

Comparative analysis and optimization of an enhanced DenseNet model for multi-modal medical image classification

Zichun Wei

School of Computing and Data Science, Xiamen University Malaysia, Sepang, Malaysia

2371069501@qq.com

Abstract. Medical image classification models often lack validation across diverse datasets, limiting their generalization in clinical settings. This study evaluates and optimizes an enhanced DenseNet-121 model, integrating dilated convolutions and Squeeze-and-Excitation (SE) blocks, for multi-modal medical image classification. We assess its robustness across Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and histopathology datasets, focusing on cross-domain and cross-modality performance. Experiments reveal strong in-domain results but significant degradation in cross-modality tasks (e.g., MRI-to-CT accuracy drops to ~ 0.5). To address this, we propose two strategies: (1) multi-modal joint training, which boosts cross-modality accuracy to 0.87, and (2) Cycle-Consistent Adversarial Networks (CycleGAN)-based modality translation, improving performance to 0.7. Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations confirm the model's focus on clinically relevant regions, enhancing interpretability. Findings highlight the superiority of multi-modal training while demonstrating CycleGAN's utility when target-domain data is scarce. Future work should explore larger multi-center datasets and advanced domain adaptation to further improve robustness.

Keywords: DenseNet optimization, cross-modality generalization, medical image classification, CycleGAN, Explainable AI (Grad-CAM)

1. Introduction

In recent years, the rapid development of neural networks has exhibited exceptional performance across diverse domains, such as computer vision and natural language processing. As the demand for nonlinear representation and complex function approximation keeps increasing, network architectures have evolved toward deeper and more extensive scales to strengthen feature representation and generalization capacities. However, deeper models do not always outperform shallower ones, primarily due to training challenges such as vanishing or exploding gradients, which hinder model convergence.

To address this issue, He et al. [1] proposed the Residual Network (ResNet), which introduces skip connections on top of traditional convolutional layers. This effectively mitigated the vanishing gradient problem, enabling the successful training of networks with hundreds of layers and significantly improving model performance [1]. Building on ResNet, DenseNet further enhances feature utilization by concatenating outputs from all preceding layers, thus leveraging both low-level and high-level features.

Numerous recent models have been developed based on these foundational architectures, enabling deep learning to address tasks more efficiently and effectively. For instance, Mao et al. [2] proposed an improved model based on DenseNet-121, integrating Squeeze-and-Excitation (SE) blocks and dilated convolutions, which significantly improved classification accuracy on brain tumor MRI images.

However, similar to Mao et al.'s model, many newly proposed architectures are only evaluated on a single dataset, lacking validation of their generalization ability and practical applicability. This limitation is particularly crucial in medical imaging, where variations in hospitals, imaging devices, and patient populations result in differences in data characteristics and distributions. Consequently, success on one dataset does not guarantee broad applicability. The lack of cross-dataset validation remains a major limitation in current research. The lack of cross-dataset validation remains a major limitation in current research [3].

To enhance model robustness and cross-domain generalization, we perform generalization tests on the improved DenseNet model proposed by Mao et al. [2], optimize its architecture to improve cross-domain performance, and integrate post-hoc Explainable AI (XAI) techniques to enhance model trustworthiness.

2. Materials and methods

2.1. Datasets and preprocessing

2.1.1. Dataset curation for generalization assessment

Given that the original model was trained and tested on brain tumor MRI images for multi-class classification, our primary objective is to first evaluate its classification performance and generalization capability. To this end, we collected experimental datasets based on combinations of data domain and task, including the same domain with different tasks, different domains with the same task, and different domains with different tasks [4].

The original dataset utilized by the authors is a publicly accessible dataset from Kaggle, comprising 7,023 T1-weighted brain MRI images acquired from three anatomical views. These images are classified into four categories: glioma, meningioma, pituitary tumor, and no tumor.

To test generalization under the same domain with a slightly different task, we additionally collected a classic brain tumor MRI dataset from Kaggle—Br35H. This dataset comprises two balanced categories: "yes" (tumor present) and "no" (no tumor), with 1,500 images per category. (Owing to its limited data source, Br35H is a single-center dataset with single-modality and single-view images, which makes it suitable for evaluating model generalization.)

