Available Online: 1 July 2025 DOI: 10.54254/2977-3903/2025.24722

Multi-modal fusion and transferable deep learning for rare disease detection: a CNN-Transformer framework with cross-domain adaptation on limited CT and MRI data

Jingyu Tang

University of Sydney, Sydney, Australia

rara481846778@gmail.com

Abstract. Medical imaging diagnosis of rare diseases faces the dual challenge of scarce labeled data and significant differences in equipment. This study proposes a hybrid framework integrating CT and MRI modalities, combining CNN and Transformer architectures and introducing an adversarial domain adaptation mechanism. The dedicated CNN encoder extracts fine-grained local features, while the Transformer module captures long-range cross-modal dependencies. The gradient inversion domain discriminator aligns the feature distribution of different scanning devices to ensure the device independence of the model. On the two public datasets of neurological diseases, intracranial hemorrhage and demyelinating lesions, the average accuracy rate of this model reached 91.3%, and the F1 score was 0.89, which is 5 to 10 percentage points higher than the single-mode and pure Transformer baseline. Ablation experiments confirmed that the domain-based adversarial transformer component and training contributed to significant performance gains. In the cross-domain (CT to MRI) experiment, the domain adaptation technique increased the F1 score from 0.74 to 0.84. These results highlight the effectiveness of local feature extraction, global context modeling, and collaborative adversarial alignment schemes in multi-agency scenarios with sparse data. Future research will be extended to three-dimensional volume data, integrate semi-supervised learning of unlabeled images, and optimize the reasoning process of real-time clinical decision support.

Keywords: rare disease detection, multi-modal fusion, cnn-transformer, domain adaptation, transfer learning

1. Introduction

Accurate and rapid diagnosis of rare neurological diseases is crucial for clinical intervention and improved prognosis. However, most deep learning methods for medical images rely on large-scale labeled data, which is often difficult to achieve in the field of rare diseases. CT and MRI each have complementary advantages: CT excels at imaging high-density anatomical structures, while MRI offers excellent soft tissue contrast. Theoretically, dual-modal fusion can improve pathological characterization capabilities. However, simple fusion strategies (such as channel splicing) not only lack the benefits of modal specificity but are also susceptible to disruption by the dominant modal, leading to imbalance.

Differences in imaging equipment, acquisition protocols, and patient groups can lead to domain shifts, resulting in a decrease in the generalization ability of the single-mechanism training model in new scenarios. This inconsistency between different devices is particularly pronounced in the field of rare diseases—limited sample size hinders retraining or fine-tuning for new domains. Therefore, a new type of architecture is urgently needed, one that can not only effectively integrate multimodal information but also adaptively shift domains to maintain the robustness of multi-mechanism deployment [1].

This study proposes a hybrid CNN-transformer architecture integrating CT and MRI and combining the adversarial domain adaptation mechanism to achieve rare disease detection. The dedicated dual-mode CNN encoder extracts local detail features, and after tokenization, the long-range cross-modal correlation is captured by the multi-head self-attention transformer. The adversarial domain discriminator connected by the gradient inversion layer aligns and fuses the feature distribution to improve device independence. In the unbalanced, low-sample scenarios of the two major public neurological disease datasets, this method significantly outperforms single-modal baselines and pure transformers. Ablation experiments verified the key contributions of the Transformer module and adversarial training in the domain. Cross-domain experiments show that the domain adaptation strategy increases the cross-modal F1 score by 10 percentage points (0.74 → 0.84). Section 2 reviews related research on multimodal medical imaging, rare disease detection using CNNs, and the application of the transformer; Section 3 details the methodology, including architecture design, adaptation strategies, and integration modules; Section 4 explains the dataset, preprocessing, evaluation parameters, and training process; Section 5 presents the results, ablation analysis, and robustness experiments. Finally, Section 6 summarizes the contributions and considers future research directions.

2. Literature review

2.1. Multi-modal learning in medical imaging

The multimodal learning framework is dedicated to integrating data from different image sources to utilize complementary information. Existing methods cover a variety of strategies from simple concatenation of feature vectors to the use of blocked networks to balance the input. However, most schemes fail to address the modal imbalance problem. When training signals are unevenly distributed, it is easy to make the network over-rely on a single modal. In the field of rare disease detection, research on adaptive fusion strategies capable of dynamically adjusting modal weights is still insufficient [2].

