Proceedings of CONF-SEML 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA33119

Multimodal-based Thyroid Nodule Classification Prediction

Yi Zhao[1], Yufan Liu[1], Yinyue Fang[1], Shukuan Sun[1], Yuting Liu[1*]

[1]School of Medical Information, Wannan Medical University School, Wuhu, China
**Corresponding Author. Email: liuyuting@wnmc.edu.cn*

Abstract. Early diagnosis of thyroid diseases has challenges due to limitations of single-modal data. Benign and malignant thyroid nodules overlap in ultrasonographic features, and clinical data is crucial for imaging interpretation and practice. This study proposes a multimodal fusion strategy that integrates thyroid ultrasound images with clinical data features to improve nodule prediction accuracy. An EfficientNet - b4+BERT fusion model with multi - head self - attention mechanisms is developed for dynamic weighted feature vector fusion, which enhances image - text feature alignment efficiency by leveraging inter - modal complementarity. Comparative experiments show the model outperforms single models in accuracy and stability. The proposed technology performs well in high - risk thyroid tumor prediction and has significant clinical application value.

Keywords: EfficientNet-b4 + BERT Fusion Model, Multi-Head Self-Attention.,Thyroid nodule prediction

1. Introduction

Thyroid nodules (TNs) are common thyroid disorders.with approximately 95% being benign and not requiring immediate intervention, but about 5% are malignant [1]. Early diagnosis of malignant nodules is crucial for preventing cancer progression and improving survival rates. Fine-needle aspiration biopsy (FNAB) is the gold standard for diagnosis but is invasive and requires skilled personnel [2]. High-resolution ultrasound, preferred for its high sensitivity and non-invasiveness, evaluates parameters like nodule size, margins, and calcification [3]. The Thyroid Imaging Report and Data System (TI-RADS) aids in treatment planning and prognosis assessment [4]. However, imaging features of grade 3-5 nodules often overlap, complicating visual differentiation [5]. Ultrasound findings are operator-dependent, leading to subjective variability and lower accuracy for small lesions, impacting treatment decisions [6]. Thus, developing accurate diagnostic techniques is clinically significant.

2. Related work

Deep learning models using multimodal fusion techniques can make nodule segmentation more accurate. Also, doctors can make better diagnosis and treatment plans [7].This is supported by some studies: for example, Zhou et al. [8] developed a deep learning radiomics model.has better performance than basic CNN and TL models for differentiating benign and malignant nodules. Also, Wu et al. [9] constructed a deep multimodal learning model. obtained 0.973 AUC for predicting

DOI: 10.54254/2755-2721/2026.BA33119

cervical lymph node metastasis. Also, Chen et al. [10] used ultrasound radiomics-based model to predict papillary thyroid microcarcinoma in TI-RADS 3 nodules. Accuracy can be improved when combining radiomics features with clinical information. The reason is that radiomics has become a focus in clinical research recently. By extracting quantitative imaging features, it can reveal disease characteristics and provide auxiliary evidence for diagnosis and prognosis prediction.

3. Method

Our model is based on EfficientNet and BERT. It is worth noting that we incorporate MSA to fuse the combined feature vectors, because in thyroid nodule grading prediction the main goal is to identify progression levels of nodules and integrate ultrasound image information with clinical data at the same time, which can help improve prediction performance.

3.1. Feature extraction of ultrasound images

Figure 1 shows the specific architecture. When extracting features from input image, we first apply 3×3 convolution. Then the features are processed through multiple MBConv modules. At last, the output should pass through 1×1 convolutional layers and global average pooling layers, and then it goes to fully connected layers. In fact, this design uses EfficientNet-B4 as a variant of EfficientNet-B0. It is worth noting that the network uses a unified compound scaling method. This method can expand the width and depth, and also resolution parameters at the same time using compound coefficients. So the performance can be improved.

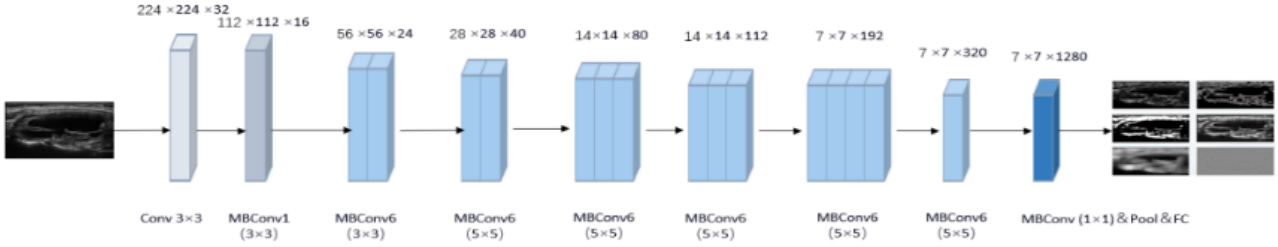

Figure 1. Structure diagram of the EfficientNet model network

3.2. Clinical data feature extraction

BERT takes word, position, and segment embeddings as input. The input passes through stacked Transformer encoder modules with multi - head self - attention mechanisms, which capture context information from different positions and build word representations. These modules assist the model in learning from data during pre - training. BERT, a pre - trained language model using self - supervised learning on large - scale corpora, can generate high - quality word feature representations. The specific structure is shown in Figure 2.