For the same domain but a different task, we acquired a publicly annotated dataset of brain MRI images for Alzheimer's disease classification. This dataset includes approximately 7,000 images labeled into four severity levels, which provides additional evidence for assessing the model's generalization capability.

To explore generalization across different domains with the same task, we collected brain tumor CT images from public datasets. These images are acquired from three anatomical views and include approximately 5,000 samples, classified into "healthy" and "tumor" categories.

Finally, for different domains and different tasks, we collected histopathological images of small intestinal cancer, totaling 10,000 images, which are divided into "healthy" and "unhealthy" categories.

The features and relationship between the datasets are presented in Table 1.

Table 1. Comparison of feature and relationship between datasets

Data Set	Domain	Task	Relationship with original dataset
Original Dataset	MRI	4-categories classification on Brain Tumor	Baseline
Br35H	MRI	2-categories classification on Brain Tumor	in-domain, similar-task
Alzheimer MRI Disease Classification Dataset	MRI	4-categories classification on Alzheimer Disease	in-domain, different-task
Brain tumor CT image	CT	2-categories classification on Brain Tumor	different-domain, same-task
LC25000-colon histopathological images	Histopathological images	2-categories classification on colon adenocarcinomas	different-domain, different-task

Prior to model training, we applied image preprocessing techniques consistent with those in the original paper, including the removal of irrelevant regions, image resizing, and Gaussian blurring. These steps improve data diversity and expand the dataset size, which helps prevent model overfitting. The training set was split into a 9:1 ratio for training and validation, where the validation subset is used for model selection and hyperparameter tuning to avoid overfitting to the training data. For the test set, we adopted a strategy of generating 10 random crops per image and averaging the corresponding predictions, which reduces randomness and improves model robustness.

2.1.2. Feature of dataset

Dataset characteristics, such as modality, dimensionality, sample size, class balance, acquisition center, and protocol differences, can all affect model training processes and outcomes. Since all our datasets are publicly accessible and composed of 2D images, we exclude dimensionality and protocol factors etc. from our consideration. Instead, we focus on modality and sample size, given that modality differences can significantly impact feature representation, preprocessing procedures, and transferability. What's more, small sample sizes may lead to overfitting, while class imbalance can cause inflated accuracy due to majority class dominance [5].

2.2. Enhanced densenet model architecture

Our model is improved based on DenseNet-121 by integrating dilated convolution and Squeeze-and-Excitation (SE) mechanisms. Specifically, we modify the dense blocks by applying dilation rates based on the Hybrid Dilated Convolution strategy to the originally non-dilated convolutional layers. This expands the receptive field without significantly increasing the number of parameters or computational complexity, thereby enabling more effective multi-scale feature extraction.

Additionally, we insert SE blocks after the convolutional layers within each dense block. This introduces a channel-wise attention mechanism, which explicitly models the inter-channel dependencies. By learning and applying adaptive weights, the model recalibrates the importance of each feature map, thus enhancing its representational capacity along the channel dimension.

2.3. Experimental methodology

Our experiments consist of four main steps:

1. In-domain evaluation: We trained and tested the model on datasets from different diseases within the same domain (MRI) to assess its robustness and generalization across disease types.

2. Cross-dataset evaluation: We trained the model on one brain tumor MRI dataset, followed by testing it on another brain tumor MRI dataset and a brain tumor CT dataset. This setup evaluates both cross-dataset and cross-modality generalization.

3. Cross-modality optimization: We explore two strategies to improve cross-modality generalization.

The first employs CycleGAN-based modality translation to reduce appearance discrepancies between MRI and CT, thereby enhancing the transferability of the model. The CycleGAN architecture includes two generators (MRI→CT and CT→MRI) and two discriminators, which are trained adversarially with cycle-consistency loss to preserve anatomical structure. [6]. We trained CycleGAN using CT data and a small subset of MRI images (from Br35H), then applied the MRI→CT generator to produce pseudo-CT images for testing CT-trained classifiers

The second strategy involves multi-modal joint training, aiming to improve generalization when multiple modalities are available.

4. Post-hoc explainability: We integrate Grad-CAM to visualize model attention. The last convolutional layer is selected as the target layer, and heatmaps are generated using ClassifierOutputTarget function. Random samples from the test set (YES/NO) are used to produce Grad-CAM visualizations, which are overlaid on the original images to analyze model focus areas. This enhances clinical interpretability and trust in the model's decisions.