2.2. Cnn-based rare disease detection

Figure 1 shows a typical CNN-based medical image classification process. The original CT or MRI sections first undergo preprocessing and data enhancement operations—including cropping, rotation, intensity normalization, etc.—to expand the effective training set and reduce sensitivity to scanning noise. The processed images are input into pre-trained convolutional networks (such as AlexNet, ResNet-18, DenseNet-201, SqueezeNet). These networks learn the underlying edges, textures, and higher-order anatomical patterns through convolution and stepwise pooling layers, and finally output the classification results (normal or lesion) through the fully connected layer [3]. Although CNNs are good at capturing spatially hierarchical features, they face a dual challenge in typical scenarios of small sample imbalances in rare diseases: first, limited training samples are likely to allow the network to capture spurious artifacts (such as fixed station positions or device-specific noise), resulting in overfitting rather than true lesion recognition; second, although the deep architecture expands the nominal receptive field, its essence still has locality—a single neuron can only "observe" a limited number of pixel neighborhoods simultaneously, making it difficult to model long-range dependencies (such as subtle morphological distortions across the entire organ).

Transfer learning reduces the sparsity of some data by initializing pre-training weights from natural image datasets and integrating universal edge and texture detectors. However, differences in disease manifestations and scanning equipment parameters often prevent direct fine-tuning. In practice, without additional strategies (such as domain adaptation or attention modules), CNN performs poorly in cross-agency data generalization [4]. Figure 1 clearly reveals that even with data augmentation and state-of-the-art pre-trained backbone networks, pure CNN solutions still struggle to robustly address the challenges posed by new devices or rare lesions.

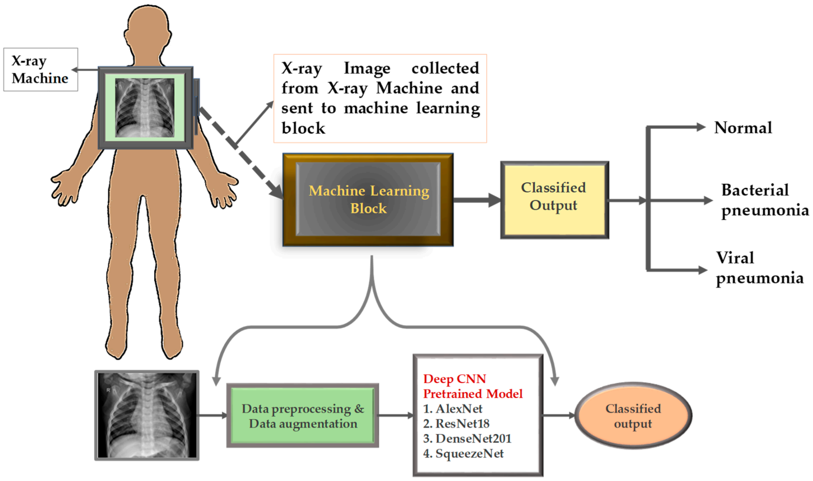

Figure 1. CNN-based medical image classification pipeline (source:https://pub.mdpi-res.com/applsci/applsci-10-03233/article_deploy/html/images/applsci-10-03233-g008.png?1590214485)

2.3. Transformer applications in biomedical imaging

The Visual Transformer introduces a self-attention mechanism to learn the global correlation between image blocks, thus modeling global contextual information. Its early application in medical imaging focused on classification or segmentation tasks in mature domains, demonstrating a significant advantage for diffuse lesions. However, pure Transformer models generally require massive data support and lack the innate inductive bias for local textures—crucial for detecting tiny lesions. Although the hybrid architecture combining CNN's local perception and Transformer's global reasoning has potential, it has not yet been widely tested in sparse-data multimodal scenarios [5].

3. Methodology

3.1. Cnn-transformer hybrid architecture

This framework first processes each modality through a dedicated convolutional backbone: customizing the residual network for CT sections and optimizing the dense connection network for MRI volume data. Each backbone network generates high-dimensional feature maps that preserve spatial details and simultaneously abstract underlying textures and edges. These feature maps were segmented into fixed-size blocks, flattened into labeling sequences, and then the CT and MRI labelings were spliced together [6]. Pre-learned classification labels are used to aggregate global information. The combined label sequence is input to the multi-layer transformer encoder, and its multi-head self-attention layers simultaneously learn intra-modal dependencies (such as the internal models of a single CT portion) and inter-modal correlations (such as the correspondence between CT and MRI features). Location integration ensures coordinated retention of spatial locality and global context.

