LLM-guided Multi-modal Multiple Instance Learning for 5-year Overall Survival Prediction of Lung Cancer

Kyungwon Kim[1][0009−0005−4760−8886], Yongmoon Lee[2], Doohyun Park[1][0000−0001−6237−468X], Taejoon Eo[1][0000−0002−3546−0184], Daemyung Youn[3], Hyesang Lee[3], and Dosik Hwang[1,4,5,6][0000−0002−2217−2837]

[1] School of Electrical and Electronic Engineering, Yonsei University, Seoul, Republic of Korea
{yskgw93, dosik.hwang}@yonsei.ac.kr

[2] Department of Pathology, Dankook University School of Medicine, Cheonan, Republic of Korea

[3] DoctorWorks Co., Ltd., Seoul, Republic of Korea

[4] Department of Radioogy and Center for Clinical Imaging DataScience, Yonsei University, Seoul, Republic of Korea

[5] Department of Oral and Maxillofacial Radiology, College of Dentistry, Yonsei University, Seoul, Republic of Korea

[6] Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology, Seoul, Republic of Korea

Abstract. Accurately predicting the 5-year prognosis of lung cancer patients is crucial for guiding treatment planning and providing optimal patient care. Traditional methods relying on CT image-based cancer stage assessment and morphological analysis of cancer cells in pathology images have encountered challenges in terms of reliability and accuracy due to the complexity and diversity of information within these images. Recent rapid advancements in deep learning have shown promising performance in prognosis prediction, however utilizing CT and pathology images independently is limited by their differing imaging characteristics and the unique prognostic information. To effectively address these challenges, this study proposes a novel framework that integrates prognostic capabilities of both CT and pathology images with clinical information, employing a multi-modal integration approach via multiple instance learning, leveraging large language models (LLMs) to analyze clinical notes and align them with image modalities. The proposed approach was rigorously validated using external datasets from different hospitals, demonstrating superior performance over models reliant on vision or clinical data alone. This highlights the adaptability and strength of LLMs in managing complex multi-modal medical datasets for lung cancer prognosis, marking a significant advance towards more accurate and comprehensive patient care strategies. The code is publicly available on https://github.com/KyleKWKim/LLM-guided-Multimodal-MIL.

Keywords: Survival Prediction · Multi-modality · Multiple Instance Learning · Large Language Model · Lung Cancer.

1 Introduction

Recently, with advancements in early diagnosis and treatment of cancer, the survival rates for cancer patients have seen a gradual increase. For non-small cell lung cancer, the likelihood of surviving more than 5 years post-diagnosis was previously very small, under 20%. However, with appropriate surgery or treatment, this survival rate can soar to over 60% [7,12]. Thus, accurately predicting prognosis and determining the risk for cancer patients to provide optimal treatment options is very important.

Treatment for lung cancer patients is determined depending on the stage, the type of carcinoma, and other patient-specific circumstances. The TNM stage, which is derived from the analysis of CT image and biopsy for pathological examination, plays the most representative role in prognosis determination, where T, N, and M stages refer to the size or direct extent of the primary tumor, the degree of spread to regional lymph nodes, and the presence of distant metastasis, respectively [1]. However, the translation of key image-based cancer features into simple text-based clinical information has led to a loss of image domain information. Therefore, accurate precision medicine should consider not only TNM stage information but also CT image analysis to utilize the tumor's morphological characteristics, as well as the properties of cancer cells visible in histopathological images [13].

To address the above problems, research attempts to assess the condition of cancer patients and predict their prognosis using not only clinical information but also CT and histopathological images. Recent developments in deep learning technology have demonstrated its efficacy in automated quantitative image analysis. Research using the vision transformer (ViT) [3,5] model for CT-based overall survival prediction have achieved an AUC of 0.78 [9]. Other research incorporating non-image clinical information alongside CT images has shown an AUC of 0.76 [21]. The first attempt to predict prognosis using pathology images through deep learning was by DeepConvSurv [22], achieving an AUC of 0.64. Although this was considered low, it marked the first step and set the baseline for future models. Following this, DeepCorrSurv [20] introduced a multi-modal approach that included molecular data. However, these methods are not as effective because they do not consider both CT and pathological images, which limits their accuracy in the detailed prognosis prediction process typically conducted in clinical settings.