3. Results and discussion

3.1. In-domain classification performance

We conducted in-domain evaluations on four types of datasets: brain tumor MRI, brain tumor CT, Alzheimer's MRI, and colon histopathological images. The model performed well across all tasks. Training typically converged within 10 epochs, with the loss approaching zero and the accuracy nearing 1.0. Validation and training curves were closely aligned, which indicates the model's strong learning and generalization capabilities.

For example, the colon dataset achieved an F1 score, recall, and accuracy of up to 0.99, while brain tumor MRI and CT datasets achieved F1 scores between 0.92 and 0.95. Detailed results are shown in Figure 1.

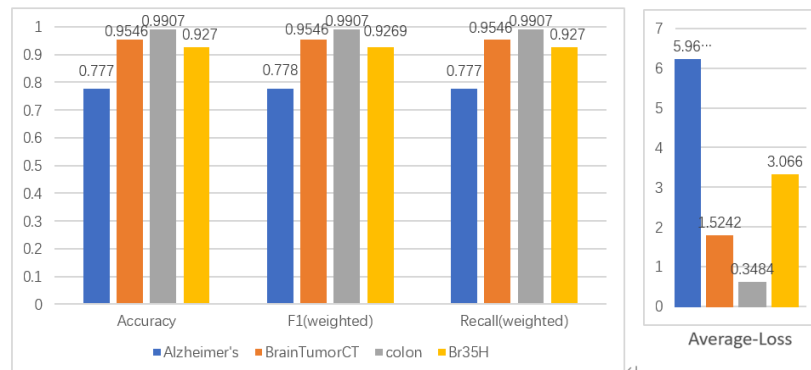


Figure 1. Comparison of test performance and average loss: Alzheimer's MRI dataset (blue), brain tumor CT dataset (orange), colon histopathology dataset (gray), and Br35H dataset (yellow)

3.2. Cross-dataset and cross-modality generalization analysis

In our cross-dataset evaluation, the source domain consisted of brain tumor MRI images (Br35H), while the target domains included another brain tumor MRI dataset and a brain tumor CT dataset. Results showed that the model maintained strong classification performance in the MRI-to-MRI setting, with an F1 score and accuracy of up to 0.98. However, the model's performance dropped significantly in MRI-to-CT and CT-to-MRI mappings, with the accuracy and F1 scores close to 0.5—a result indicating poor feature transfer and near-random classification.

This comparison highlights that models trained on one modality perform much worse when being tested on a different modality, suggesting limited cross-modality generalization. Therefore, we proceeded to optimize the model's cross-modality generalization capability.

3.3. Efficacy of cross-modality optimization strategies

In clinical practice, brain tumors are often imaged using CT, a modality that is more accessible and cost-effective than MRI. From the patient's perspective, CT scans are also more affordable. If our model could be trained on CT images and still perform well on MRI data, it would offer practical benefits for both patients and hospitals, thereby potentially improving diagnostic efficiency and overall healthcare quality.

3.3.1. Multi-modal joint training

We first trained the model using mixed modalities—4,000 CT images and 800 MRI images—and tested it on MRI data. The performance improved significantly, with accuracy rising from 0.5 to 0.87, demonstrating effective optimization of our model and its strong generalization capability.

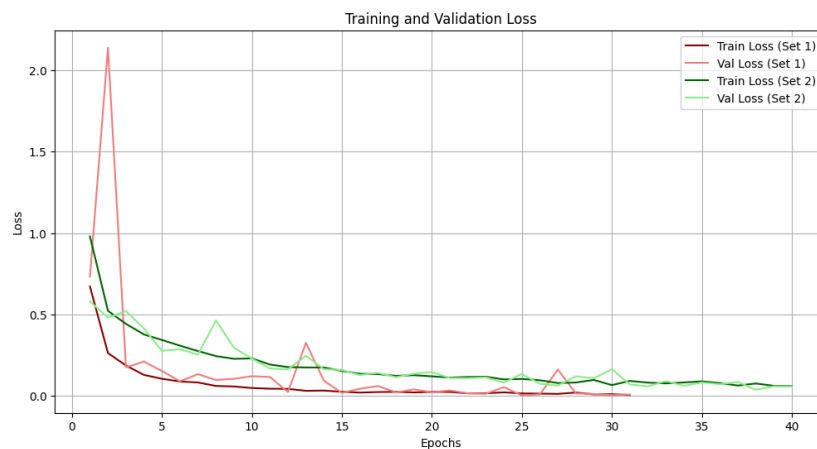


Figure 2. Loss curve comparison: Set1 (trained with pure MRI data) vs. Set2 (trained with mixed CT and MRI data)

