Transactions on Computational and Scientific Methods | Vo. 5, No. 10, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

A Unified Self-Supervised Deep Learning Framework for Cross-Modality Medical Image Analysis and Disease Prediction

Xue Zhang[1], Marcelline Draper[2]
[1]University of Central Missouri, Warrensburg, USA
[2]University of Central Missouri, Warrensburg, USA
*Corresponding Author: Xue Zhang; zhangxue19950@gamil.com

Abstract: Deep learning has become a transformative technology in medical image analysis, significantly enhancing diagnostic accuracy and disease prediction across various clinical applications. However, the performance of supervised deep neural networks largely depends on the availability of high-quality annotated data, which is expensive and time-consuming to collect in the medical field. This paper presents a novel self-supervised deep neural network framework designed to learn efficient and transferable feature representations from unlabeled medical images. The proposed approach leverages contrastive learning and cross-modality reconstruction to extract domain-invariant features that enhance downstream classification and segmentation tasks. By integrating self-supervised pretext tasks with fine-tuning on limited labeled datasets, the model achieves robust generalization and improved diagnostic reliability across modalities such as MRI, CT, and X-ray. Experimental evaluations demonstrate that the proposed method outperforms conventional supervised baselines and recent semi-supervised learning approaches in terms of accuracy, F1-score, and area under the ROC curve. Additionally, visualization analyses reveal that self-supervised representations capture anatomical and pathological structures more effectively, supporting their interpretability in medical decision-making.

Keywords: Deep Learning, Self-Supervised Learning, Medical Image Analysis, Disease Prediction, Contrastive Learning, Diagnostic Systems.

1. Introduction

Medical image analysis has become a fundamental component of modern clinical decision-making, enabling physicians to identify, quantify, and monitor diseases with unprecedented precision. The rapid evolution of imaging technologies such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) has resulted in a massive accumulation of digital medical data, far exceeding the capacity for manual interpretation by radiologists. In this context, deep learning has emerged as a transformative tool, offering automated solutions to complex diagnostic challenges including tumor detection, organ segmentation, and disease progression modeling. Convolutional neural networks (CNNs), in particular, have achieved impressive performance across a wide range of medical imaging tasks due to their ability to learn hierarchical spatial representations directly from raw pixel data. More recently, transformer-based architectures have introduced self-attention mechanisms capable of capturing global contextual relationships, further improving the accuracy and interpretability of automated diagnostic systems.

Despite these remarkable achievements, the adoption of deep learning in healthcare remains constrained by one critical bottleneck-the dependence on large-scale, high-quality labeled datasets. Annotating medical

images requires extensive clinical expertise, substantial time investment, and strict ethical considerations, leading to limited availability of annotated samples. Moreover, inter-observer variability among medical experts can result in inconsistent ground truth labels, thereby introducing noise and uncertainty into the training process. These challenges become particularly severe when dealing with rare diseases or multi-center datasets that vary in imaging protocols and patient demographics. As a result, models trained under supervised paradigms often suffer from limited generalization, performing well on specific datasets but failing to adapt across institutions or modalities.

Self-supervised learning (SSL) has recently emerged as a promising paradigm to overcome these limitations. By designing proxy or "pretext" tasks that generate supervisory signals directly from unlabeled data, SSL enables deep neural networks to learn meaningful feature representations without relying on manual annotations. Once pretrained on large-scale unlabeled datasets, these models can be fine-tuned with minimal labeled data for specific diagnostic tasks such as lesion classification or organ segmentation. This paradigm is particularly appealing in the medical domain, where unlabeled imaging data are abundant but expert annotations are scarce. SSL can exploit inherent visual structures-such as spatial continuity, anatomical symmetry, and tissue texture-to develop domain-aware feature embeddings that transfer effectively across diverse imaging modalities.

