Cross Attention-based Fusion of Deep and Radiomics Features for Lung Nodule Invasiveness Prediction

Sadaf Khademi
Concordia Inst. for Info. Syst. Eng.
Concordia University
Montreal, Canada
sa_khad@encs.concordia.ca

Anastasia Oikonomou
Dep. of Medical Imaging
Sunnybrook Health Sciences Centre
Toronto, Canada
anastasia.oikonomou@sunnybrook.ca

Konstantinos N. Plataniotis
Dep. of Elec. and Comp. Eng.
University of Toronto
Toronto, Canada
kostas@comm.utoronto.ca

Arash Mohammadi
Concordia Inst. for Info. Syst. Eng.
Concordia University
Montreal, Canada
arash.mohammadi@concordia.ca

Abstract*—This paper presents an attention-based fusion algorithm that effectively combines two distinct sets of features obtained from chest Computed Tomography (CT) images to predict lung nodule invasiveness. Lung cancer has the highest number of cancer-related deaths worldwide and Lung Adenocarcinoma (LUAC) including three possible invasiveness stages for nodules is discovered as the most prevalent histologic subtype. In this case, developing a proper treatment plan requires accurate information about the invasiveness level of nodules to prevent unwarranted surgeries as pre-invasive/non-invasive nodules can be monitored over time by periodic imaging tests while invasive nodules demand surgical resections. The most reliable method of evaluating nodule invasiveness stage is pathological sampling which is not preferred as the primary diagnostic method due to its invasive nature and potential complications. Currently, chest CT scan is used as an early diagnostic tool by radiologists. However, radiologists' analysis based on CT images is subjective and highly error-prone. The proposed fusion framework has two parallel feature extraction paths: (i) Deep Learning-based path that extracts deep features via a hierarchical vision transformer using shifted windows known as SWin transformer, and; (ii) Radiomics-based path that consists of extracting quantitative features related to nodule specifications appeared in CT slices. Extracted deep and radiomics features are then fused through a Cross-Attention mechanism acts as the fusion center of our framework to predict nodules invasiveness. Experimental results on our in-house dataset of 114 pathologically proven subsolid nodules present the superiority of the proposed fusion framework among its standalone models of parallel paths and achieved an accuracy of 89.46%, sensitivity of 83.99%, and specificity of 94.99% using 10-fold cross-validation.

***Index Terms*—Lung Adenocarcinoma, Transformer, Subsolid Nodule, Self-Attention, Radiomics**

I. Introduction

Lung cancer is a major concern for public health globally since it remains among the top reasons for cancer-related deaths across the world. According to the latest report published by World Health Organization (WHO), lung cancer accounts for 1.8 million deaths annually [1]. Detecting lung cancer at an early stage is crucial in improving the probability of successful treatment and favorable health outcomes. Early diagnosis allows for prompt medical intervention and implementation of effective treatment plans, which can significantly enhance the chances of recovery and increase the patient's overall survival rate. Non-Small-Cell Lung Cancer (NSCLC) is the most common type of lung cancer and among its subtypes, Lung Adenocarcinoma (LUAC) is the most prevalent [2]. SubSolid Nodules (SSNs) are a specific type of pulmonary nodule that can be typically characterized by a mixture of Ground-Glass Opacity (GGO) and solid components, making them distinct from purely solid nodules. Research study performed in [3] demonstrates that the malignancy rate of SSNs is significantly higher than the rate observed in solid nodules, which leads SSNs to be identified as a significant clinical concern due to their association with early-stage LUAC. The importance of SSNs diagnosis lies in the fact that its degree of invasiveness can greatly impact the likelihood of malignancy and the appropriate course of treatment. The diagnosis of SSNs can be challenging due to their complex and heterogeneous nature, however, there are several diagnostic approaches to identify and characterize SSNs.