DOI: 10.54254/2755-2721/2026.BA33119

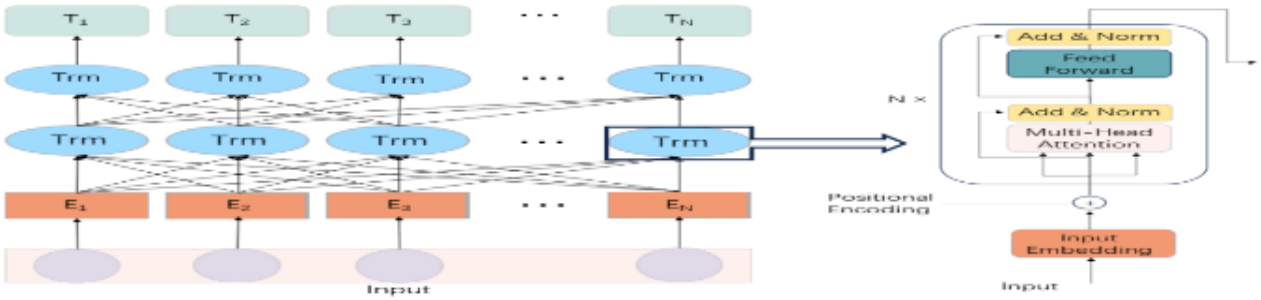

Figure 2. Schematic diagram of BERT structure

3.3. Feature fusion

This study uses a dynamic weighted fusion strategy, combining hierarchical feature fusion and multi - head self - attention mechanisms, to deeply integrate image and text features. After constructing joint feature vectors, the multi - head self - attention mechanism conducts dynamic weighted fusion to capture diverse features and improve model representation. The computational process of this mechanism is shown in Equations (1) and (2).

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \cdots, head_h)W^0 \tag{1}$$

Where: Q, K, and V are query, key, and value vectors respectively; h is the number of heads; head_i represents the output of the i - th head; W^0 is the output transformation matrix. Each head's output head_i can be expressed as:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

Where: W_i^Q, W_i^K, W_i^V

represent the query, key, and value transformation matrices for the i-th head, respectively.

4. Experiment

To comprehensively evaluate the proposed algorithm's performance in thyroid nodule grading prediction, this study designed comparative experiments. First, under the same conditions, three neural networks (ResNet50, DenseNet121, and EfficientNet - b4) were compared to find the optimal model architecture for image feature extraction. Then, the most effective feature extraction model was integrated with BERT. Next, all models were retrained on our shared dataset with default parameters. Finally, the experimental results were thoroughly analyzed.

4.1. Datasets

This retrospective study included patients with complete imaging data who had thyroid contrast - enhanced ultrasound examinations in a tertiary hospital's ultrasound department from January 2024 to April 2025. Clinical data like patient age, gender,margins, morphology, and TI - RADS classification of thyroid nodules were extracted from ultrasound reports. 776 samples were obtained and divided into a training set (620 cases) and a test set (156 cases).

DOI: 10.54254/2755-2721/2026.BA33119

4.2. Results

Under specific parameter settings, we imported training datasets for comparative analysis to find the optimal image feature extraction model. We calculated performance metrics on the training set, compared key indicators across experimental groups, and summarized the results in Table 1.

Table 1. Comparison table of single visual model performance

Model	Accuracy	Precision	Recall	Loss
ResNet50	0.9411	0.9120	0.9322	0.3017
DenseNet121	0.9242	0.9250	0.9242	0.0330
EfficientNet-b4	0.9577	0.9579	0.9577	0.0090

After the comparative experiments, EfficientNet - b4 was chosen as the best model for thyroid image feature extraction. Also, this study assessed the combined performance of EfficientNet - b4 and BERT. Table 2 shows the single - modal vs multi - modal performance comparison.