This study aimed to extract sufficient information from both CT and pathology images to predict the prognosis of lung cancer patients more accurately. Similar to clinical prognosis prediction process, we enhanced our approach by incorporating clinical information, thus effectively extracting data on TNM stages and the morphology of cancer cells from CT and pathology images. With the advance of Large Language Models (LLMs) [14,16], which offer rich semantic language representations, we utilized LLM to interpret clinical information and connect its relevance to the CT and pathology images. The features extracted from CT and pathology, guided by clinical information, were integrated using

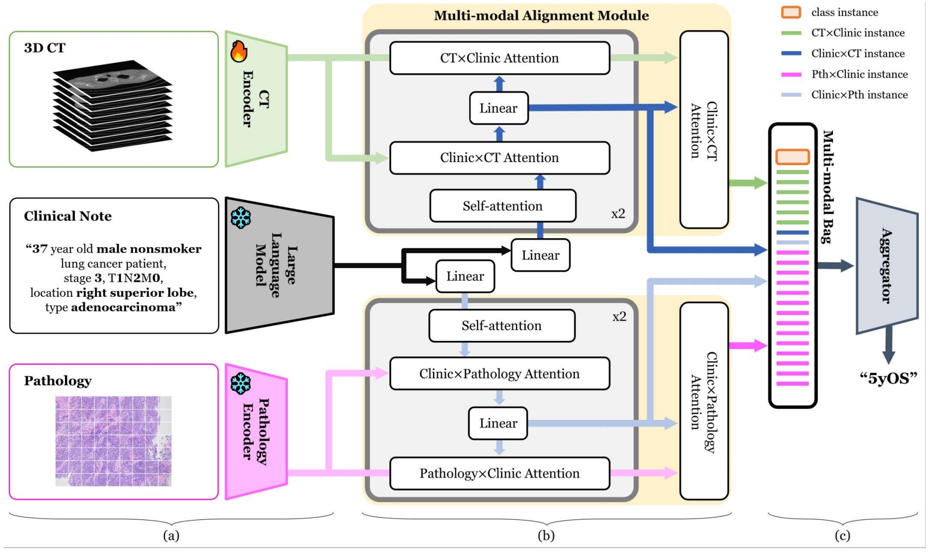

Fig. 1: Overview of our proposed multi-modal MIL framework. Our model consists of 3 steps: (a) encoding features, (b) aligning the encoded features with LLM-guidance, and (c) aggregating features. The name of sub-module in (b), Modality1×Modality2 Attention, means the cross-attention of Modality1 embedding (as query) to Modality2 embedding.

multiple instance learning techniques, achieving an AUC of 0.877, $p \leq 0.05$, and a recall of 0.964, on the external validation set.

2 Method

Fig. 1 illustrates the overview of our proposed framework for LLM-guided multi-modal multiple instance learning (MIL) for 5-year overall survival prediction.

2.1 Multiple Instance Learning

The binary MIL problem is predicting the label $Y \in \{0,1\}$ for a bag of instances of X. In a situation where the instance-level label, $\{y_1, ..., y_n\}$, for each instance is unknown, a bag can be considered positive if it contains at least one positive instance, a prediction $\hat{Y} = \mathbf{S}(X)$ where $\mathbf{S}$ is a scoring function.

The basic premise is that each instance must be independent and identically distributed (i.i.d.) [11]. However, in many real-world environments, it is very difficult to construct a bag with instances that satisfy the premise, and it has been proven that the MIL problem persists even when there is a correlation between instances [15].

MIL is usually used to process pathology images with large resolution [4,10,15], but it has been expanded and applied to a multi-modality bag rather than a pathology-only bag. In other words, each slice of CT, each patch of pathology, and clinical note were regarded as one instance, forming a multi-modality bag. For the defined instances for each modality, $X^{CT} = \{x_1^{CT}, ..., x_{N_{CT}}^{CT}\}$ and $X^{Pth} = \{x_1^{Pth}, ..., x_{N_{Pth}}^{Pth}\}$, the entire instance set becomes $X = X^{CT} \cup X^{Pth} \cup x^{text}$, and $\mathbf{S}$ can be expressed for multi-modal instances as follows.