Computed tomography (CT) imaging is the most commonly used method for detecting and evaluating SSNs, with high spatial resolution and the ability to capture detailed images of lung tissues. However, accurate diagnosis between benign and malignant nodules requires a careful analysis of imaging features such as size, shape, density, and the presence or absence of solid components that cause difficulty in achieving common agreement between radiologists about findings. In some cases, further diagnostic tests such as Positron Emission Tomography

979-8-3503-3959-8/23/$31.00 ©2023 IEEE

(PET) scanning, which measures metabolic activity in the lung tissue, may be needed to confirm the diagnosis of SSNs and assess their invasiveness. It is worth noting that the most accurate diagnostic technique to stage SSNs relates to the biopsy, in which a tissue sample is collected for microscopic examination but it is not preferred due to its invasive nature and potential complications [4] which calls for further research works to develop noninvasive diagnostic methods for accurate differentiation of benign and malignant SSNs.

Recently, Deep Learning (DL) based methodologies have shown promising results in improving the accuracy and efficiency of SSN diagnosis [5], [6]. DL models can autonomously identify subtle imaging features and patterns that may be difficult for human experts to detect.

Literature Review: The current literature on assessing the invasiveness of SSNs can be broadly divided into three categories: radiomics-based [7], [8], DL-based [9], and hybrid solutions [10] that integrate both approaches. Radiomics-based frameworks involve the extraction of quantitative features about the nodule's shape, texture, and heterogeneity, while DL-based frameworks automatically learn patterns and extract features from the images. Alternatively, hybrid frameworks overcome the limitations related to stand-alone techniques such as the need for large uniform pathology-confirmed annotated datasets. Moreover, DL algorithms can learn to identify relevant features even in low-quality images, thus reducing the dependence of radiomics features on the quality and resolution of imaging equipment while leading to a more accurate and reliable evaluation of SSN malignancy. Volumetric CT scans are made up of a sequence of 2D images, which together create a 3D vision of the body tissue. In other words, each slice can present a different view of the nodule being imaged which helps the radiologist to get a sense of nodule variations in size or shape by looking at the sequence of slices. DL models are capable of analyzing both 2D and 3D visions of CT scans. As an example of hybrid frameworks built on a 3D processing scheme, a fusion algorithm that combines handcrafted features into the features learned from a 3D Convolutional Neural Network (CNN) was developed to predict lung nodule malignancy [11]. In the aforementioned study, the Support Vector Machine (SVM) was coupled with the Sequential Forward feature Selection (SFS) method to select the optimal feature subset among features combination and construct the final classifier. The important point about 3D-based DL models is that although they are more accurate in extracting image details but are computationally demanding and require larger training datasets. On the other hand, 2D-based DL models are more efficient in terms of computational complexity and recently a surge of interest [12] appeared in using sequential deep models for analyzing 2D CT scans.

When DL models are applied to analyze CT scans, it is imperative to focus on key important components/regions that are essential for accurate diagnosis. This is where attention-based mechanisms like Transformer architectures can be useful, as they enable the model to selectively focus on specific regions of the image. Transformer concept was initially introduced in 2017 and has since been applied to a wide range of tasks such as natural language processing [13]–[17]. Transformer models are capable of capturing long-range dependencies in the input sequence more effectively than traditional Recurrent Neural Networks (RNNs) and CNNs [18]. In 2020, a new variant of the transformer model called Vision Transformer (ViT) [19] was presented for image processing applications. ViT splits the input image into a sequence of fixed-size patches and then applies the same attention mechanism used in the Transformer model. Even though it has not been around for very long, ViT has shown great promising results in the field of computer vision and is considered the fundamental component of our DL model.

Contributions: The main contribution of this article is to demonstrate effectiveness of combining radiomics and deep features in improving the performance and robustness of Lung Nodule Invasiveness (LNI) prediction. This goal has been achieved by fusing hierarchical features generated by Shifting Window (SWin) Tansformer [20] and various quantitative measurements of nodule characteristics utilizing a cross-attention module to learn relationships between these two feature sets. In particular, a fusion model based on cross-attention can effectively provide complementary information by allowing each feature set to attend to the other set, resulting in a more comprehensive understanding of the input data. This integration allows the model to capture the interactions and dependencies between two feature sets, resulting in a higher level of performance than either set can achieve alone. Experimental results on our in-house dataset including 114 pathologically proven SSNs achieved an accuracy of 89.46% in classifying invasive/non-invasive nodules.