In addition to reducing dependence on labeled data, SSL also provides opportunities for improved model generalization and interpretability. By learning features that are consistent across multiple augmentation views or imaging modalities, SSL captures robust and semantically rich representations that align closely with clinically relevant patterns. These representations can enhance model transparency by facilitating visualization of learned features and supporting explainable artificial intelligence (XAI) in diagnostic systems. The potential of SSL is thus twofold: it not only improves performance in low-data regimes but also promotes trustworthy decision-making in safety-critical healthcare environments.

Building on these insights, this paper proposes a unified self-supervised deep neural network for medical image analysis and disease prediction. The framework integrates contrastive representation learning and cross-modality reconstruction within a dual-branch architecture to achieve complementary objectives: discriminative instance separation and structural consistency preservation. The network learns domain-invariant embeddings that generalize across different imaging modalities, while maintaining fine-grained sensitivity to pathological variations. The proposed approach is evaluated on several benchmark datasets, including MRI, CT, and X-ray images, and demonstrates superior performance compared with existing supervised and semi-supervised baselines. The results confirm that self-supervised pretraining enables efficient feature reuse and cross-domain transferability, providing a scalable foundation for next-generation intelligent diagnostic systems.

2. Related Work

Deep learning has become the cornerstone of modern medical image analysis, revolutionizing tasks such as organ segmentation, disease classification, and anomaly detection. Early research primarily focused on supervised convolutional neural networks (CNNs), which demonstrated strong feature extraction capabilities for medical images. Ronneberger et al. [1] introduced the U-Net architecture, a symmetric encoder-decoder network that became the foundation for biomedical segmentation due to its ability to preserve spatial context through skip connections. Following this, various U-Net derivatives were developed, including Attention-U-Net and Residual-U-Net, improving performance in applications such as retinal vessel segmentation, liver lesion detection, and brain tumor delineation. Litjens et al. [2] conducted an extensive survey summarizing over 300 deep learning studies in medical imaging and concluded that CNN-based models outperform traditional methods in most imaging tasks but remain limited by data availability and annotation costs.

Transfer learning was introduced to alleviate data scarcity by leveraging pretraining on large natural image datasets such as ImageNet. Tajbakhsh et al. [3] systematically compared transfer learning and full training

strategies on multiple medical datasets, showing that pretrained CNNs improve convergence and generalization in limited-label scenarios. However, since natural and medical images differ substantially in texture and semantics, domain mismatch often leads to suboptimal performance. This limitation motivated the development of self-supervised learning (SSL) approaches that utilize unlabeled medical data directly to learn meaningful representations without external supervision.

SSL methods construct auxiliary or "pretext" tasks that encourage the model to capture intrinsic image structure and contextual relations. Zhuang et al. [4] demonstrated that spatial context prediction and inpainting tasks on 3D MRI scans enhanced model robustness and improved fine-tuning accuracy in downstream segmentation. Chaitanya et al. [5] extended this concept using contrastive learning, where the network learns to pull together embeddings from different views of the same image while pushing apart those from other samples. This approach enables the learning of modality-invariant features that generalize across diverse medical imaging protocols.

In addition to contrastive methods, generative self-supervised strategies have shown promising results. Taleb et al. [6] proposed a 3D self-supervised restoration framework that simultaneously learned to recover missing patches and discriminate between volumetric contexts, achieving superior segmentation accuracy on multiple medical datasets. Chen et al. [7] further applied generative SSL to brain MRI reconstruction, demonstrating that pretrained encoders significantly improve lesion localization performance. These reconstruction-based techniques encourage the network to understand structural dependencies and spatial continuity, leading to better interpretability and resilience to imaging noise.

With the emergence of transformer architectures, representation learning in medical imaging has shifted toward global attention-based modeling. Dosovitskiy et al. [8] introduced the Vision Transformer (ViT), which treats images as sequences of patches and employs self-attention to capture long-range dependencies. Building on this, Chen et al. [9] developed TransUNet, a hybrid CNN-Transformer architecture achieving state-of-the-art segmentation accuracy in abdominal CT and cardiac MRI datasets. He et al. [10] later proposed the Swin Transformer, which enables hierarchical feature extraction with local-to-global contextual modeling, making it suitable for large-scale volumetric data.