$$M = \{\mathbf{E}(x_i)\}_{i=1}^{N} \tag{1}$$

$$Z^{image}, Z^{text} = \mathbf{MAM}(M^{image}, \mathbf{fc}(M^{text})) \tag{2}$$

$$\mathbf{S}(X) = \mathbf{G}(Z) \tag{3}$$

where $\mathbf{E}$ is a feature encoder from each modality, $\mathbf{MAM}$ is a multi-modality alignment module, $\mathbf{fc}$ is a single fully-connected(fc) layer and $\mathbf{G}$ is a feature aggregator. M and $Z = Z^{CT} \cup Z^{Pth} \cup Z^{text}$ refer to encoded feature maps that passed $\mathbf{E}$ and $\mathbf{MAM}$, respectively, and $Z^{text \times CT}$ and $Z^{text \times Pth}$ represent encoded text features aligned to CT and pathology image, respectively. N represents the number of instances per each modality.

2.2 Feature Encoders

CT Encoder For the 3D volume CT, we utilized a 3D convolution-based model structure, the 3D ResNet (MC3) [17], as the encoder. To preserve the number of instances in M^{CT}, we only used up to the layer before 3D global average pooling in MC3. Thus, for the input $X^{CT} \in \mathbb{R}^{1 \times N_{slice} \times H_{CT} \times W_{CT}}$, we perform 2D global average pooling over the spatial dimensions of output feature of MC3 which has the size of $D \times N_{slice} \times H'_{CT} \times W'_{CT}$ to create a set of 1D vector instances of length D which is identical for all modalities, denoted as $M^{CT} \in \mathbb{R}^{N_{slice} \times D}$.

Pathology Encoder Patch-wise processing of pathology images involves feature encoding for each patch. Considering computational costs, it is common to use the patch feature encoder in a frozen state. In this study, we used CTransPath [19], which is a state-of-the-art pretrained feature extractor trained via unsupervised contrastive learning on H&E stained pathological images. It extracts a feature vector of size $1 \times D_0$ for each input $x_i^{Pth} \in \mathbb{R}^{3 \times H_{patch} \times W_{patch}}$. To align these features with those from other modalities, we add an additional fc layer, resulting in $M^{Pth} \in \mathbb{R}^{N_{patch} \times D}$.

Text Encoder Recent advancements have seen the emergence of LLMs capable of capturing the deep semantic meanings, nuances, and relationships within text. We utilized the pretrained text encoder of CLIP [14] as our text encoder. Although CLIP, a leading model in the Visual-Language Model domain, is primarily specialized in visual-language associative learning, its pretrained text encoder possesses the complex capability required for processing and understanding language data, thus fulfilling the role of an LLM. Our clinical information, consisting of a single sentence, is converted into a list of integers through a tokenizer from CLIP and then transformed into $M^{text} \in \mathbb{R}^{1 \times D}$ by the text encoder.

2.3 Multi-modal Alignment Module (MAM)

To efficiently extract information from CT and pathology image modalities, **MAM** employs text features derived via LLM, providing text prompts for mapping image features, closely mirroring the approach of the Segment Anything Model [8]. This foundation model for segmentation maps image and prompt embeddings using a lightweight mask decoder which is a strategy we adapt by incorporating its modified transformer structure with omitting the mask generator. In the **MAM**, we set the dimension of each modality's encoded features to D, generating LLM-guided features for each modality $Z^{CT} \in \mathbb{R}^{N_{slice} \times D}$ and $Z^{Pth} \in \mathbb{R}^{N_{patch} \times D}$, alongside modality-aligned text features $Z^{text \times CT}, Z^{text \times Pth} \in \mathbb{R}^{1 \times D}$, as in Eq.(2).

2.4 Feature Aggregator

Features updated through the **MAM** are concatenated back into the bag with uniform dimensions. An aggregator **G** has a structure that repeats the self-attention mechanism of the transformer [18] twice with additional class token where the information of the entire bag is summarized. Subsequently, the class token is passed through the fc layer to determine the final survival output.

3 Experiments and Results

3.1 Dataset

We use lung cancer prognosis prediction fusion datasets from 'The Open AI Dataset Project (AI-Hub, S. Korea)', and all data information can be accessed through 'AI-Hub'. The dataset was collected from a total of 10,000 patients at 6 university hospitals. While the clinical information is organized into approximately 40 categories, only essential information that is necessarily obtained from patients was selected for use in the study. The refined clinical information includes nine specific items: sex, age, smoking history, overall TNM stage, T stage, N stage, M stage, location of cancer, and type of cancer cells. Pathology images were classified into five cancer types, but this research focused only on adenocarcinoma and squamous cell carcinoma, which constitute the majority of lung cancer cases. The pathology images were acquired either through biopsy or resection, with low quality images being excluded. Both CT and pathology images used in the study were taken before treatment.

