A Multi-Information Dual-Layer Cross-Attention Model for Esophageal Fistula Prognosis

Jianqiao Zhang[1] [0000-0002-1718-8711], Hao Xiong[2] [0000-0002-6842-1667], Qiangguo Jin[3], Tian Feng[4], Jiquan Ma[5], Ping Xuan[6], Peng Cheng[1], Zhiyuan Ning[7], Zhiyu Ning[7], Changyang Li[8], Linlin Wang[9], and Hui Cui[1] (✉) [0000-0001-8224-4698]

[1] Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia
l.cui@latrobe.edu.au
[2] Centre for Health Informatics, Macquarie University, Sydney, Australia
[3] School of Software, Northwestern Polytechnical University, Shaanxi, China
[4] School of Software Technology, Zhejiang University, Zhejiang, China
[5] Department of Computer Science and Technology, Heilongjiang University, Harbin, China
[6] Department of Computer Science, Shantou University, Shantou, China
[7] School of Computer Science, The University of Sydney, Sydney, Australia
[8] Sydney Polytechnic Institute, Sydney, Australia
[9] Department of Radiation Oncology, Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan, China

Abstract. Esophageal fistula (EF) is a critical and life-threatening complication following radiotherapy treatment for esophageal cancer (EC). Albeit tabular clinical data contains other clinically valuable information, it is inherently different from CT images and the heterogeneity among them may impede the effective fusion of multi-modal data and thus degrade the performance of deep learning methods. However, current methodologies do not explicitly address this limitation. To tackle this gap, we present an adaptive multi-information dual-layer cross-attention (MDC) model using both CT images and tabular clinical data for early-stage EF detection before radiotherapy. Our MDC model comprises a clinical data encoder, an adaptive 3D Trans-CNN image encoder, and a dual-layer cross-attention (DualCrossAtt) module. The Image Encoder utilizes both CNN and transformer to extract multi-level local and global features, followed by global depth-wise convolution to remove the redundancy from these features for robust adaptive fusion. To mitigate the heterogeneity among multi-modal features and enhance fusion effectiveness, our DualCrossAtt applies the first layer of a cross-attention mechanism to perform alignment between the features of clinical data and images, generating commonly attended features to the second-layer cross-attention that models the global relationship among multi-modal features for prediction. Furthermore, we introduce a contrastive learning-

J. Zhang and H. Xiong — Equal first-author contribution.

enhanced hybrid loss function to further boost performance. Comparative evaluations against eight state-of-the-art multi-modality predictive models demonstrate the superiority of our method in EF prediction, with potential to assist personalized stratification and precision EC treatment planning.

Keywords: Multi-modal Data Fusion · Attention Networks · Predictive Model

1 Introduction

Esophageal cancer (EC) was ranked sixth in global cancer mortality, and the five-year survival rate is between 15% and 20% [10]. The main treatments of EC involve chemotherapy and radiation therapy [16, 1, 15], and only few patients have the potential for complete recovery [16]. However, these therapeutic interventions may inadvertently cause a life-threatening complication known as Esophageal Fistula (EF) [21]. Patients diagnosed with EF face a grim challenge, for which the survival time is typically limited to a few months. Therefore, the early identification of EF enables timely intervention and treatment, probably offering patients improved treatment outcomes and enhanced quality of life. In EC treatment planning, Computed Tomography (CT) images and other types of clinical data are often used to facilitate esophageal tumor detection and patient examination. In general, the clinical data includes complementary information on therapy and patient profiles [17]. However, the combination of both images and clinical data substantially augments its complexity and introduces challenges to EF identification. Recently, deep learning have been successfully applied to fields including computer vision and natural language processing [13], with superiority to discover underlying correlations among complex data and build reliable mappings between data and tasks. Hence, EF prognosis using deep learning methods have demonstrated paramount significance and urgency.