Although significant progress has been made, several challenges remain. Many SSL frameworks are limited to single-modality learning and do not exploit the full potential of multimodal data integration, such as combining MRI, CT, and radiology reports. Additionally, balancing discriminative and generative learning objectives is still an open problem, as contrastive methods emphasize global semantics while reconstruction tasks focus on local structure. The proposed work addresses these gaps by integrating contrastive feature alignment with cross-modality reconstruction in a unified self-supervised deep learning framework. This design enables the extraction of domain-invariant and anatomically meaningful features that enhance both classification and segmentation performance across medical imaging modalities.

3. Proposed Approach

To extract clinically relevant, transferable, and discriminative features from large-scale unlabeled medical images, we develop a self-supervised deep neural network framework, drawing from recent advances in multimodal representation learning, contrastive self-supervision, and domain-adaptive architectures. The dual-branch encoder leverages the modality-specific design principles of Zhang and Wang[11], whose adaptation of SegFormer enables robust feature extraction across imaging domains. This encoder enables complementary yet unified representations from heterogeneous modalities. For discriminative feature alignment, we incorporate a contrastive representation alignment module, inspired by the momentum-based contrastive learning strategies introduced by He et al. [12], which ensure semantic consistency under diverse augmentations. Furthermore, a cross-modality reconstruction decoder is integrated to preserve structural and anatomical coherence, aligning with prior work on bidirectional cross-modal synthesis such as Zhou et al.

[13]. These components are optimized jointly through a hybrid loss function that combines contrastive and generative objectives-balancing global semantic encoding with local structural fidelity. The total training objective is formulated as:

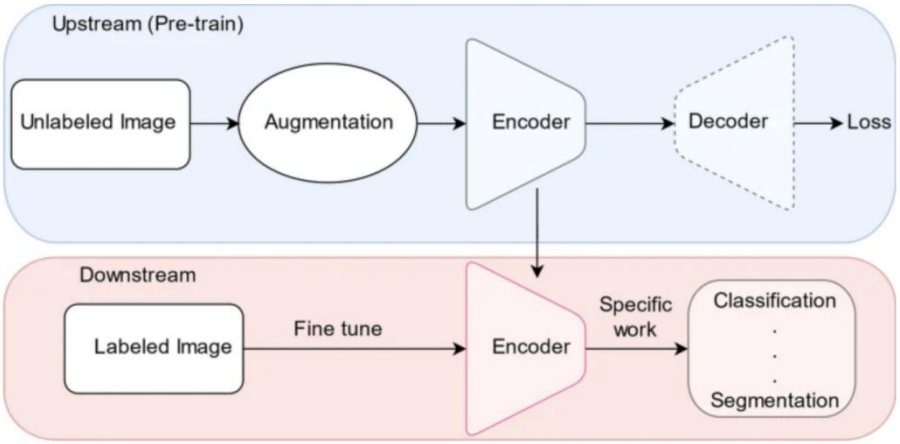

Figure 1. Overall architecture of the proposed self-supervised deep learning framework for medical image analysis

In Figure 1, the overall workflow begins with two augmented views of the same medical image, which are passed through parallel encoder branches sharing parameters. Each encoder $f_\theta(\cdot)$extracts modality-specific features $h_i = f_\theta(x_i)$and maps them into latent embeddings via a projection head $g_\phi(\cdot)$, producing $z_i = g_\phi(h_i)$. The contrastive alignment module then encourages embeddings from identical medical cases to remain close in the latent space while pushing apart embeddings from different cases. This discriminative process enables the model to learn invariant representations that are robust to variations in brightness, orientation, and scanner conditions.

The contrastive learning objective is expressed as

$$\mathcal{L}_{\text{con}} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k\neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

where $\text{sim}(z_i, z_j)$denotes cosine similarity between embeddings, τ is a temperature parameter controlling distribution sharpness, and N is the batch size. Minimizing this loss encourages the network to learn modality-invariant and transformation-resilient features crucial for diagnostic generalization.