Data unsuitable for 5-year overall survival (5yOS) analysis, such as those lost to follow-up or whose death was not due to cancer, were excluded. Following these criteria, the total dataset of CT, pathology, and clinical information paired data consisted of 908 cases, with the number of cases from each institution being 294, 267, 133, 111, 67, and 36, respectively. The data from the last three hospitals were combined to comprise 23% of the total dataset, which was designated as the external validation set. Distribution of all clinical indicators has been prepared in the Supplementary Materials.

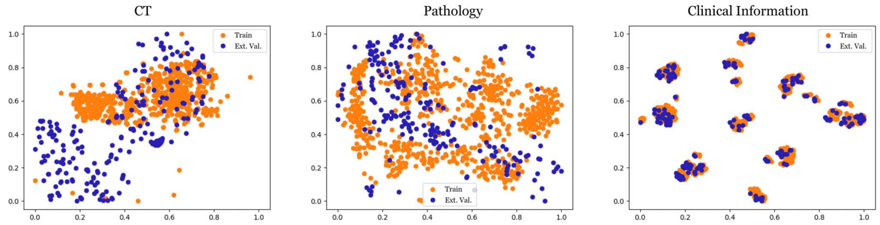

Fig. 2: t-SNE of data distribution for each modality. The distribution variance between the train set and external validation set is minimal for clinical information, yet it is relatively large for image domain modalities, CT and pathology images.

3.2 Preprocessing

CT Pixel values were clipped to remain within the Hounsfield unit range of [-1000, 1000] and then rescaled to [0,1]. Voxel size was normalized to [3.0, 0.69, 0.69] mm^3, and input volumes adjusted to 160 × 512 × 512. For images with spatial dimensions larger than 512 × 512, center cropping was performed to maintain the focus on relevant regions, while smaller images were padded to meet this size requirement. In instances where the number of slices exceeded 160, typically indicating inclusion of the pelvic area, cropping was focused on the lower sections. Conversely, for volumes with fewer slices, zero-padding was applied to the lower end of the CT volume.

Pathology There are typically large variance in staining due to differences in organs and the objectives of observation during the H&E staining process (Fig. 2). Therefore, stain normalization is an essential preprocessing step, and in this study, we performed it using StainNet [6], which has shown great performance. Also, each pathology image was divided into patches of size $3 \times 224 \times 224$, excluding those patches corresponding to the background. On average, approximately 3,000 patches were generated.

Clinical Information Each item is provided in binary or continuous scalar form. However, considering that the text encoder of CLIP was pretrained in the form of words rather than general numbers, every item except age and stage was converted to clinical words. Below is an example of clinical notes. The bold text corresponds to clinical information.

> *"**37** year old **male nonsmoker** lung cancer patient, stage **3**, **T1N2M0**, location **right superior lobe**, type **adenocarcinoma**."*

Then, it is converted into a 1D vector consisting of integers through CLIP's tokenizer.

3.3 Implementation Details

The framework is implemented on PyTorch using 2 NVIDIA RTX A6000 GPUs. The dimension size of extracted patch-wise pathological feature D_0 and each instance feature D are set to 768 and 512, respectively. We optimized the models using the Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of $1e$-7, and learning rate of $1e$-5 for 300 epochs. The binary cross-entropy loss is adopted to train the network.

3.4 Results

Comparison with Existing Methods Both [20,21] attempted multi-modal integration by extracting features from each modality and simply concatenating these features before aggregating them through a fc layer. The results of their comparative experiments are summarized in Table 1 where the method 'concat.+fc' represents the corresponding methodology. The method of previous studies achieved AUCs of 0.648 and 0.699, without and with clinical information, respectively. For effective feature encoding, an additional experiment was conducted where each encoder was first trained on a single modality before fine-tuning on multi-modality. In this case, AUCs of 0.818 and 0.821 were achieved with and without clinical information, respectively. In contrast, AUCs of our method were 0.824 and 0.877.