3.2. Cross-domain transfer learning strategy

To address the domain shift caused by the difference between the scanning device and the acquisition protocol, we integrate the adversarial adaptation module. The fused features output by the transformer are input to the domain discriminator, which is trained to distinguish features from CT sources from those from MRI sources. The discriminator is connected to the backbone network through the gradient inversion layer to invert the gradient signal during backpropagation. The feature extractor thus learns to confuse the discriminator and generates mode-independent embeddings that retain disease-related information while eliminating device artifacts. This conflicting interaction strikes a balance between maximizing classification accuracy and minimizing domain differences, ultimately forming a cross-mechanism generalization model [7].

3.3. Multi-modal fusion module

After transformer encoding, the label streams for each modal and the global classification labels are extracted, respectively. Each label flows through a small feedforward subnetwork to refine the modal-specific and fusion representations. The attention gating unit calculates modal weights by applying sigmoid activation to the linear projections of the classification labels. These weights dynamically adjust the contribution degrees of the CT and MRI branches, allowing the network to focus on modalities with more information for specific cases (e.g., focusing on MRI when comparing wall tissues is critical). The weighted sum of the branch outputs is finally generated by the Softmax classifier to produce disease predictions [8]. This post-fusion design ensures flexibility and robustness for handling variable image quality and lesion characterization.

4. Experimental setup

4.1. Datasets and preprocessing

The evaluation adopted two rare neurological disease datasets: intracranial hemorrhage and demyelinating lesions. Each dataset contains paired CT and MRI scans from different hospitals. Preprocessing includes rigid registration of CT to MRI, intensity normalization to mean zero unit variance, and resampling to a unified pixel size of 1 cubic millimeter. Two-dimensional slices are extracted and adjacent slices are stacked as three-channel inputs [9].

4.2. Evaluation metrics and baselines

Model performance is evaluated by the accuracy rate, precision rate, recall rate, and F1 score of the reserved test set. Three benchmarks are compared: ResNet50, trained solely on CT, DenseNet121, trained solely on MRI, and an independent visual transformer with fused inputs. All benchmarks were adjusted based on ImageNet's pre-trained weights and the same training program was adopted.

4.3. Training details and implementation

The training process is performed over 100 rounds using the Adam optimizer with an initial learning rate of 1e-4, and decays by 0.1 times every 30 rounds. The early stopping strategy is implemented based on the F1 score of the verification set. The batch size is set to 16 slices per mode, and data augmentation is performed by random flipping, rotation, and intensity perturbation. All experiments were conducted in the PyTorch 1.10 environment using the NVIDIA Tesla V100 GPU.

5. Results and discussion

5.1. Performance comparison with baselines

As shown in Table 1, the bimodal model achieved an average accuracy rate of 91.3% and an F1 score of 0.89, which exceeds the baseline of the multimodal model by 7 to 10 percentage points. Although the independent visual transformer achieves reasonable overall modeling (with an accuracy rate of 87.1% and F1= 0.84), its F1 score still lags behind the hybrid architecture by about 5 percentage points. These results confirm the synergistic advantages of CNN's local feature extraction and Transformer's long-distance association capturing capabilities, which are particularly crucial when training data is sparse [10].

Table 1. Performance comparison across models on held-out test sets

Model	Accuracy (%)	Precision	Recall	F1-score
CT-only CNN (ResNet-18)	84.5	0.82	0.80	0.81
MRI-only CNN (DenseNet-201)	83.2	0.81	0.78	0.79
ViT (pure transformer)	87.1	0.86	0.85	0.84
CNN-Transformer (proposed)	91.3	0.90	0.89	0.89

5.2. Ablation study

We quantify the contributions of each component through the targeted suppression module. Removing the transformer encoder reduced the accuracy rate to 84.5% (F1 = 0.82), confirming the critical importance of the overall image block interaction (row 1 of Table 2). Not adapting the adversarial domain reduced the accuracy rate to 86.0% (F1 = 0.84), indicating that aligning the CT and MRI distributions significantly improved performance (row 2 of Table 2).

