Multi-view Contrastive Learning with Additive Margin for Adaptive Nasopharyngeal Carcinoma Radiotherapy Prediction

Jiabo SHENG  
jia-bao.sheng@connect.polyu.hk  
The Dept. of Health Technology and Informatics  
Research Institute for Smart Ageing  
The Hong Kong Polytechnic University  
Hong Kong, China

Zhe LI  
lizhe.li@connect.polyu.hk  
The Dept. of Electrical and Electronic Engineering  
The Hong Kong Polytechnic University  
Hong Kong, China

SaiKit LAM  
saikit.lam@polyu.edu.hk  
The Dept. of Biomedical Engineering  
Research Institute for Smart Ageing  
The Hong Kong Polytechnic University  
Hong Kong, China

Jiang ZHANG  
jiang.zhang@connect.polyu.hk  
The Dept. of Health Technology and Informatics  
The Hong Kong Polytechnic University  
Hong Kong, China

Xinzhi TENG  
xinzhi.teng@connect.polyu.hk  
The Dept. of Health Technology and Informatics  
The Hong Kong Polytechnic University  
Hong Kong, China

Yuanpeng ZHANG  
y.p.zhang@ieee.org  
The Dept. of Health Technology and Informatics  
Shenzhen Research Institute  
The Hong Kong Polytechnic University  
Hong Kong, China

Jing CAI  
jing.cai@polyu.edu.hk  
The Dept. of Health Technology and Informatics  
Research Institute for Smart Ageing  
The Hong Kong Polytechnic University  
Hong Kong, China

ABSTRACT

The accurate prediction of adaptive radiation therapy (ART) for nasopharyngeal carcinoma (NPC) patients before radiation therapy (RT) is crucial for minimizing toxicity and enhancing patient survival rates. Owing to the complexity of the tumor micro-environment, a single high-resolution image offers only limited insight. Furthermore, the traditional softmax-based loss falls short in quantifying a model’s discriminative power. To address these challenges, we introduce a supervised multi-view contrastive learning approach with an additive margin (MMCon). For each patient, we consider four medical images to form multi-view positive pairs, which supply supplementary information and bolster the representation of medical images. We employ supervised contrastive learning to determine the embedding space, ensuring that NPC samples from the same patient or with the same labels stay in close proximity while NPC samples with different labels are distant. To enhance the discriminative ability of the loss function, we incorporate a margin into the contrastive learning process. Experimental results show that this novel learning objective effectively identifies an embedding space with superior discriminative abilities for NPC images.

KEYWORDS

Medical Image Analysis, Multi-view, Nasopharyngeal Carcinoma, Contrastive Learning

1 INTRODUCTION

The planning of intensity-modulated radiotherapy (IMRT) for nasopharyngeal carcinoma (NPC) relies on medical imaging guidance. Previous research has indicated that the geometry of both target volume (TV) and organs-at-risk (OAR) depicted in images can undergo significant changes throughout the IMRT process [2, 11]. To minimize unnecessary radiation exposure during treatment, it is crucial to incorporate medical image analysis into the clinical workflow, aiding physicians in determining the necessity of adaptive radiotherapy (ART).

In contrast to other medical image classification tasks, such as tumor identification [14, 15], and cancer diagnosis [1, 17], the prediction task for NPC ART focuses on analyzing tumor properties to discern the necessity for short-term radiotherapy replanning. Owing to tumor heterogeneity [5, 13], the volume, shape, and texture of tumor regions may differ among patients, with numerous factors potentially contributing to these variations.

In a previous study, [12] employed manually extracted magnetic resonance imaging (MRI) features to investigate the planning of radiation therapy for NPC. However, a single manually extracted omics signature is insufficient to fully capture the information of NPC samples [9]. Existing sample learning methods [9] have demonstrated that manually extracted multi-omics feature representation...
The resurgence of contrastive learning has led to significant advancements in image representation learning [16, 19–21]. The study by [16] showcased the benefits of data augmentation based on various views of the same sample for enhancing visual representation. However, [3, 10] demonstrated limitations of softmax-based contrastive learning loss in classification tasks. This is because softmax loss excels at optimizing inter-class differences (i.e., separating different classes) but falls short in minimizing intra-class variations (i.e., consolidating features of the same class). Although their work proposes a framework to enhance classification performance in contrastive learning, the learned features prove separable for closed-set classification problems. Still, they lack sufficient discriminative power for open-set cancer medical image recognition challenges.