Table 1: Comparison results on methods for aggregating the features of multi-modality. The text is converted into sentence form as the input into CLIP's visual encoder. * indicates that the text is converted into a single vector and extracted by a linear fc layer. AUC is for external validation set. The highest results are **bolded**.

CT	Pathology	Text	Method	AUC
✓	✓	-	concat. + fc	0.648
✓	✓	-	**G**	0.824
✓	✓	✓	concat. + fc	0.699
✓	✓	✓	**G**	0.807
✓	✓	✓*	**MAM** + **G**	0.829
✓	✓	✓	**MAM** + **G**	**0.877**

Ablation Studies AUC, accuracy, precision, and recall for both internal and external validations are presented in Table 2. The internal validation set was randomly split as 25% of the training set, with accuracy, precision, and recall derived by applying the threshold value that maximizes Youden's index in the internal validation.

Without the use of LLM-guidance, the combination of CT and pathology images achieved an AUC of 0.824, accuracy of 0.832, and recall of 0.934, outperforming single modality approaches. Incorporating the guidance improved the AUC for the external validation set improved by 0.079 for CT images, 0.085 for pathology images, and 0.053 for the combined use of CT and pathology.

The model utilizing all three modalities with this guidance recorded the highest values, with an AUC of 0.877 ($p<0.05$), accuracy of 0.836, and recall of 0.964.

Additionally, without LLM-guidance, the AUC for pathology images in external validation decreased by 0.091 compared to internal validation, whereas with it, the decrease was only 0.008. Other modality combinations also showed smaller performance decreases when the guidance was applied compared to when it was absent.

To further demonstrate its effectiveness, we aggregated clinical information into a single vector and trained a linear layer without employing the CLIP text encoder. Despite using the same feature alignment and aggregation as our model, this method achieved an AUC of 0.829 (Table 1).

Table 2: Quantitative results of AUC, accuracy, precision, and recall for internal and external validation of 5yOS prediction. The increase in the evaluation metrics of the results with text guidance is written below each metric value (▲ indicates an increase, and ▼ indicates a decrease). In the external validation, p-values were calculated by the DeLong's test [2] between with and without text guidance, respectively. * indicates *p*-value is less than 0.05. The highest results are **bolded**.

Modality			**Internal Validation** (N=178)				**External Validation** (N=214)			
CT	**Pathology**	**Text**	AUC	Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall
		✓	0.836	0.775	0.869	0.683	0.823	0.757	0.919	0.753
✓			0.814	0.730	0.819	0.673	0.796	0.678	**0.945**	0.621
✓		✓	0.869	0.792	0.802	0.842	0.875*	0.818	0.926	0.831
			▲0.055	▲0.062	▼0.017	▲0.168	▲0.079	▲0.140	▼0.019	▲0.211
	✓		0.867	0.758	**0.876**	0.772	0.776	0.757	0.885	0.789
	✓	✓	0.868	0.787	0.825	0.792	0.860*	0.822	0.868	0.910
			▲0.001	▲0.028	▼0.052	▲0.020	▲0.085	▲0.065	▼0.017	▲0.120
✓	✓		0.843	0.775	0.770	**0.861**	0.824	0.832	0.861	0.934
✓	✓	✓	**0.898**	**0.798**	0.849	0.832	**0.877***	**0.836**	0.898	**0.964**
			▲0.054	▲0.023	▲0.079	▼0.030	▲0.053	▲0.005	▲0.037	▲0.030

4 Discussion

In this study, we significantly advanced lung cancer prognosis predictions by integrating CT image, pathology slide, and clinical note with LLM via MIL. Our findings, validated internally and externally, showcase the model's enhanced accuracy, precision, recall, and AUC metrics, reflecting a substantial improvement over traditional single-modality diagnostic methods. This multi-modal approach, leveraging advanced artificial intelligence technologies, aligns closely with the nuanced, holistic nature of clinical decision-making.

LLM-guidance was effectively applied across all image modalities to achieve higher evaluation metrics when successfully aggregating various unique pieces of information through MIL, facilitating accurate bag-level predictions. Moreover, it is evident that LLM-guidance is particularly beneficial in external validation contexts. The performance metrics in external validation did not significantly decline compared to internal validation. Since clinical information involves standardized labeling of structured items, there is minimal distribution variance across cohorts and patients (Fig. 2). Consequently, leveraging clinical information has enhanced the model's generalization capacity in external validation scenarios. In the medical field, since prediction of risk group is considered more important, evaluation of recall is inevitable, and our model achieved the highest recall value.