Table 2. Comparison table of unimodal and multimodal performances

Model	Accuracy	Precision	Recall	Loss
BERT	0.9500	0.9423	0.9400	0.1165
EfficientNet-b4	0.9577	0.9579	0.9577	0.0090
BERT+EfficientNet-b4	0.9673	0.9714	0.9721	0.0070

The experimental results demonstrate that the BERT+EfficientNet-b4 model outperforms single-modal . The specific performance metrics of the model are presented in Table 3 and Figure 3:

Table 3. Performance of the EfficientNet-b4+BERT model

label	Precision	Recall	F1-score
TI-RADS:Category 3	0.971	0.985	0.943
TI-RADS:Category 4a	1.000	0.956	0.951
TI-RADS:Category 4b	0.986	1.000	0.993
TI-RADS:Category 4c	0.960	0.945	0.953
TI-RADS:Category 5	0.940	0.965	0.977

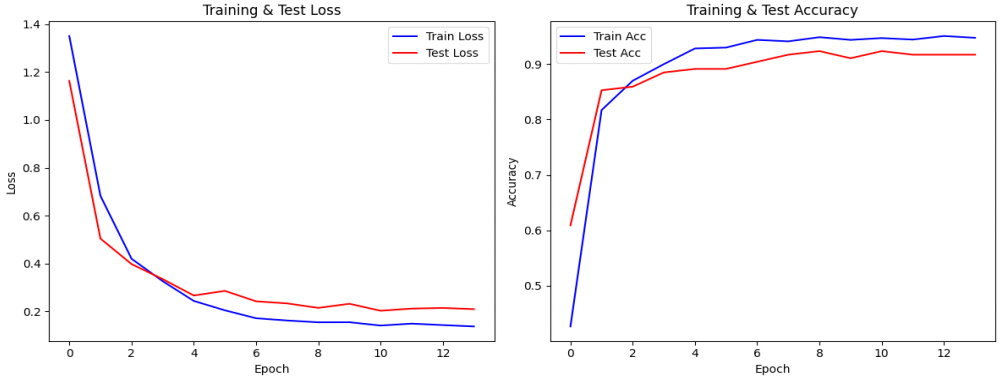

Figure 3. Precision and loss variation diagram

DOI: 10.54254/2755-2721/2026.BA33119

5. Conclusion

This study integrates thyroid ultrasound image features extracted by the EfficientNet - b4 network model with clinical information features from the Bert model. Using feature fusion, we constructed a multimodal thyroid grading prediction model to improve the accuracy of benign/malignant nodule classification. This approach prevents overdiagnosis of low - risk tumors and enables precise prediction of high - risk tumors, showing significant clinical application value.

Acknowledgments

We sincerely thank Anhui University Student Innovation Training Program (Nos. 202510368002 & 202510368064) and Wannan Medical University's Horizontal Research Project (No. H202530) for funding. Also, gratitude to our supervisors, peers, and families for support.

References

[1] Chen, L. J., & Wang, L. (2025). Research progress of radiomics in qualitative diagnosis of thyroid nodules. Magnetic Resonance Imaging, 16(2), 165–171.
[2] Haugen, B. R., Alexander, E. K., Bible, K. C., et al. (2016). 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. Thyroid, 26(1), 1-133.
[3] China Integrated Diagnosis and Treatment Guidelines for Cancer (CACA) - Thyroid Cancer. (2025, January 10). https: //cacaguidelines.cacakp.com/pdflist/detail?id=417
[4] Xu, Q. J., Huang, P. F., Sun, H., et al. (2023). Diagnostic efficacy of AI TI-RADS combined with clinical indicators in differentiating benign and malignant thyroid nodules. Chinese Journal of Ultrasound Medicine, 39(11), 20–27.
[5] Zhang, M., Jin, Z., Zhao, H. L., Li, C. C., & Cao, J. Y. (2024). Prediction model for benign and malignant thyroid nodules based on multimodal ultrasound combined with clinical features. Chinese Journal of Medical Imaging, 32(1), 14–20.
[6] Topcuoglu, M. O., Uzunoglu, B., Orhan, T., et al. (2025). A real-world comparison of the diagnostic performances of six different TI-RADS guidelines, including ACR-/Kwak-/K-/EU-/ATA-/C-TIRADS. Clinical Imaging, 117, 1103-1103.
[7] Mayerhoefer, M. E., Materka, A., Langs, G., et al. (2020). Introduction to radiomics. Journal of Nuclear Medicine, 61(4), 488-495.
[8] Zhou, H., Jin, Y., Dai, L., et al. (2020). Differential Diagnosis of Benign and Malignant Thyroid Nodules Using Deep Learning Radiomics of Thyroid Ultrasound Images. European Journal Of Radiology, 127, 108992-108992.
[9] Wu, X., Li, M., Cui, X. W., et al. (2022). Deep multimodal learning for lymph node metastasis prediction of primary thyroid cancer. Physics In Medicine And Biology, 67(3), 035008-035008.
[10] Chen, Z., Zhan, W., Wu, Z., et al. (2024). The ultrasound-based radiomics-clinical machine learning model to predict papillary thyroid microcarcinoma in TI-RADS 3 nodules. Translational Cancer Reserch, 13(1), 278