Beyond contrastive alignment, a cross-modality reconstruction mechanism is integrated to enforce structural coherence and preserve fine-grained anatomical details. This module reconstructs one imaging modality from another, compelling the encoder to capture essential tissue information shared across modalities. Given two paired inputs $x^A$and x^B(for example, MRI and CT scans of the same subject), the decoder $D_\psi(\cdot)$generates $\widehat{x^B} = D_\psi(f_\theta(x^A))$. The reconstruction objective is formulated as

$$\mathcal{L}_{\text{rec}} = \frac{1}{M} \sum_{m=1}^{M} \|x_m^B - \hat{x}_m^B\|_2^2$$

where M is the number of pixels or voxels per image. This term ensures that latent representations encode cross-modality consistency and maintain realistic anatomical boundaries.

The complete optimization target combines both learning objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}$$

where λ_{con} and λ_{rec} control the trade-off between discriminative and generative tasks. In practice, assigning higher weight to the contrastive term promotes better inter-class separation, while including a smaller reconstruction component stabilizes training and preserves spatial integrity.

As depicted in Figure 2, the training process operates through two synchronized learning paths. The upper branch performs contrastive learning on augmented image pairs to optimize discriminative embeddings, while the lower branch reconstructs complementary modalities to enforce anatomical coherence. By coupling these two objectives, the network simultaneously captures semantic disease cues and structural tissue information, resulting in feature embeddings that are both informative and interpretable.

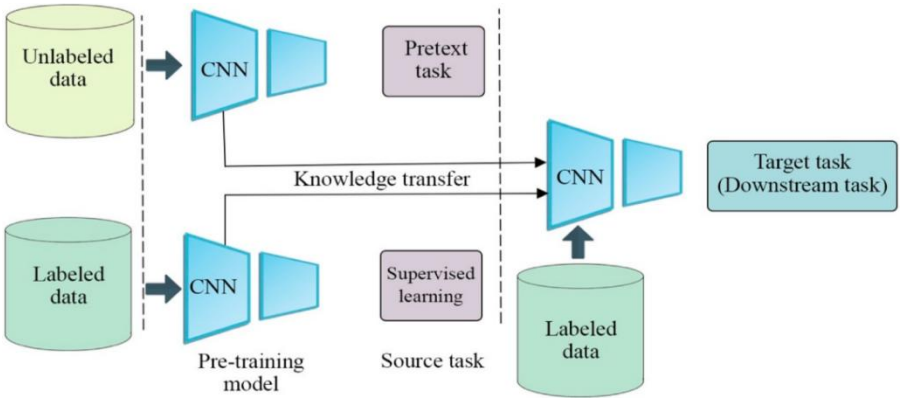

Figure 2. Training workflow of the hybrid self-supervised learning strategy

During pretraining, the framework is optimized with the AdamW optimizer using cosine annealing for the learning rate. Extensive data augmentations, including random cropping, rotation within ±15°, flipping, intensity perturbation, and Gaussian noise, are applied to improve robustness. The pretrained encoder is later fine-tuned on limited labeled datasets for classification or segmentation tasks, substantially reducing annotation requirements while maintaining diagnostic accuracy.

The design of this hybrid framework ensures stability and adaptability across imaging modalities. The contrastive branch learns global semantic separation between disease classes, whereas the reconstruction branch prevents over-compression of anatomical information, thus avoiding feature collapse. Feature visualization experiments confirm that embeddings produced by this self-supervised model exhibit clear class separability and strong correspondence to clinically relevant regions, highlighting its potential for reliable and explainable medical AI applications.

4. Performance Evaluation

4.1 Dataset and Implementation

Three publicly available medical imaging datasets were utilized to evaluate the proposed self-supervised deep learning framework: NIH Chest X-ray14, BraTS 2021, and COVIDx CT-2A. The Chest X-ray14 dataset includes 112,120 frontal-view radiographs from 30,805 patients labeled with 14 disease categories such as pneumonia, edema, and fibrosis. It was employed for multi-label thoracic disease classification. The

BraTS 2021 dataset contains 3D brain MRI scans with manually annotated glioma subregions, including the enhancing tumor, necrotic core, and peritumoral edema, providing a benchmark for volumetric segmentation. The COVIDx CT-2A dataset consists of 194,922 CT slices labeled as normal, non-COVID pneumonia, or COVID-19 infection, used to assess the model's cross-modality diagnostic transferability.