To alleviate the above challenge, we introduce multiview margin contrastive (MMCon) learning, a medical image representation learning approach. As illustrated in Figure 1, given a set of NPC image views, a deep representation is learned by drawing views of the same class patient closer together in the embedding space while simultaneously pushing views of different class patients further apart. We present an example of a learned representation for 4-views (T1, T2, CT, and Dose). The embedding vector for each view is concatenated to create a comprehensive representation of a patient.

The main contributions can be summarized as follows:

- We incorporated multi-view medical images to develop an NPC representation, which aims to maximize the mutual information between different views of the same class patient by utilizing positive pairs from various views.
- We introduce a margin between distinct target class regions to enhance the discriminative capability for samples with unclear boundaries, achieved by extending the conventional contrastive learning loss. With this additional margin, MM-Con demonstrates improved discriminative power and noise tolerance within the embedding space.

## 2 DATASET COLLECTION

We collected samples from 502 NPC patients who underwent radiotherapy in Hong Kong to create the NPC-GTV dataset. Each patient has four different views, including CECT-T1w (T1 image), T2 MR (T2 image), CT images, and dose. All planning images were retrospectively collected in Digital Imaging and Communications in Medicine (DICOM) format and archived using an image archiving and communication system (PACS). Patients with clinical records indicating the need for ART implementation were labeled as 1; otherwise, patients were labeled as 0. The statistics of the NPC-GTV dataset are presented in Table 1.

The imaging data comprised planning CT images and pretreatment T1 and T2 MR images. The treatment-related data included dose fractionation schemes. The outcome data encompassed replanning status and any replanning-related medical records. Attending radiation oncologists input all enrolled patients’ clinical records, which were meticulously examined to determine the binary prediction outcome for this study. All CT and MR images were resampled to a voxel size of 1x1x1 mm³ to minimize the impact of differing image acquisition parameters among patients.

### Table 1: The statistic of NPC-GTV dataset.

| Organ | Views | Non-necessitating ART | Necessitating ART | Samples | Total images |
|-------|-------|-----------------------|-------------------|---------|--------------|
| GTVs  | T1    | 364                   | 138               | 502     | 2,008        |
|       | T2    | 364                   | 138               | 502     |              |
|       | CT    | 364                   | 138               | 502     |              |
|       | Dose  | 364                   | 138               | 502     |              |

## 3 METHODOLOGY

In this section, we first introduce the representation learning framework utilized in this study. Next, we propose a margin contrastive loss function with enhanced discriminative capabilities. Finally, we conclude by comparing the framework in this study and the contrastive loss function’s classification ability with multi-view to previous work. Our objective is to train a feature embedding network using labeled medical images. Embeddings for patient samples with same diseases should be close to each other, while those from patients with different diseases should be far apart. The framework is illustrated in Figure 2.

### 3.1 Representation Learning Network

Given a batch of input samples, we construct positive samples using different types of NPC images (T1, T2, CT, Dose) of the same organ, considering them as multi-view medical images. The patient features of the embedding vector should remain consistent across various viewpoints for the same sample, while embeddings from different samples should differ. As illustrated in Figure 1, the multi-view samples are input to the encoder network. At the network’s output, a margin contrastive loss is computed.
Multi-view Data We associate each input query sample with three distinct medical image views, each offering a unique perspective of the data. The positive samples are collected from the T1, T2, CT, and Dose images from the same patient and samples of the same class, as opposed to self-supervised learning in which positive samples come from the anchor’s augmented data only. Samples from different classes of patients are considered negative samples. We denote these views as \( M \), where \( M = V_1, V_2, \ldots, V_m \).

Encoder Network Our objective is to train an encoder network \( f_0(\cdot) \) from a set of labeled images \( X = \{x_1, x_2, \ldots, x_i\} \), \( f_0(\cdot) \) transforms the input image \( x_i \) to a low-dimensional embedding vector \( h_i = f_0(x_i) \in \mathbb{R}^d \), where \( d \) is the output dimension. Both original and different view samples are independently fed into the same type of encoder, yielding in four representation vectors.