Recently, multi-modal data has been exploited by deep learning methods to various disease-related analyses. For instance, Chauhan *et al.* [2] employed a joint function with separate encoders and classifiers to assess pulmonary oedema from chest radiographs and radiology reports. Li *et al.* proposed a multi-modal network combining supervised and unsupervised learning for cancer survival prediction [12] using CT images and clinical data. Meanwhile, Yap *et al.* exploited a late fusion technique for integration of macroscopic images, dermatoscopic images and histopathological diagnosis data to classify skin lesion [20]. Vale-Silva *et al.* proposed a multi-modal deep-learning model aggregating imaging, clinical and molecular data for long-term cancer survival prediction [18]. With the popularity of attention networks [19] a few attention-based integration models have been developed. For instance, with CT images and clinical data, a cross-modal self-attention model [9] was proposed for EC prediction and a graph attention model [5] was designed for lymph node metastasis prediction. Likewise, Fu *et al.* [8] proposed a multi-modal graph-based network for breast cancer survival prediction using imaging mass cytometry and patient variables. The multi-modal transformer model proposed by Zheng *et al.* predicted the survival rate of nasopharyngeal carcinoma patients [22] from CT images, segmentation map and tabular clinical data. Similarly, [23, 24] leveraged multi-modal transformer networks using clinical text data and a mix of imaging data for the visual acuity prediction of cataract surgery and pulmonary disease diagnosis. Due to the heterogeneous nature of data from different modalities, the existing heterogeneity may hinder the learning capacity of the model and thus affect the performance of downstream tasks. However, existing approaches do not explicitly relieve the heterogeneity among multi-modal data.

To address above-mentioned issues, we propose an adaptive multi-information dual-layer cross-attention (MDC) model for EF diagnosis. Specifically, MDC comprises three major components. First, a clinical data encoder is adopted to extract features from tabular clinical data. Second, an adaptive 3D Trans-CNN image encoder is designed to extract latent image features. The image encoder utilizes both CNN and transformer to extract multi-level local and global features. To facilitate the mitigation of heterogeneity, the image encoder further introduces a global depth-wise adaptive fusion module that aims to filter out redundant information within extracted local and global features by identifying the importance of these features and then adaptively fusing them. This is because image is high-dimensional data itself, and both the locally and globally extracted features further introduce more redundant and noisy information. Third, we design a novel dual-layer cross-attention (DualCrossAtt) module to mitigate the cross-modal heterogeneity and fuse the multi-modal features. DualCrossAtt consists of the first layer of a cross-attention mechanism which utilizes clinical data feature as query so as to find its attended areas within image features. By doing that, we are able to align image features and clinical data features, identifying attended regions shared by both features and thus relieving the heterogeneity among them. Concatenated with image features, the identified attended features are then passed to the second layer of cross-attention mechanism to build global relationship of multi-modal features for final detection. To more effectively train the model, we devise the hybrid loss function combining a cross-entropy loss and a contrastive loss. The cross-entropy loss is utilized for our classification task, and the contrastive loss aims to render features from same class more similar while increasing the dissimilarity of those from different classes.

Our contributions can be summarized as: (1) a multi-sourced information strategy that includes features from different views of CT images and integrates clinical data from another modality, thereby augmenting the performance of EF prognosis. (2) an adaptive fusion strategy to generate more effective and unified features of image modality by combining its local and global information. (3) a new dual-layer attention mechanism aims to align the cross-modal features to extract attended features across all modalities and meanwhile model the global relationship among multi-modal features. (4) the hybrid loss with contrastive learning further enhances our performance. Comprehensive validation against eight state-of-the-art (SOTA) multi-modal predictive models and ablation study demonstrated the effectiveness of our technical innovations and contributions.

Our work is critical as early prediction of this radiation therapy caused EF enables to develop more personalized treatment plans, thereby improving the quality of life for patients with EC.