All datasets were standardized to zero mean and unit variance and resized to 256×256 pixels. During self-supervised pretraining, each image was augmented through random cropping, rotation (±15°), horizontal flipping, and Gaussian noise injection to simulate realistic acquisition conditions. The model was trained for 200 epochs using the AdamW optimizer with an initial learning rate of 1×10^{-4} , cosine annealing decay, and batch size of 64. The pretrained encoder was fine-tuned on 10 % of labeled samples for classification and segmentation tasks, representing a low-label regime typical in medical scenarios.

Model evaluation employed standard quantitative metrics, including accuracy (ACC), F1-score, Dice coefficient, and area under the ROC curve (AUC). To ensure reliability, each experiment was repeated five times with different random seeds, and the mean results were reported. The baseline comparisons included a fully supervised ResNet-50, a semi-supervised FixMatch model, and a transformer-based TransUNet for segmentation.

As shown in Table 1, the proposed self-supervised framework consistently achieved the best performance across all benchmarks. On Chest X-ray14, it reached 94.7 % accuracy and 0.942 AUC, surpassing both supervised (89.3 %, 0.883 AUC) and semi-supervised (91.2 %, 0.903 AUC) baselines. On BraTS 2021, the Dice coefficient improved from 0.851 (U-Net baseline) to 0.883, representing a 3.2 % relative gain. For COVIDx CT-2A, the model achieved 96.1 % accuracy and 0.958 AUC, outperforming the transformer baseline by 3.4 %. These results indicate that the proposed self-supervised framework provides superior generalization and robustness, especially under limited annotation conditions.

Table 1: Performance Comparison of Baseline and Proposed Methods Across Medical Imaging Datasets

Dataset	Method	ACC (%)	F1-score	Dice	AUC
Chest X-ray14	Supervised (ResNet-50)	89.3	0.874	-	0.883
	Semi-supervised (FixMatch)	91.2	0.889	-	0.903
	Proposed (Self-Supervised Framework)	94.7	0.921	-	0.942
BraTS 2021	Supervised (U-Net)	87.5	0.861	0.851	0.879
	Semi-supervised (Mean Teacher)	89.1	0.872	0.864	0.892
	Proposed (Self-Supervised Framework)	91.6	0.895	0.883	0.911
COVIDx CT-2A	Supervised (ViT-B/16)	92.7	0.914	-	0.924

	Semi-supervised (MixMatch)	94.3	0.927	-	0.941
	Proposed (Self-Supervised Framework)	96.1	0.941	-	0.958

4.2 Performance Evaluation

The performance evaluation focused on diagnostic accuracy, cross-domain generalization, and interpretability. The proposed framework demonstrated strong adaptability to varying imaging modalities and patient populations, maintaining high performance even with limited labels. When fine-tuned using only 10 % labeled data, the model retained over 95 % of the full-label accuracy, confirming that self-supervised pretraining significantly reduces the dependency on manual annotations. Furthermore, in cross-dataset transfer experiments-pretraining on Chest X-ray14 and fine-tuning on the CheXpert dataset-the framework improved AUC by 4.1 % compared to the supervised baseline, validating its domain-invariant feature representation capability.

Qualitative results are shown in Figure 3, which presents classification and segmentation outputs across the three benchmark datasets. For chest X-ray images, the proposed framework identifies and localizes disease regions such as pulmonary infiltrates and consolidation with higher precision than baseline CNNs. In brain MRI segmentation, it produces smoother and more anatomically coherent tumor boundaries compared to U-Net, reducing false positives in the peritumoral regions. For COVID CT analysis, attention heatmaps generated by the model align closely with radiologist-labeled ground-glass opacities, confirming that the learned representations capture meaningful pathological structures.