Nevertheless, this study has limitations. First, the study's reliance on manually generated prompts for LLM-guidance introduces performance variability. Optimal prompt making is crucial, as the quality and structure of the prompts significantly impact the model's performance. Combining all clinical information into a single sentence prompt has proven more effective than applying each piece individually (e.g., "a photo of {information}"). Additionally, there is potential to further enhance performance through learnable prompt techniques. Second, it highlights a limitation in detailing specific feature correlations within images and beyond clinical notes. These underscore the need for further research into automated prompt generation and a deeper exploration of the independent contributions of each imaging modality.

In conclusion, the use of LLMs for feature alignment and MIL for data aggregation marks a notable innovation in processing complex medical datasets. Our approach efficiently captures the intricate interplay between different medical data types, significantly enhancing the predictive model's performance. Furthermore, the model demonstrated robust generalization across diverse clinical settings, indicating its potential for widespread clinical application despite challenges in variability across patient populations.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2021R1C1C2008773, 2022R1A2C2008983), Artificial Intelligence Graduate School Program Yonsei University (RS-2020-II201361), the KIST Institutional Program (Project No.2E32983), Yonsei Signature Research Cluster Program of 2023 (2023-22-0008), Seoul R&BD Program(CT-230034) through the Seoul Business Agency(SBA) funded by Seoul Metropolitan Government, and the Starting growth Technological R&D Program (RS-2023-00257885) funded by the Ministry of SMEs and Startups(MSS, Korea).

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

References

1. Buccheri, G., Ferrigno, D.: Prognostic factors in lung cancer: tables and comments. European Respiratory Journal **7**(7), 1350–1364 (1994)

2. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics pp. 837–845 (1988)
3. 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)
4. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
5. Jang, J., Hwang, D.: M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20718–20729 (2022)
6. Kang, H., Luo, D., Feng, W., Zeng, S., Quan, T., Hu, J., Liu, X.: Stainnet: a fast and robust stain normalization network. Frontiers in Medicine **8**, 746307 (2021)
7. Kang, M.J., Won, Y.J., Lee, J.J., Jung, K.W., Kim, H.J., Kong, H.J., Im, J.S., Seo, H.G.: Cancer statistics in korea: incidence, mortality, survival, and prevalence in 2019. Cancer Research and Treatment: Official Journal of Korean Cancer Association **54**(2), 330–344 (2022)
8. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
9. Lian, J., Deng, J., Hui, E.S., Koohi-Moghadam, M., She, Y., Chen, C., Vardhanabhuti, V.: Early stage nscls patients' prognostic prediction with multi-information using transformer and graph neural network model. Elife **11**, e80547 (2022)
10. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
11. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. Advances in neural information processing systems **10** (1997)
12. Miller, K.D., Nogueira, L., Devasia, T., Mariotto, A.B., Yabroff, K.R., Jemal, A., Kramer, J., Siegel, R.L.: Cancer treatment and survivorship statistics, 2022. CA: a cancer journal for clinicians **72**(5), 409–436 (2022)
13. Park, D., Oh, D., Lee, M., Lee, S.Y., Shin, K.M., Jun, J.S., Hwang, D.: Importance of ct image normalization in radiomics analysis: prediction of 3-year recurrence-free survival in non-small cell lung cancer. European Radiology **32**(12), 8716–8725 (2022)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
15. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
16. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
17. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)

18. 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)
19. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)
20. Yao, J., Zhu, X., Zhu, F., Huang, J.: Deep correlational learning for survival prediction from multi-modality data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 406–414. Springer (2017)
21. Zheng, S., Guo, J., Langendijk, J.A., Both, S., Veldhuis, R.N., Oudkerk, M., van Ooijen, P.M., Wijsman, R., Sijtsema, N.M.: Survival prediction for stage i-iiia non-small cell lung cancer using deep learning. Radiotherapy and oncology **180**, 109483 (2023)
22. Zhu, X., Yao, J., Huang, J.: Deep convolutional neural network for survival analysis with pathological images. In: 2016 IEEE international conference on bioinformatics and biomedicine (BIBM). pp. 544–547. IEEE (2016)