W., Huang, D. P., Johnson, P. J.: Molecular prognostication of nasopharyngeal carcinoma by quantitative analysis of circulating Epstein-Barr virus DNA. Cancer Research **60**(24), 6878–6881 (2000)
10. Lv, J., Wu, C., Li, J., Chen, F., He, S., He, Q., . . . , Lin, L.: Improving on-treatment risk stratification of cancer patients with refined response classification and integration of circulating tumor DNA kinetics. BMC Medicine **20**(1), 268 (2022)

11. Lv, J., Chen, Y., Zhou, G., Qi, Z., Tan, K. R. L., Wang, H., . . . , Sun, Y.: Liquid biopsy tracking during sequential chemo-radiotherapy identifies distinct prognostic phenotypes in nasopharyngeal carcinoma. Nature Communications **10**(1), 3941 (2019)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . , Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems **34**, 18932–18943 (2021)
14. Caffarelli, L. A., McCann, R. J.: Free boundaries in optimal transport and Monge-Ampere obstacle problems. Annals of Mathematics **673**–730 (2010)
15. Montesuma, E. F., Mboula, F. M. N., Souloumiac, A.: Recent advances in optimal transport for machine learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
16. Lu, J., Gong, P., Ye, J., Zhang, J., Zhang, C.: A survey on machine learning from few samples. Pattern Recognition **139**, 109480 (2023)
17. Li, J., Li, M.: Few-shot Learning: Methods and Applications. In: ITM Web of Conferences, vol. 70, p. 02012. EDP Sciences (2025)
18. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems **30** (2017)
19. Knight, P. A.: The Sinkhorn–Knopp algorithm: convergence and applications. SIAM Journal on Matrix Analysis and Applications **30**(1), 261–275 (2008)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
21. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K. Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
22. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019)
23. Gheini, M., Ren, X., May, J.: Cross-attention is all you need: Adapting pretrained transformers for machine translation. arXiv preprint arXiv:2104.08771 (2021)
24. Qin, R., Wang, Z., Jiang, L., Qiao, K., Hai, J., Chen, J., . . . , Yan, B.: Fine-grained lung cancer classification from PET and CT images based on multidimensional attention mechanism. Complexity **2020**(1), 6153657 (2020)
25. Liu, L., Liu, S., Zhang, L., To, X. V., Nasrallah, F., Chandra, S. S.: Cascaded multi-modal mixing transformers for Alzheimer's disease classification with incomplete data. NeuroImage **277**, 120267 (2023)
26. Wu, P., Wang, Z., Zheng, B., Li, H., Alsaadi, F. E., Zeng, N.: AGGN: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion. Computers in Biology and Medicine **152**, 106457 (2023)