II. Fusion of Deep & Radiomics Features for LNI Prediction

The proposed framework, which fuses deep and Radiomics features for the task of LNI prediction, utilizes two parallel processing paths to extract deep and radiomics features from CT slices. Extracted features are then combined using a fusion strategy known as cross-attention to perform classification. In what follows, first the dual path feature extraction procedures, which create desired inputs for the fusion module are described in detail. Afterwards, the fusion technique and classification mechanism will be presented.

A. Deep Learning-based Features

The proposed DL model is established on the transformer architecture. The main constructive component of the transformer encoder that computes relationships between different instances of input in a sequence is called Self-Attention. The self-attention mechanism linearly transforms the input sequence into three representative vectors known as query (Q), key (K), and value (V) with dimensions d_k, d_k, and d_v. Then, the attention weights are computed by taking the dot product between the query vector and the key vector for each element in the sequence. These dot products are then scaled by a factor that depends on the dimensionality of the query and key vectors and passed through a softmax activation function.

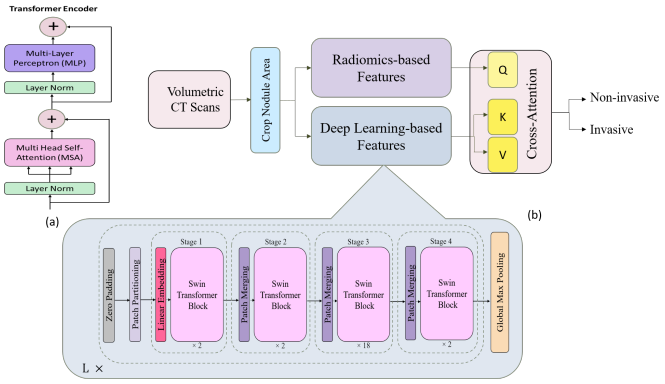

Fig. 1. (a) Transformer encoder architecture, (b) Pipeline of the fusion framework. L represents the number of parallel SWin-Transformer paths per the number of slices available for each patient.

Finally, the weighted sum of the value vectors is computed using the attention weights, resulting in a new representation for each element in the sequence, i.e.,

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

where superscript ^{T} indicates the transpose of the given matrix. In order to increase the model's ability to capture more complex dependencies, the input sequence can be split into multiple representations, or "heads" by applying different linear transformations. Afterward, the self-attention mechanism is applied to each head independently in parallel, with different sets of weights learned for each attention head. The output of each head is then concatenated and passed through a linear transformation to produce the final output. The described process which is known as Multi-head Self Attention (MSA) allows the model to attend to multiple aspects of the input sequence simultaneously and can be expressed as follows

$$MSA(Q, K, V) = Concat\big(head_1, \cdots, head_h\big)W^O,$$
$$\text{where } head_i = Attention(QW_i^Q, KW_i^K, VW_i^V), \quad (2)$$

where the projections are calculated based on the parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$. In addition to the MSA layer, the transformer encoder also includes Layer Normalization, a Multi-Layer Perceptron (MLP), and residual connections as shown in Fig. 1. The combination of these components in the transformer encoder enables the model to learn more meaningful representations of the sequence data.

The DL feature extraction path of our framework is designed according to a SWin transformer that has a hierarchical attention-based structure to discover informative spatial features from nodule patches. The superiority of this model over the basic ViT architecture is addressing computational complexity and fixed-scale patches as two major limitations faced in previous designs by utilizing hierarchical feature maps and the shifted window attention mechanism. The overall architecture of the SWin transformer including four processing stages is shown in Fig. 1.