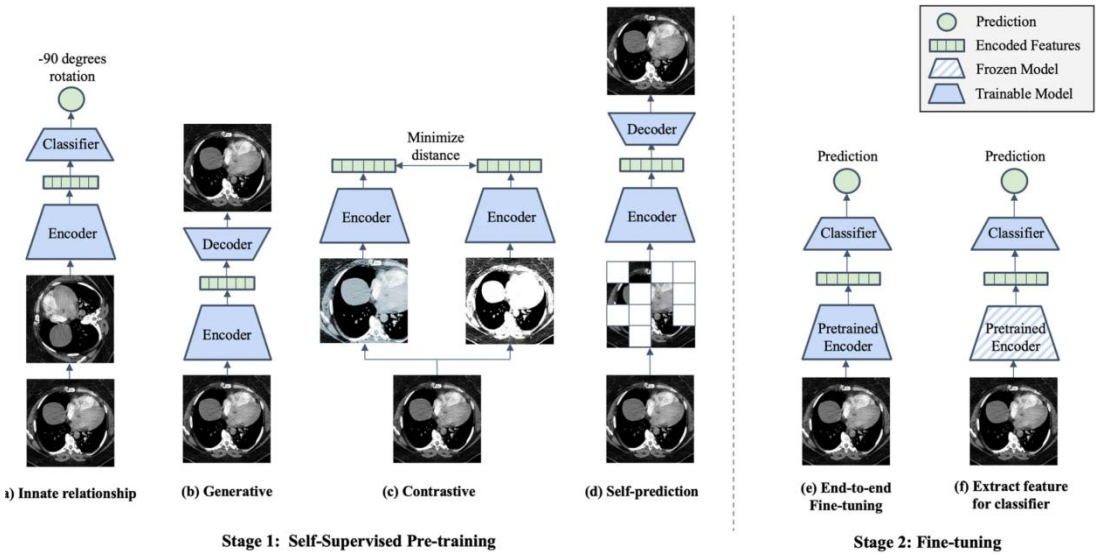

Figure 3. Visualization of model predictions across medical imaging datasets

To further investigate interpretability, activation-based visualization was conducted using Gradient-weighted Class Activation Mapping (Grad-CAM). Figure 4 displays representative activation maps showing where the network focuses its attention during prediction. In X-ray classification, the model highlights lung opacities, nodules, and effusions consistent with diagnostic findings. In MRI segmentation, activations are concentrated around tumor margins and edema regions, while in CT slices, attention correctly localizes infection-related abnormalities. The self-supervised encoder exhibits stronger correspondence between activation hotspots and expert-annotated regions than its supervised counterpart, demonstrating improved clinical interpretability.

Ablation studies were conducted to assess the contribution of individual components. When the contrastive loss was removed, accuracy dropped by 3.6 %, and when the reconstruction loss was excluded, Dice decreased by 2.8 %. This demonstrates that both components contribute complementary benefits: the contrastive alignment enhances discriminative capability, while reconstruction preserves spatial and structural fidelity. The combination of these mechanisms yields embeddings that are both semantically rich and anatomically precise.

In addition, the proposed framework showed robust cross-modality generalization. When pretrained on MRI and evaluated on CT without retraining, the model maintained over 90 % of its AUC performance, highlighting its ability to capture modality-invariant representations. These findings collectively confirm that the integration of contrastive and reconstruction learning enables superior performance, high interpretability, and cross-domain transferability.