2 Material and Method

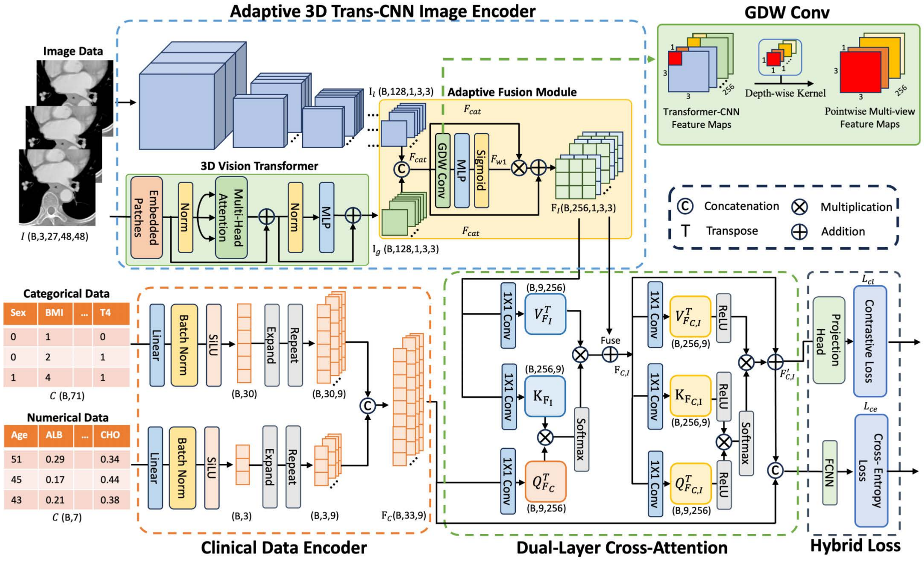
2.1 Dataset

The dataset includes 553 EC patients collected from Shandong Cancer Hospital between 2014 and 2019. This study was approved by the ethical committee of Shandong Cancer Hospital, China. There are 367 patients diagnosed with EF after radiotherapy treatment, while the remaining 186 cases are the patients without diagnosed EF. This dataset provides both clinical data and CT images. We cropped the original CT scans into individual cubes of size 201 pixels × 201 pixels × 27 slides at the centre of the tumour, and these cubes were then resized to 48 pixels × 48 pixels × 27 slides as the input of our model. The clinical data has general information such as patients' ID, age, gender, drinking, and smoking history. It also contains medical treatment details, including radiation dose types and volumes, history of radiotherapy and inhibitor therapy. In total, there are 34 variables in clinical data, with 7 numerical variables and 27 categorical variables. One-hot embedding was then applied to categorical variables, resulting in 71-dimensional categorical data. As a result, the input clinical data has 78 dimensions including 7 numerical and 71-dimentional categorical data. To prevent overfitting, we applied data augmentation to the training data by shifting along different axes with [-5, 0, 5] pixels, generating 3142 samples with 1674 EF cases and 1468 non-EF ones.

2.2 MDC Esophageal Fistula Prognosis Method

The overall architecture of our MDC method is shown in Fig. 1. Briefly, the clinical data encoder and the adaptive 3D Trans-CNN image encoder aim to extract features from tabular clinical data and CT images, respectively. After that, the DualCrossAtt module aligns the extracted multi-modal features to relieve the heterogeneity across multi-modal data and also builds long range dependencies among them. To further enhance accuracy, we present the hybrid loss by integrating the cross-entropy loss with the contrastive learning.

Clinical Data Encoder and Adaptive 3D Trans-CNN Image Encoder. The clinical data encoder takes tabular clinical data $C \epsilon R^{D_c}$ as inputs, where $D_c = 78$ refers to the dimension of clinical data. Since the clinical data includes numerical variables and categorical variables, we utilize two separate encoders to process these two types of variables. Here, each encoder consists of a linear layer, a batch normalisation layer and a SiLU activation function. To generate clinical data features F_c, the features extracted from these two types of variables are then expanded, duplicated and concatenated along channels.



Sex	BMI	...	T4
0	1		0
0	2		1
1	4		1

Age	ALB	...	CHO
51	0.29		0.34
45	0.17		0.44
43	0.21		0.38

Fig. 1. Our method consists of the clinical data encoder, the adaptive 3D Trans-CNN image encoder, the dual-layer cross-attention (DualCrossAtt) module and a hybrid loss. The encoders aim to extract features of data from their corresponding modalities, and the dual-layer cross-attention module aligns the extracted multi-modal features and also builds the global relationship among multi-modal data. The hybrid loss with contrastive learning further enhances the capacity of our model.

Compared to low-dimensional clinical data, the 3D CT images $I \epsilon R^{D_I \times H \times W}$ contain much richer information. Here, D_I, H, W are respectively the depth, height and width of the image. We therefore propose the adaptive 3D Trans-CNN image encoder, including a CNN encoder based on the U-Net [11] and a Vision Transformer [6], to extract image features from multiple scales and views. That is, the CNN encoder focuses on the local features $I_l \epsilon R^{C \times D_I \times H \times W}$ (C is the number of channels), whilst the transformer can capture global features $I_g \epsilon R^{C \times D_I \times H \times W}$. We then perform channel-wise concatenation on I_l to I_g generate $F_{cat} \epsilon R^{2 \times C \times D_I \times H \times W}$.