3.2 Contrastive Loss Function

3.2.1 Supervised Contrastive Loss. Supervised contrastive loss (SupCon)[8] can handle the situation where multiple samples are known to belong to the same class due to the presence of labels:

\[
L_{\text{SupCon}} = \sum_{i=1}^{N} \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in \mathcal{A}(i)} \exp(z_i \cdot z_a / \tau)}
\]

In Eq. 1, \( \mathcal{P}(i) \) contains the indices of positive samples in the augmented batch (original + multi-view data) concerning \( x_i \) and \( |\mathcal{P}(i)| \) is the cardinality of \( \mathcal{P}(i) \). \( z_i \) is an anchor. It belongs to the query samples. \( z_a \) are negative samples, \( z_p \) are positive samples, and \( \mathcal{A}(i) \) is the index set of negative samples.

3.2.2 Angular Margin based Contrastive Learning. In Eq. 1, the angular \( \theta_{i,p} \) is as follows:

\[
\theta_{i,p} = \arccos \left( \frac{z_i \cdot z_p}{\|z_i\| \cdot \|z_p\|} \right)
\]

(2)

The decision boundary for \( z_i \) is \( \theta_{i,p} = \theta_{i,a} \). A tiny perturbation around the decision boundary may result in an inaccurate conclusion if an insufficient decision margin is present. To alleviate the problem, we proposed a new training objective for representation learning by adding an additive angular margin \( m \) between positive pair \( z_i, z_p \), and negative pair \( z_i, z_a \), which can be formulated as follows:

\[
L_{\text{MarginCon}} = \sum_{i=1}^{N} \frac{-1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp((\cos \theta_{i,p} - m) / \tau)}{\sum_{a \in \mathcal{A}(i)} \exp((\cos \theta_{i,a} - m) / \tau)}
\]

(3)

In this loss, \( m \) is the margin between the decision boundary, the decision boundary for \( z_i \) is \( \theta_{i,p} = m = \theta_{i,a} \). It increases the compactness of organ feature representation with the same semantics and enlarges the discrepancy of different semantic representations. This help enhances the alignment and uniformity properties, which are two key measures of representation quality related to contrastive learning, indicating how close between positive pair embeddings is and how well the embeddings are uniformly distributed.

3.2.3 Multi-view Margin Contrastive Loss. Let multi-view samples as \( M = \{V_1, V_2, \ldots, V_m\} \), and divide them into three parts, which are query sample representation vector \( z_i \), positive samples representation vectors \( z_p \), and negative samples representation vectors \( z_a \). We bring samples from different views into the \( L_{\text{MarginCon}} \).
We evaluated the effectiveness of the proposed MMCon loss on the NPC-GTV dataset, which consisted of CT, T1, T2, and Dose images that were registered using image registration techniques. For the encoder network, we experimented with three popular architectures, including ResNet [6], Vision Transformer (ViT) [4], and DenseNet [7], and compared three different loss functions: the conventional contrastive learning loss [8], cross-entropy loss, and MMCon loss.

The results in Table 2 indicate that the performance of supervised contrastive learning is the poorest of the three encoders. It partitions the samples into query, positive, and negative groups, and trains their differences. However, the prediction task in this study is a fuzzy boundary classification problem. The results demonstrate that the samples on the fuzzy boundary cannot be well classified even after clustering.

Cross-entropy achieves good performance by effectively utilizing the label information to ensure that samples of the same class are closely clustered. However, using only one view is insufficient for representation in NPC medical images. Therefore, MMCon leverages multi-view information and obtains better results than other losses when combined with each encoder. MMCon adds a margin to the original contrastive learning loss function to ensure a discriminative separation of target and nontarget classes.

### Table 2: The experiment of three different encoders and loss functions

| Encoder       | Loss Function | Accuracy(%) | Precision(%) | Recall(%) | F1(%)  |
|---------------|---------------|-------------|--------------|-----------|--------|
| ResNet50 [6]  | Cross Entropy | 82.97       | 86.21        | 86.21     | 86.23  |
|               | MMCon         | 90.67       | 82.20        | 90.67     | 86.23  |
| DenseNet [7]   | Cross Entropy | 80.79       | 85.94        | 82.76     | 84.32  |
|               | MMCon         | 88.90       | 83.94        | 91.14     | 86.91  |
| ViT [4]        | Supcon        | 80.83       | 75.17        | 79.31     | 77.74  |
|               | Cross Entropy | 86.90       | 80.72        | 89.87     | 85.08  |
|               | MMCon         | 91.28       | 83.42        | 91.33     | 87.20  |

### 4 EXPERIMENT

#### 4.1 Implementation Details

We conducted experiments on the NPC-GTV dataset using K-fold cross-validation with $k = 10$. The mini-batch size for training was set to 50, and the contrastive learning temperature $τ$ was set to 0.07. The learning rate was set to 0.001, and we trained our model for 300 epochs using an SGD optimizer to optimize parameters. For training, we utilized 3 A40 GPUs with 48G memory. We evaluated our proposed MMCon method using accuracy, precision, recall, and F1 score as metrics for binary classification.