As is shown in Figure 2 that the red lines represent the model trained with only MRI images, while the green lines show the model trained with mixed-modality images. As shown, the mixed-modality model converges more slowly, with a higher initial loss value. This is likely due to the richer and more diverse feature information in mixed data, which requires more learning time. However, both models eventually converge to a low loss value, indicating that both models can effectively learn from different modalities. Test results also confirm that the mixed-modality model performs well in multi-modal classification tasks.

3.3.2. CycleGAN-based modality translation

Next, we introduced a CycleGAN model to translate MRI images into CT-like representations. These pseudo-CT images were then used to test a model trained solely on CT data. The F1 score and accuracy improved from 0.5 to 0.7, indicating that the CycleGAN effectively reduced the domain gap between MRI and CT modalities. Through unsupervised image-to-image translation, the MRI images became more similar to CT images in terms of grayscale distribution and texture features, thereby enhancing cross-modality recognition performance.

3.3.3. Comparative analysis of optimization strategies

The following Figure 3 shows the test results of two optimized models and the base model (trained on CT images). By comparing the results of the two optimization strategies, we found that mixed-modality training outperformed the CycleGAN-based approach, but it requires access to both CT and MRI data during training. This finding highlights that synthetic domain adaptation (CycleGAN) can improve cross-modality generalization only when the model has no direct exposure to the target domain. When sufficient multimodal data are available, direct multi-domain training remains the superior strategy, and the use of synthetic intermediate images may even be detrimental due to distributional inconsistencies between synthetic and real data.

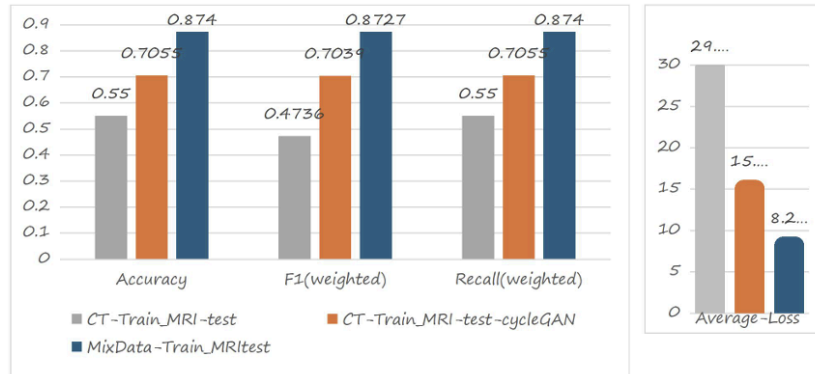


Figure 3. Comparison of test performance and average loss: CT-trained model on MRI (gray), CycleGAN-transformed MRI (orange), and mixed-modality training (blue)

3.4. Model interpretability via Grad-CAM

Finally, we applied the post-hoc XAI method Grad-CAM to visualize the model's decision-making process during testing. This helped us understand the specific regions the model focused on when making predictions. The specific heatmaps and original images are shown in Figure 4 and Figure 5.

For "YES" samples in the test set, the Grad-CAM heatmap (right) revealed a strong activation in the upper-right region, which corresponded to the suspected lesion area. This suggests that the model relied on tumor-related regions when predicting "YES," and that its attention aligned closely with clinically relevant areas, thereby demonstrating strong interpretability.

Conversely, for "NO" samples, although some high-response points appeared, they were dispersed around normal brain tissue edges rather than concentrated in typical tumor regions. This implies that the model did not identify tumor-like features and instead conducted a holistic assessment of the image, which supports the "NO" prediction.