As F_{cat} is the feature fused from multi-views, which is likely to contain redundant information, we introduce an adaptive fusion module into the 3D Trans-CNN image encoder so that the encoder is able to extract more effective multi-view features with less redundancy. To achieve that, the adaptive fusion module involves a Global Depth-wise Convolution (GDWConv) [4] layer, a Linear layer and Sigmoid activation function:

$$F_{w1} = Sigmoid(Linear(GDWConv(F_{cat}))) \tag{1}$$

In Eqn. 1, F_{w1} represents the channel-wise weight of F_{cat}, for which the weight indicates the importance of feature and its value range is between 0 and 1. It is also noteworthy that we exploit GDWConv rather than ordinary convolutions to better utilize the features across different views in F_{cat} for weights calculation. Next, we multiply F_{cat} by the weights F_{w1} to identify more important features and add back the results to F_{cat}:

$$F_1 = F_{cat} + F_{cat} \times F_{w1} \tag{2}$$

Dual-Layer Cross-Attention (DualCrossAtt) Module. The features of the two modalities are inherently different, since they provide patient information from different perspectives. We therefore propose the DualCrossAtt module exploiting a first layer of cross-attention to mitigate the discrepancy among multi-modal features and take a further step to model the global relationship of multi-modal features with a second layer of cross-attention. Unlike a single attention mechanism, our dual-layer cross-attention module strengthens both the consistency and global dependency among multi-modal features. For the first layer cross-attention, we use the clinical data feature Q_{F_C} as query, image features K_{F_I}, V_{F_I} as key and value. By doing that, it performs alignment of clinical data and images, allowing the clinical data to find its consistent information in image features. As a result, the attended regions generated by the first layer cross-attention are the important features shared by both modalities:

$$F_{C,I} = Softmax(Q_{F_C}^T \times K_{F_I}) \times V_{F_I}^T + F_I \tag{3}$$

where Q_{F_c} is obtained by applying the convolution operation to F_c. Similarly, we utilize two convolutions upon F_I to obtain K_{F_I} and V_{F_I}, respectively.

In Eqn. 3, the output $F_{C,I}$ of first layer cross-attention contains information of image features and the attended features of both modalities. To better exploit $F_{C,I}$, we introduce the second layer cross-attention to strengthen the long-range dependency among $F_{C,I}$. Specifically, we apply three separate convolutions to $F_{C,I}$, obtaining the query $Q_{F_{C,I}}$, the key $K_{F_{C,I}}$ and the value $V_{F_{C,I}}$.

$$F'_{C,I} = Softmax(Q_{F_{C,I}}^T \times K_{F_{C,I}}) \times V_{F_{C,I}}^T + F_{C,I} \tag{4}$$

Here, $Q_{F_{C,I}}$, $K_{F_{C,I}}$ and $V_{F_{C,I}}$ are from the same source $F_{C,I}$. In essence, the second layer cross-attention is the self-attention [19, 14]. The generated $F'_{C,I}$ with strengthened global relationship is utilized for final prediction.

Hybrid Loss. Our hybrid loss involves a cross-entropy loss and a contrastive loss as:

$$L_h = L_{ce} + \omega L_{cl} \tag{5}$$

where $\omega = 1$ refers to the weight, L_{ce} is the cross-entropy loss for the learning of classification task. Besides, we add the contrastive loss L_{cl} that aims to enhance the clustering degree of the features. This is primarily because different patients tend to have varied image appearances, textures and clinical data values, even if they are from the same group (e.g. patients with EF). As a result, the features extracted from CT images and clinical data are also varied among these patients.

To facilitate classification, we need to learn more unified feature representations for each class, so that features from the same class are clustered together and those from different classes are kept as far as possible. Inspired by current studies of contrastive learning [3], we introduce the contrastive loss as:

$$L_{cl} = \Sigma_{i=1}^{p}((1-y^i)\frac{1}{2}\left\|f(x_1^i)-f(x_2^i)\right\|_2 + y^i\frac{1}{2}(max(0,m-\left\|f(x_1^i)-f(x_2^i)\right\|_2)) \tag{6}$$

where p represents the number of training pairs, y^i is a binary label indicating if the pair of features x_1^i and x_2^i is from the same class or not. Meanwhile, $f()$ is the fully-connected layer based projection head that reduces the dimension of features for ease of contrastive learning, and $\|\|_2$ measures the l2 distance between projected features. In addition, $m > 0$ is the margin that the measured distance less than it will contribute to the contrastive loss.