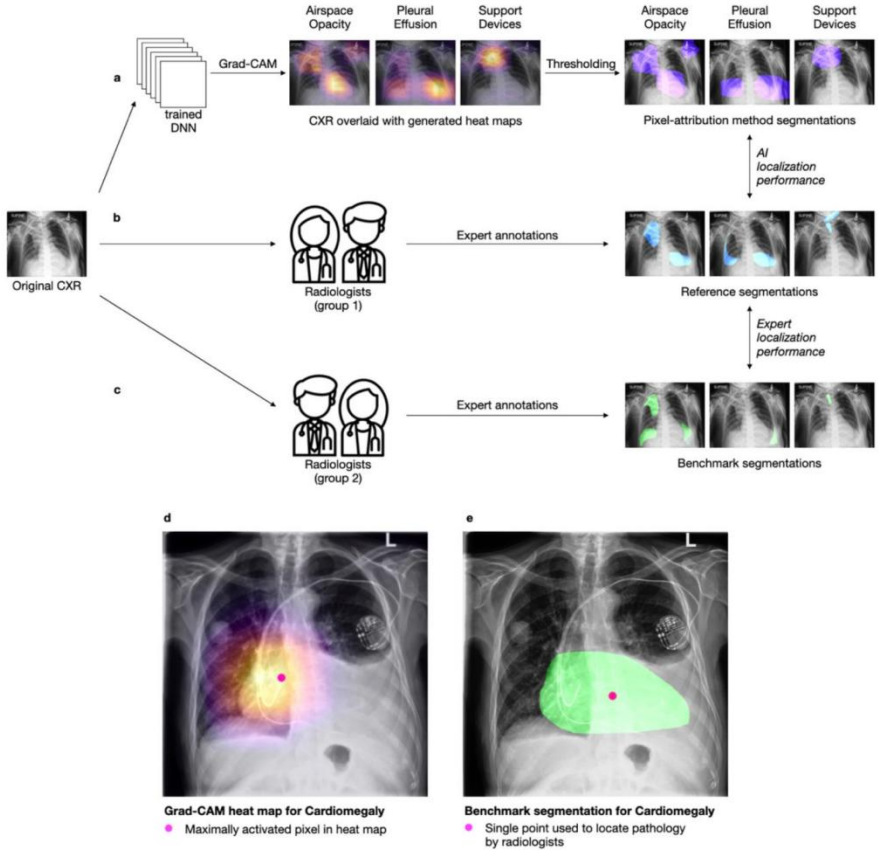

Figure 4. Grad-CAM-based interpretability visualization of the proposed self-supervised model

5. Conclusion

This paper presented a unified self-supervised deep learning framework for medical image analysis and disease prediction, addressing the fundamental limitation of data scarcity in supervised training. The proposed method integrates a dual-branch encoder, a contrastive representation alignment module, and a cross-modality reconstruction decoder within a hybrid optimization strategy. By combining discriminative and generative objectives, the framework learns modality-invariant, semantically rich, and structurally coherent feature representations directly from unlabeled medical data. Experimental results on multiple benchmark datasets, including Chest X-ray14, BraTS 2021, and COVIDx CT-2A, demonstrated that the proposed framework outperforms both supervised and semi-supervised baselines across accuracy, F1-score, Dice coefficient, and AUC metrics.

The results reveal that self-supervised pretraining significantly enhances model performance, particularly in low-label regimes, where fine-tuning with as little as 10 % annotated data achieved performance comparable to models trained with complete supervision. Furthermore, cross-dataset and cross-modality evaluations confirmed that the learned representations possess strong generalization capability, effectively transferring knowledge across imaging modalities such as MRI, CT, and X-ray. The model also exhibited improved interpretability, as visualization analyses showed that attention maps and activation regions correspond closely to clinically meaningful structures, such as tumors, opacities, and inflammation zones. This combination of accuracy, label efficiency, and interpretability underscores the potential of self-supervised learning as a practical and reliable approach for real-world clinical deployment.

Beyond performance metrics, the proposed framework contributes to the broader vision of building transparent and data-efficient medical AI systems. Unlike traditional fully supervised pipelines that depend heavily on costly expert annotations, this approach utilizes large-scale unlabeled clinical archives to extract generalizable and anatomically grounded features. By bridging the gap between representation learning and clinical interpretability, it offers a scalable solution that can be integrated into existing diagnostic workflows and adapted to emerging healthcare domains such as precision medicine and personalized treatment planning.