At first, the segmented area surrounding each nodule resulting from the annotation coordinates in pre-processing the dataset is zero-padded to $(224, 224)$ pixels and divided into 4×4 patches. In Stage 1, the linear embedding layer projects each patch's feature dimension into 128 and then SWin Transformer blocks are applied to the features of each patch. SWin transformer block layers are identical to the original transformer encoder architecture except for the MSA module, which is modified by a shifting window mechanism to compute self-attention within local windows. This modified design has two consecutive transformer encoder units by replacing the MSA module with Window MSA (W-MSA) and Shifted Window MSA (SW-MSA) modules, respectively, to consider cross-window connections and decrease computational complexity. W-MSA module computes self-attention in non-overlapping windows of size 7×7 patches. Since each window contains a fixed number of patches throughout the network, window-based MSA provides a linear computational complexity with respect to input image size. In the following unit, the SW-MSA module calculates self-attention in shifted W-MSA windows generated by a shift factor of $\frac{M}{2}$. The hierarchical feature maps are achieved by the patch merging technique, which is the process of combining the representations generated for multiple patches of the input signal in the SWin transformer to produce a single representation for the entire input signal. This technique concatenates extracted features of neighboring patches. We implemented the SWin-B architecture of SWin transformer that has 2, 2, 18, and 2 layers for Stages 1 to 4, respectively. The output feature vectors generated for the slices of each subject from Stage 4 were fed to a Global Max Pooling (GMP) layer to aggregate the spatial variations of nodules per participant [21]. Due to the small size of our dataset, we fine-tuned a pre-trained SWin-B transformer trained on ImageNet-21k dataset [22], which provides a feature map consisting of $1,024$ features for each patient.

B. Radiomic-based Features

The second path of our framework depends on a combination of quantitative features including geometric and CT attenuation parameters extracted from CT scans that generally result in 48 features. The geometric parameters were automatically generated after nodule segmentation including specifications based on nodule volume, mean, minimum and maximum diameter of the nodule, and lesion irregularity index defined as the maximum to minimum diameter ratio. Finally, analyzing the CT attenuation histogram provided parametric (mean, standard deviation, skewness, and kurtosis) and non-parametric features (quantile-based values) to complete our feature set.

C. Attention-based Fusion Module

In this step, the output feature maps generated by each of our processing paths are passed through a cross-attention module. In contrast to the self-attention mechanism that attends to different parts of a single set of features, the cross-attention mechanism attends to two different feature sets and computes

the attention weights similarly to self-attention except in comparing the query vectors from one set to the key vectors from the other set which improves model performance by selectively attend to relevant parts of both sets of features. To accomplish this, extracted radiomics and DL features are projected to a common dimension of 100 and passed to the query and key-value vectors, respectively. The classification task was done on the output of the MLP head of the cross-attention module using a softmax activation function to produce probability scores related to binary classes.

III. Experimental Results

In this section, we evaluate performance of the proposed fusion framework. In what follows, first we introduce the dataset used for model assessment and then present the evaluation results.

A. Dataset

CT scans used in this study were acquired from an in-house dataset initially presented in [23] collected from the latest scan before the surgery session of patients with technical parameters of 100 - 135 (kVp) and 80 - 120 (mAs). The original data includes 109 SSNs confirmed and labeled by a subspecialty pulmonary pathologist after surgical resections into three groups of pre-invasive, minimally invasive, and invasive. Following the same approach utilized in [23], we combined pre-invasive and minimally invasive nodules as a single class against the invasive nodules and added five additional cases from the same institution to enhance the data balance. As a result, the final dataset consisted of 58 non-invasive and 56 invasive cases.

For each patient, two chest radiologists who were not informed of the pathology results independently evaluated the CT slices and selected the slices where the nodule was visible. Afterward, the commercial software Vitrea v7.3 was used to automatically create a preliminary segmented region around each SSN. Following this, the segmented regions were manually reviewed and corrected by chest radiologists to generate nodule annotations.