3 Experiments, Results and Discussions

3.1 Implementation Details

We implemented our model in PyTorch framework and trained it using one NVIDIA Corporation GV100 GPU. The Adam optimizer was adopted with learning rate of 0.01. The batch size was set to 50, and the total number of epochs was 250. We evaluated all experimental methods using the same 5-fold cross validation in which each fold contains 4/5 and 1/5 of data as the training and test sets, respectively. Out of the training set, 1/5 of data was utilized as the validation set.

3.2 Comparison Methods and Evaluation Measures

We compare our method to eight state-of-the-art baselines, including the image based method Lungnet [7], and multi-modality based methods – Chauhan [2], Yap [20], MultiSurv [18], VisText [9], IRENE [23], MSA [24] and DMGN [8]. We performed evaluation in terms of Area Under the Curve (AUC), Balanced Accuracy (BAC), F1 Score, and Sensitivity (SEN).

3.3 Comparison Results

As shown by results in Table 1, our proposed method outperforms the other baselines with respect to all metrics.

It is noteworthy that AUC is a comprehensive metric that considers both sensitivity and specificity. To better illustrate it, we also show the comparison of ROC curves among our method and the other baselines in Fig. 2. It can be seen that the ROC curve of our method with the AUC score of 75.1 is consistently above the ROC curves of all the other baselines, gaining 6.4% improvement over

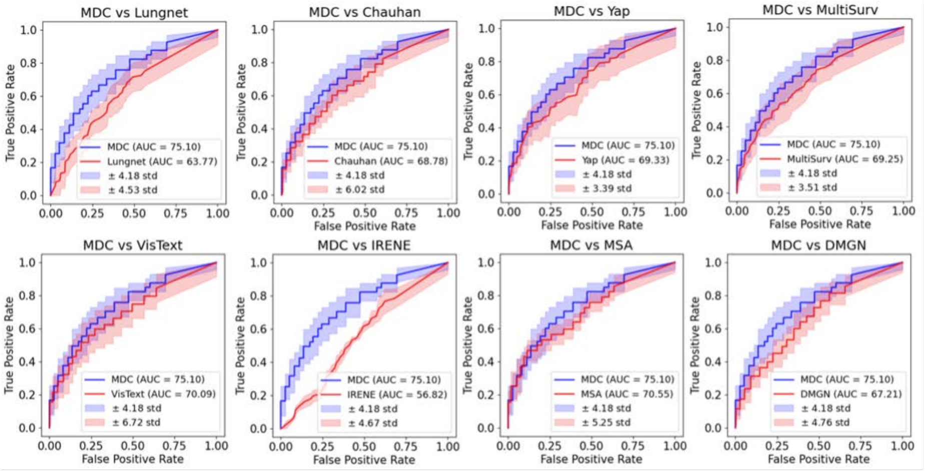


Fig. 2. Mean ROC curves of 5-fold cross validation on all the 8 compared methods with mean AUC values ± standard deviation (std) on the EF prognosis.

Table 1. Quantitative comparison of our method and 8 baselines on EF prognosis. We utilize mean value ± standard deviation to show the variances in 5-fold cross validation.