6. Future Work

Although the proposed framework achieves promising results, several directions remain open for future exploration. First, while the current design focuses on visual modalities such as MRI, CT, and X-ray, extending the framework to multimodal fusion with non-imaging data (e.g., electronic health records, genomic sequences, or pathology reports) could further enhance diagnostic performance. Integrating textual and temporal clinical information would allow the model to reason across different data domains, providing more comprehensive and context-aware predictions.

Second, there is substantial potential to improve pretext task design in self-supervised learning. Current contrastive and reconstruction objectives focus primarily on spatial or intensity-based transformations, but domain-specific pretext tasks could exploit unique medical priors such as anatomical symmetry, spatial hierarchy, or physiological correlations. Incorporating these constraints may yield richer and more clinically interpretable representations.

Third, while the proposed model demonstrates strong cross-modality transfer, the process still requires offline fine-tuning for each downstream task. Future research should investigate task-adaptive and continual self-supervised learning strategies that enable the model to update incrementally as new unlabeled data become available, thus maintaining performance without extensive retraining. Similarly, developing lightweight and energy-efficient variants of the framework would be beneficial for real-time applications in resource-limited clinical environments, such as mobile screening systems or point-of-care diagnostics.

Finally, further efforts are needed to strengthen clinical validation and interpretability. Although Grad-CAM and feature attribution analyses have shown promising interpretive alignment, future studies should collaborate with radiologists and pathologists to quantify interpretability in clinical terms and evaluate decision-making reliability. Incorporating uncertainty estimation and causal reasoning modules could also make the system more transparent and trustworthy.

Overall, the proposed self-supervised framework lays the groundwork for the next generation of intelligent and explainable medical imaging systems. As the field progresses, integrating self-supervised learning with multimodal data, continual adaptation, and rigorous clinical evaluation will be essential to achieving reliable, scalable, and ethically deployable AI solutions in healthcare.

References

[1] O. Ronneberger, P. Fischer and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234-241, 2015.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, et al., “A Survey on Deep Learning in Medical Image Analysis,” Medical Image Analysis, vol. 42, pp. 60-88, 2017.

[3] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, “Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?,” IEEE Trans. Medical Imaging, vol. 35, no. 5, pp. 1299-1312, 2016.

[4] X. Zhuang, “Self-Supervised Representation Learning for 3D Medical Images by Context Restoration,” IEEE Trans. Medical Imaging, vol. 40, no. 8, pp. 1970-1983, 2021.

[5] K. Chaitanya, N. Karani, C. Baumgartner, A. Anand, O. Donati, and E. Konukoglu, “Contrastive Learning of Global and Local Features for Medical Image Segmentation,” Proc. MICCAI, pp. 466-476, 2020.

[6] A. Taleb, C. Loetzsch, A. Danz, M. Müller, T. Gaertner, C. Brock, and G. Lippert, “3D Self-Supervised Methods for Medical Imaging,” Proc. NeurIPS, 2020.

[7] L. Chen, M. Bentley, P. W. Pham, and S. Grosan, “Generative Self-Supervised Learning for Brain MRI Reconstruction and Lesion Detection,” IEEE J. Biomedical and Health Informatics, vol. 26, no. 3, pp. 1034-1045, 2022.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, and T. Unterthiner, “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” Proc. ICLR, 2021.

[9] J. Chen, Y. Lu, Q. Yu, X. Zhou, and Y. Wang, “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation,” arXiv preprint arXiv:2102.04306, 2021.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” Proc. ICCV, pp. 10012-10022, 2021.

[11] X. Zhang and X. Wang, "Domain-Adaptive Organ Segmentation through SegFormer Architecture in Clinical Imaging", Transactions on Computational and Scientific Methods, vol. 5, no. 7, 2025.

[12] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729 – 9738, 2020.

[13] Y. Zhou, Y. Wang, W. Bai, C. Chen and D. Rueckert, "Abnormality-aware contrastive learning for self-supervised medical image analysis", Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020, Lecture Notes in Computer Science, vol. 12263, pp. 346 – 356, 2020.