To investigate the effect of the margin $m$ in the MMCon loss, we conduct an ablation experiment by varying $m$ from 0 degrees to 1 degree, increasing by 0.1 degrees at each step. We tune the hyperparameter and determine the optimal value of $m$ to be 0.5. This value aligns with our intuition that a small $m$ may have little effect, while a large $m$ may negatively impact the modeling of positive pair relations.

#### 4.2 Results and Analysis

We evaluated the effectiveness of the proposed MMCon loss on the NPC-GTV dataset, which consisted of CT, T1, T2, and Dose images that were registered using image registration techniques.

#### 4.3 Ablation Study

Table 3 presents a comparison between single-view images and multi-view images. In the single-view experiment, only one type of NPC image was used. Three different encoders were used to conduct comparative experiments on T1 single-view, CT single-view, and T1+CT+T2+dose multi-view images under the same loss function. Due to space constraints, we omit the experimental results obtained for T2 single-view and dose single-view images as they do not provide much analytical value. Notably, all three encoders show better results under the multi-view approach than under the single-view approach.

#### 4.4 The Effect of Contrastive Learning

To demonstrate that our proposed contrastive learning tasks can assist the model in learning common features from multimodal input, we conducted a visualization experiment using the NPC-GTV dataset. We employed the TSNE dimensionality reduction approach to generate a 2-dimensional feature vector, which we then visualized, as shown in Figure 3. Figure 3 a depicts the visualization of the fusion result output from our model, while Figure 3 b illustrates the visualization of the cross-entropy output from our model.
Table 3: The experiment of different views by using three different encoders with MMCon. For each encoder result, the best results in each metric are bold.

| Loss   | Encoder | Views         | Accuracy(%) | Precision(%) | Recall(%) | F1(%)  |
|--------|---------|---------------|-------------|--------------|-----------|--------|
|        | MMCon   | ResNet50[6]   | T1 T2 CT Dose | 74.28        | 40.00     | 34.80  | 46.24  |
|        |         | CT            | CT T2 CT Dose | 90.67        | 82.20     | 90.67  | 86.23  |
|        | MMCon   | DenseNet[7]   | T1 T2 CT Dose | 52.70        | 40.32     | 55.17  | 46.59  |
|        |         | CT            | T1 T2 CT Dose | 62.88        | 43.16     | 66.74  | 52.42  |
|        | MMCon   | ViT[4]        | CT           | 64.20        | 50.37     | 58.29  | 54.04  |
|        |         | T1 T2 CT Dose | 91.28        | 83.42        | 91.33   | 87.20  |

As can be seen from Figure 3, contrastive learning effectively increases the distance between positive and negative samples in the vector space, highlighting the degree of data aggregation. This finding indicates that the model can differentiate between data in the vector space based on the characteristics shared by the same-class data, thereby improving its performance. Therefore, contrastive learning can aid the model in acquiring common same-class sample-related traits, thereby improving its ability to generalize to new data.

Figure 3: Cluster visualization of MVSA-Single. Cross entropy loss is typically effective at optimizing the inter-class difference (i.e., separating different classes) but less effective at reducing the intra-class variation (i.e., making features of the same class compact).

5 CONCLUSION

In this study, we propose a classification-capable supervised contrastive representation learning framework. We incorporate multi-view discrimination and an angular margin into the supervised contrastive learning loss to model the NPC image representation, thereby enhancing its discriminative ability. Our experiments demonstrate that our framework outperforms previous baselines on the NPC-GTV dataset, showcasing its efficacy in improving classification accuracy.