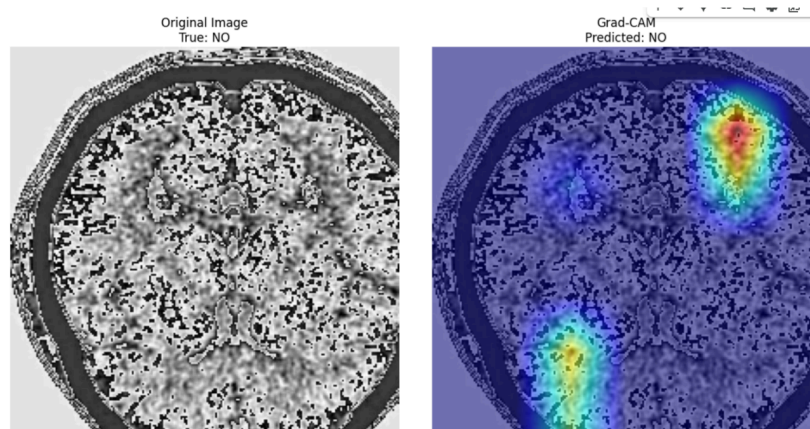


Figure 4. "NO" type heatmap and original figure

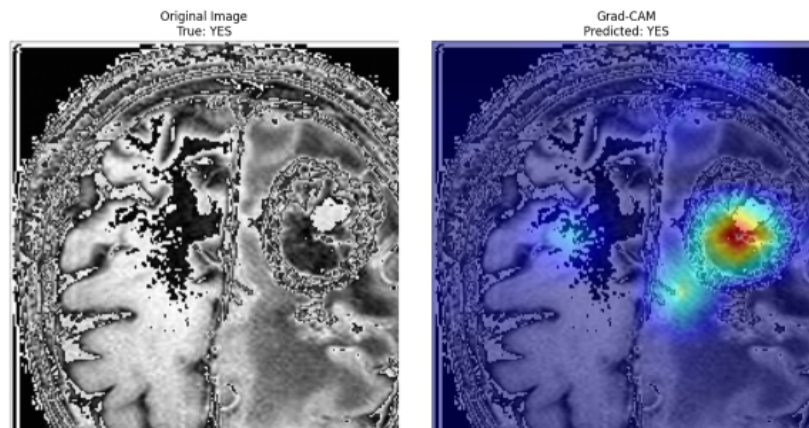


Figure 5. "YES" type heatmap and original figure

4. Conclusion

This work evaluated and optimized the cross-domain and cross-modality generalization capabilities of a previously proposed DenseNet-121 variant enhanced with dilated convolution and squeeze-and-excitation mechanisms. Building on this improved backbone, we carried out extensive experiments across multiple medical imaging datasets to test its robustness beyond the original single-dataset setting. In-domain results confirmed strong performance when training and testing within the same modality, but accuracy decreased sharply when transferring between MRI and CT, highlighting the challenge of domain shift in real clinical scenarios.

To mitigate this gap, we explored two optimization strategies for improvement. Multimodal joint training, using both MRI and CT data, achieved the greatest gain, elevating cross-modality accuracy from approximately 0.5 to 0.87. CycleGAN-based modality translation also improved performance, to around 0.7, by reducing the appearance discrepancies between modalities. Grad-CAM visualization further showed that the model consistently focused on tumor-related regions, thereby supporting the reliability of its predictive outputs..

Our study, however, has limitations. The datasets still lack sufficient size and diversity, and the improved DenseNet was tested within a constrained parameter space. Moreover, CycleGAN outputs were only used for testing rather than being integrated into the training process. Future work could explore combining real CT images with CycleGAN-generated pseudo-CT images for joint training to enhance feature learning effectiveness, or apply advanced domain adaptation and larger multi-center datasets to enhance robustness. These directions would further improve the optimized model's generalization ability and its practical utility in heterogeneous medical imaging environments.

References

- [1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). <https://doi.org/10.1109/CVPR.2016.90>
- [2] Mao, Y., Kim, J., Podina, L., & Kohandel, M. (2025). Dilated SE-DenseNet for brain tumor MRI classification. *Scientific Reports*, *15*(1), 1–10. <https://doi.org/10.1038/S41598-025-86752-Y>; SUBJMETA
- [3] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, *15*(11), e1002683. <https://doi.org/10.1371/JOURNAL.PMED.1002683>
- [4] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.
- [5] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. <https://doi.org/10.1016/J.MEDIA.2017.07.005>
- [6] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2242–2251). <https://doi.org/10.1109/ICCV.2017.244>