B. Results

In this regard, the 10-fold cross-validation technique is used to assess the impact of cross-attention fusing technique in improving SSNs malignancy prediction considering both DL and radiomics features. The pre-trained SWin-B transformer was re-trained using the AdamW optimizer with a learning rate of $1e-5$, a weight-decay of 0.05, and 50 epochs. Then, the feature maps obtained by the DL-path and radiomics-path were fed to the fusion center. The fusion center was trained by the Adam optimizer with a learning rate of $1e-3$, and 50 epochs. It should be noted that dropout layers and early stopping training strategy were incorporated in our model to prevent potential over-fitting issues. Furthermore, the loss function utilized for the entire model was cross-entropy. Table I presents the classification performance of the proposed hybrid framework with its individual standalone models of parallel paths and other algorithms. The classification results show that our hybrid framework outperformed both standalone DL-based and radiomics-based models. Moreover, the proposed hybrid framework achieved the best-performing metrics compared to other DL models such as CNN, CNN-LSTM, and CAET-SWin [24]–[26].

TABLE I
Classification results obtained by the proposed fusion framework and its counterparts.

Model	Accuracy	Sensitivity	Specificity
Radiomics	81.00%	80.00%	81.80%
CNN	68.40%	59.33%	77.99%
CNN-LSTM	72.95%	65.66%	79.66%
CAET-SWin	82.65%	83.66%	81.66%
SWin	78.10%	76.66%	79.66%
SWin-Radiomics	**89.46%**	**83.99%**	**94.99%**

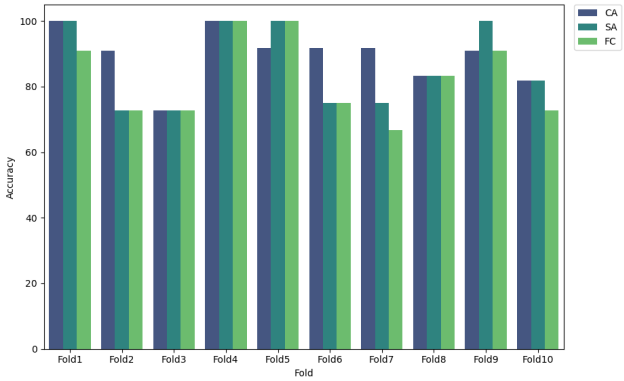

Fig. 2. Barplot of 10-fold cross-validation results for three feature fusion algorithms. CA, SA, and FC represent Cross-Attention, Self-Attention, and Fully-Connected, respectively

In order to assess the importance of the fusion method used in our study, we conducted additional experiments where the cross-attention mechanism in the fusion center is replaced with Fully-Connected (FC) layers and self-attention, respectively. This was conducted by concatenating two feature sets and feeding them as input to the fusion center. We then analyzed the classification accuracy of these three models over 10 folds and presented the results in Fig. 2.

The average classification accuracy for the models using cross-attention, self-attention, and FC layers was 89.46%, 86.06%, and 82.50%, respectively. These results indicate that simultaneous attention to both categories of features is essential for the model to classify malignant/benign SSNs.

IV. Conclusion

This paper highlighted the significance of integrating DL and radiomics feature extracting models using an effective fusion module for LNI prediction. Our findings suggest that cross-attention mechanism used in the fusion center can discover the full relation of multi-modal data by capturing and combining relevant features from both DL and radiomics feature sets, leading to superior classification performance compared to alternative fusion methods.

REFERENCES

[1] KC. Thandra, A. Barsouk, K. Saginala, JS. Aluru, A. Barsouk, "Epidemiology of lung cancer," *Contemp Oncol (Pozn)*, vol. 25, no. 1, pp. 45-52, Feb 2021.

[2] Z. Chen, CM. Fillmore, PS. Hammerman, CF. Kim, KK. Wong, "Non-small-cell lung cancers: a heterogeneous set of diseases," *Nat Rev Cancer*, vol. 14, no. 8, pp. 535-546, Aug 2014.

[3] CI. Henschke, DF. Yankelevitz, R. Mirtcheva, G. McGuinness, D. McCauley, OS. Miettinen, "CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules," *AJR Am J Roentgenol*, pp. 1053-7, May 2002.

[4] H. Zhang, S. Wang, Z. Deng, Y. Li, Y. Yang, H. Huang, "Computed tomography-based radiomics machine learning models for prediction of histological invasiveness with sub-centimeter subsolid pulmonary nodules: a retrospective study," *PeerJ*, Jan 2023.