ACKNOWLEDGMENTS

This work was supported in part by the Project of RISA (P004001) of The Hong Kong Polytechnic University, Shenzhen-Hong Kong-Macau S&T Program (Category C) (SGDX20201103095002019), Shenzhen Basic Research Program (JCYJ20210524130209023) of Shenzhen Science and Technology Innovation Committee, Project of Strategic Importance (P0035421), the NSF of Jiangsu Province (No. BK20201441), Jiangsu Post-doctoral Research Funding Program (No. 20202020), and the NSFC (Grant No. 82072019).

REFERENCES

[1] Chen Chen, Yong Wang, Jianwei Niu, Xuefeng Liu, Qingfeng Li, and Xuantong Gong. 2021. Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. *IEEE Transactions on Medical Imaging*, 40, 9 (2021), 2439–2451.
[2] Soon-Chool Chung, Mi-Hyun Choi, Hyung-Sil Kim, Na-Rae You, Sang-Pyo Hong, Jung-Chul Lee, Sung-Jun Park, Ji-Hye Baek, H-Jo Jeong, Ji-Hye You, et al. 2014. Effects of distraction task on driving: A functional magnetic resonance imaging study. *Bio-medical materials and engineering*, 24, 6 (2014), 2971–2977.
[3] Son D Dao, Ethan Zhao, Dinh Phung, and Jianfei Cai. 2021. Multi-label image classification with contrastive learning. *arXiv preprint arXiv:2107.11633* (2021).
[4] Alexey Dosovitskiy, Lucas Beyrer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
[5] Nader El-Sayes, Alyssa Vito, and Karen Mosman. 2021. Tumor heterogeneity: A great barrier in the age of cancer immunotherapy. *Cancers*, 13, 4 (2021), 806.
[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 77–86.
[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.
[8] Prannay Khoda, Poit Teterwak, Chen Wang, Aaron Sarra, Yonglong Tian, Phillip Isola, Aaron Maschiotto, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems* 33 (2020), 18661–18673.
[9] Sai-Kit Lam, Yuanpeng Zhang, Jiang Zhang, Bing Li, Jia-Chen Sun, Carol Yee-Tung Liu, Pak-Hei Chou, Xinzi Teng, Zong-Rui Ma, Rui-Yan Ni, et al. 2021. Multi-organ omics-based prediction for adaptive radiation therapy eligibility in nasopharyngeal carcinoma patients undergoing concurrent chemoradiotherapy. *Frontiers in oncology*, 11 (2021).
[10] Zhi Li, Man-Wai Mak, and Helen Mei-Ling Meng. 2022. Discriminative Speaker Representation via Contrastive Learning with Class-Aware Attention in Angular Space. *arXiv preprint arXiv:2210.16622* (2022).
[11] Jie Li, Yidong Ma, Jinbin Chen, Liming Wang, Guifang Zhang, Mukan Zhao, and Yong Yin. 2014. Assessment of anatomical and dosimetric changes from a deformed registration method during the course of intensity-modulated radiotherapy for nasopharyngeal carcinoma. *Journal of radiation research* 55, 1 (2014), 97–104.
[12] Xiangyu Ma, Xinyuan Chen, Jingwen Li, Yu Wang, Kuo Men, and Jianrong Dai. 2020. MRI-only radiotherapy planning for nasopharyngeal carcinoma using deep learning. *Frontiers in oncology*, 11 (2021), 71367.
[13] Andrzej Maruzyk and Kornelia Polvak. 2010. Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1805, 1 (2010), 105–117.
[14] Tarqi Sadad, Amjad Rehman, Asim Munir, Tanzila Sabir, Usman Tarqi, Noor Ayesha, and Rashid Abbasi. 2021. Brain tumor identification using deep learning techniques. *Microscopy Research and Technique* 84, 6 (2021), 1296–1298.
[15] C Saranya, J Geetha Priya, P Jayalakshmi, and E Harini Pavithra. 2021. Brain tumor detection and multi-classification using advanced deep learning techniques. *Microscopy Research and Technique* 84, 6 (2021), 1296–1298.
[16] Zhicheng Zhang, and Cheng Bian. 2022. ProCo: Prototype-Aware Contrastive Analysis for Long-Tailed Medical Image Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 173–182.
[17] Dewen Zeng, Yawen Wu, Xinrong Hu, Xiaowei Xu, Haiyun Yuan, Meiping Huang, Jian Zhan, Jingtong Hu, and Yuyu Shi. 2021. Positional contrastive learning for volumetric medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 221–230.