Method	AUC(%)	BAC(%)	F1(%)	SEN(%)
Lungnet [7]	$63.77_{\pm 04.53}$	$54.87_{\pm 04.99}$	$26.79_{\pm 04.44}$	$25.56_{\pm 28.31}$
Chauhan [2]	$68.78_{\pm 06.02}$	$64.63_{\pm 05.16}$	$53.94_{\pm 04.74}$	$54.31_{\pm 09.91}$
Yap [20]	$69.33_{\pm 03.39}$	$59.70_{\pm 01.33}$	$52.31_{\pm 07.11}$	$64.95_{\pm 19.72}$
MultiSurv [18]	$69.25_{\pm 03.51}$	$64.32_{\pm 03.24}$	$56.42_{\pm 06.17}$	$59.31_{\pm 14.66}$
VisText [9]	$70.09_{\pm 06.72}$	$64.73_{\pm 05.81}$	$58.01_{\pm 05.34}$	$55.86_{\pm 12.30}$
IRENE [23]	$56.82_{\pm 04.67}$	$53.43_{\pm 04.74}$	$46.82_{\pm 05.45}$	$58.75_{\pm 13.73}$
MSA [24]	$70.55_{\pm 05.25}$	$65.02_{\pm 02.54}$	$54.34_{\pm 01.97}$	$54.83_{\pm 12.94}$
DMGN [8]	$67.21_{\pm 04.76}$	$62.60_{\pm 05.44}$	$49.32_{\pm 04.41}$	$45.90_{\pm 11.52}$
MDC	$75.10_{\pm 04.18}$	$66.12_{\pm 05.66}$	$60.53_{\pm 11.60}$	$69.41_{\pm 23.91}$

the second-best method MSA [24]. Other than that, we enhance the F1 score, BAC, SEN by 4.3%, 1.7% and 6.9%, respectively.

We posit that the outperformance of our model is mainly attributed to the utilised multi-information and DualCrossAtt module. That is, the multi-sourced information, including local, global image features and the clinical data, integrating more comprehensive features and are clearly superior to those using fewer information. For the multi-modal heterogeneity, unlike other methods that do not explicitly address it, our DualCrossAtt module can relieve it by performing the alignment such that extracting the attended features shared by both modalities to achieve more effective prediction. Also, integrating with the contrastive loss strengthens the clustering effects of the learned features and further enhance the accuracy.

3.4 Ablation Study Results

We investigate how adaptive fusion module, DualCrossAtt and contrastive loss L_{cl} affect the performance of our model. We start with a baseline of our model without all these components and then gradually add each component.

Table 2. Ablation study results of investigating the effect of key components in our MDC model.

Adaptive Fusion	DualCrossAttn		L_{cl}	Results			
	1^{st} layer	2^{nd} layer		AUC(%)	BAC(%)	F1(%)	SEN(%)
				$72.79_{\pm 04.94}$	$64.26_{\pm 03.09}$	$54.05_{\pm 03.79}$	$55.78_{\pm 16.42}$
✓				$73.22_{\pm 05.42}$	$65.46_{\pm 03.95}$	$55.78_{\pm 03.99}$	$53.78_{\pm 20.97}$
✓		✓		$73.69_{\pm 06.48}$	$65.52_{\pm 04.92}$	$55.70_{\pm 05.10}$	$56.98_{\pm 20.58}$
✓	✓	✓		$75.02_{\pm 05.27}$	$65.56_{\pm 03.12}$	$56.93_{\pm 04.47}$	$55.34_{\pm 10.15}$
✓	✓	✓	✓	$75.10_{\pm 04.18}$	$66.12_{\pm 05.66}$	$60.53_{\pm 11.60}$	$69.41_{\pm 23.91}$

As shown in Table 2, adding adaptive fusion module enhances the performance of the baseline model with respect to most metrics. This is due to its ability to fuse more important features of image modality. In terms of AUC, BAC and F1 score, we note that the improvement is not large by combining only the second layer of DualCrossAtt to adaptive fusion module. This is presumably because without first layer attention for the alignment, the misaligned multi-modal features affect the prediction accuracy. This point is best illustrated when we integrate with the whole DualCrossAtt, for which the incorporation of first layer attention substantially improves the accuracy, especially for the metric of AUC. Lastly, our entire model achieves best results after adding contrastive loss L_{cl}.

4 Conclusion

In this work, we propose MDC, a novel deep learning model designed for the prediction of esophageal fistula in esophageal cancer patients, aiming to facilitate the development of more tailored treatment plans. MDC exploits clinical data, CT images with CNN and Transformer to aggregate multi-sourced information. Our model further introduces a dual-layer attention mechanism to align multi-modal data and capture the global relationship, combining a hybrid loss with contrastive learning to enhance the clustering effect of learned features. Experimental evaluations underscore the superiority of our approach over existing state-of-the-art multimodality-based baselines.

Disclosure of Interests. Authors have no competing interests in the paper.