[5] R. Li, C. Xiao, Y. Huang, H. Hassan, B. Huang, "Deep learning applications in computed tomography images for pulmonary nodule detection and diagnosis: a review," *Diagnostics*, Jan 2022.

[6] L. Wang, "Deep learning techniques to diagnose lung cancer," *Diagnostics*, Nov 2022.

[7] CH. Chen et al., "Radiomic features analysis in computed tomography images of lung nodule classification," *PLoS One*, Feb 2018.

[8] Y. Xu et al., "Application of radiomics in predicting the malignancy of pulmonary nodules in different sizes," *Am J Roentgenol*, Dec 2019.

[9] T. Shen et al., "Predicting Malignancy and Invasiveness of Pulmonary Subsolid Nodules on CT Images Using Deep Learning," *Front. Oncol*, 2021.

[10] K. Mehta et al., "Lung module classification using biomarkers, volumetric radiomics, and 3D CNNs," *J Digit Imaging*, pp. 647-666, 2021.

[11] S. Li et al., "Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features," *Phys Med Biol*, vol. 64, no. 17, Sep 2019.

[12] MM. Farhangi, N. Petrick, B. Sahiner, H. Frigui, AA. Amini, A. Pezeshk, "Recurrent attention network for false positive reduction in the detection of pulmonary nodules in thoracic CT scans," *Med Phys*, June 2020.

[13] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, pp. 5998-6008, 2017.

[14] Z. Hajiakhondi-Meybodi, A. Mohammadi, E. Rahimian, S. Heidarian, J. Abouei, and K. N. Plataniotis, "TEDGE-Caching: Transformer-based Edge Caching Towards 6G Networks," *IEEE International Conference on Communications (ICC)*, Apr 2022.

[15] Z. HajiAkhondi-Meybodi et al., "Multi-Content Time-Series Popularity Prediction with Multiple-Model Transformers in MEC Networks," *arXiv preprint arXiv:2210.05874*, Oct. 2022.

[16] Z. HajiAkhondi-Meybodi, A. Mohammadi, M. Hou, J. Abouei, and K. N. Plataniotis, "ViT-CAT: Parallel Vision Transformers with Cross Attention Fusion for Popularity Prediction in MEC Networks," *arXiv preprint arXiv:2210.15125*, 2022.

[17] M. Montazerin, E. Rahimian, F. Naderkhani, S. F. Atashzar, S. Yanushkevich, A. Mohammadi "Transformer-based Hand Gesture Recognition via High-Density EMG Signals: From Instantaneous Recognition to Fusion of Motor Unit Spike Trains," *arXiv preprint arXiv:2212.00743*, 2022.

[18] S. Paul, PY. Chen, "Vision transformers are robust learners," *In Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, No. 2, pp. 2071-2081, Jun 2022.

[19] A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," *International Conference on Learning Representations*, Oct 2020.

[20] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012-10022, oct 2021.

[21] L. Zhang and Y. Wen, "A transformer-based framework for automatic covid19 diagnosis in chest cts," *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 513-518, 2021.

[22] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, 2009.

[23] A. Oikonomou et al., "Histogram-based models on non-thin section chest CT predict invasiveness of primary lung adenocarcinoma subsolid nodules," *Scientific Reports*, vol. 9, no. 1, Apr 2019.

[24] W. Li, P. Cao, D. Zhao, J. Wang, "Pulmonary Nodule Classification with Deep Convolutional Neural Networks on Computed Tomography Images," *Comput Math Methods Med*, 2016.

[25] Z. Ni, Y. Peng, "A serialized classification method for pulmonary nodules based on lightweight cascaded convolutional neural network-long short-term memory," *Int J Imaging Syst Technol*, vol. 30, pp. 950–962, 2020.

[26] S. Khademi, S. Heidarian, P. Afshar, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "Spatio-Temporal Hybrid Fusion of CAE and SWIn Transformers for Lung Cancer Malignancy Prediction," *arXiv preprint arXiv:2210.15297*, 2022.