5.3. Cross-domain generalization and robustness

Further assess cross-domain transfer capability: train only with CT and test on MRI. When domain adaptation was not enabled, the model achieved an accuracy rate of only 72.3% (F1 = 0.74; Row 3 of Table 2). After introducing adversarial alignment, the performance of MRI CT increased to 82.1% (F1= 0.84;) (4) In row 2 of Table 2, it is demonstrated that this framework can learn domain invariant representations. The unillustrated heatmap confirmed that the feature attention was continuously focused on the intermodal lesion boundary.

Table 2. Ablation and cross-domain generalization results

Experiment	Accuracy (%)	F1-score
Full model w/o transformer	84.5	0.82
Full model w/o domain adaptation	86.0	0.84
Train on CT, test on MRI (no adaptation)	72.3	0.74
Train on CT, test on MRI (with adaptation)	82.1	0.84

6. Conclusion

This study proposes an innovative multimodal deep learning framework, which collaboratively integrates CNN local feature extraction, global transformer modeling, and adversarial domain adaptation techniques to achieve rare neurological disease detection. This architecture achieved an accuracy rate of 91.3% and a 0.89 F1 score on low-sampling CT and MRI datasets, which was up to 10 percentage points higher than the baseline data of single-modal and pure transformers. The ablation experiment confirmed that transformer encoder and domain adversarial training are indispensable, and the cross-modal evaluation demonstrated superior generalization ability. With the device-independent feature alignment system, this method breaks the major barriers of multi-institutional clinical deployment. Further research will be extended to three-dimensional volume data, will address semi-supervised learning of unlabeled images and will optimize the efficiency of real-time reasoning in radiological workflows.

References

[1] Lee, J., Liu, C., Kim, J., Chen, Z., Sun, Y., Rogers, J. R., Chung, W. K., & Weng, C. (2022). Deep learning for rare disease: A scoping review. *Journal of Biomedical Informatics*, 135, 104227. https: //doi.org/10.1016/j.jbi.2022.104227

[2] Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., Siddiqui, T., & Mellit, A. (2023). A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability*, 15(7), 5930. https: //doi.org/10.3390/su15075930

[3] Salem, M., Valverde, S., Rovira, À., Salvi, J., Oliver, A., Lladó, X., & Kushibar, K. (2021). Transductive transfer learning for domain adaptation in brain magnetic resonance image segmentation. *Frontiers in Neuroscience*, 15, 608808. https: //doi.org/10.3389/fnins.2021.608808

[4] Hong, J., Yu, S. C.-H., & Chen, W. (2021). Unsupervised domain adaptation for cross-modality liver segmentation via joint adversarial learning and self-learning. *arXiv preprint arXiv: 2109.05664.*

[5] Liu, X., Qiu, H., Li, M., Yu, Z., Yang, Y., & Yan, Y. (2024). Application of multimodal fusion deep learning model in disease recognition. *arXiv preprint arXiv: 2406.18546.*

[6] Segura-Bedmar, I., Camino-Perdonas, D., & Guerrero-Aspizua, S. (2021). Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts. *arXiv preprint arXiv: 2109.00343.*

[7] Zhao, Z., Yeoh, P. S. Q., Zuo, X., Chuah, J. H., Chow, C.-O., Wu, X., & Lai, K. W. (2024). Vision transformer-equipped convolutional neural networks for automated Alzheimer's disease diagnosis using 3D MRI scans. *Frontiers in Neurology*, 15, 1490829. https: //doi.org/10.3389/fneur.2024.1490829

[8] Mutnuri, M. K., Stelfox, H. T., Forkert, N. D., & Lee, J. (2024). Using domain adaptation and inductive transfer learning to improve patient outcome prediction in the intensive care unit: Retrospective observational study. *Journal of Medical Internet Research*, 26, e52730. https: //doi.org/10.2196/52730

[9] Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., … & Fedorov, A. (2021). Transfer learning for domain adaptation in brain lesion segmentation. *Frontiers in Neuroscience,* 15, 608808. https: //doi.org/10.3389/fnins.2021.608808

[10] Segura-Bedmar, I., Camino-Perdonas, D., & Guerrero-Aspizua, S. (2021). Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts. *arXiv preprint arXiv: 2109.00343*.