References

1. Borggreve, A.S., Kingma, B.F., Domrachev, S.A., Koshkin, M.A., Ruurda, J.P., van Hillegersberg, R., Takeda, F.R., Goense, L.: Surgical treatment of esophageal cancer in the era of multimodality management. Annals of the New York Academy of Sciences **1434**(1), 192–209 (2018)
2. Chauhan, G., Liao, R., Wells, W., Andreas, J., Wang, X., Berkowitz, S., Horng, S., Szolovits, P., Golland, P.: Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23. pp. 529–539. Springer (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
4. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
5. Cui, H., Xuan, P., Jin, Q., Ding, M., Li, B., Zou, B., Xu, Y., Fan, B., Li, W., Yu, J., et al.: Co-graph attention reasoning based imaging and clinical features integration for lymph node metastasis prediction. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 657–666. Springer (2021)
6. 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)
7. Faruqui, N., Yousuf, M.A., Whaiduzzaman, M., Azad, A., Barros, A., Moni, M.A.: Lungnet: A hybrid deep-cnn model for lung cancer diagnosis using ct and wearable sensor-based medical iot data. Computers in Biology and Medicine **139**, 104961 (2021)
8. Fu, X., Patrick, E., Yang, J.Y., Feng, D.D., Kim, J.: Deep multimodal graph-based network for survival prediction from highly multiplexed images and patient variables. Computers in Biology and Medicine **154**, 106576 (2023)
9. Guan, Y., Cui, H., Xu, Y., Jin, Q., Feng, T., Tu, H., Xuan, P., Li, W., Wang, L., Duh, B.L.: Predicting esophageal fistula risks using a multimodal self-attention network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 721–730. Springer (2021)
10. Hirano, H., Boku, N.: The current status of multimodality treatment for unresectable locally advanced esophageal squamous cell carcinoma. Asia-Pacific Journal of Clinical Oncology **14**(4), 291–299 (2018)
11. Jin, Q., Meng, Z., Sun, C., Cui, H., Su, R.: Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. Frontiers in Bioengineering and Biotechnology **8**, 605132 (2020)
12. Li, K., Chen, C., Cao, W., Wang, H., Han, S., Wang, R., Ye, Z., Wu, Z., Wang, W., Cai, L., et al.: Deaf: A multimodal deep learning framework for disease prediction. Computers in Biology and Medicine **156**, 106715 (2023)
13. Mahesh, B.: Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet] **9**(1), 381–386 (2020)

14. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
15. Short, M.W., Burgers, K.G., Fry, V.T.: Esophageal cancer. American family physician **95**(1), 22–28 (2017)
16. Tsushima, T., Mizusawa, J., Sudo, K., Honma, Y., Kato, K., Igaki, H., Tsubosa, Y., Shinoda, M., Nakamura, K., Fukuda, H., et al.: Risk factors for esophageal fistula associated with chemoradiotherapy for locally advanced unresectable esophageal cancer: a supplementary analysis of jcog0303. Medicine **95**(20), e3699 (2016)
17. of Uveitis Nomenclature (SUN) Working Group, S., et al.: Standardization of uveitis nomenclature for reporting clinical data. results of the first international workshop. American journal of ophthalmology **140**(3), 509–516 (2005)
18. Vale-Silva, L.A., Rohr, K.: Multisurv: Long-term cancer survival prediction using multimodal deep learning. MedRxiv pp. 2020–08 (2020)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
20. Yap, J., Yolland, W., Tschandl, P.: Multimodal skin lesion classification using deep learning. Experimental dermatology **27**(11), 1261–1267 (2018)
21. Zhang, Y., Li, Z., Zhang, W., Chen, W., Song, Y.: Risk factors for esophageal fistula in patients with locally advanced esophageal carcinoma receiving chemoradiotherapy. OncoTargets and therapy pp. 2311–2317 (2018)
22. Zheng, H., Lin, Z., Zhou, Q., Peng, X., Xiao, J., Zu, C., Jiao, Z., Wang, Y.: Multi-transsp: Multimodal transformer for survival prediction of nasopharyngeal carcinoma patients. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–243. Springer (2022)
23. Zhou, H.Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., Shao, J., Lu, G., Zhang, K., Li, W.: A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nature Biomedical Engineering **7**(6), 743–755 (2023)
24. Zhou, Q., Zou, H., Jiang, H., Wang, Y.: Incomplete multimodal learning for visual acuity prediction after cataract surgery using masked self-attention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 735–744. Springer (2023)