ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

Multi-Branch Medical Transformer for SPECT Myocardial Perfusion Imaging: A Novel Approach to Diagnosis

Fuling Zhao[1,a], Xuande Zhang[1,b], Long Xu[2,c], Xin Huang[2,d,*]

[1]School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi, China
[2]Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, Zhejiang, China
[a]fulingzzzzzz@163.com, [b]love_truth@126.com, [c]xulong1@nbu.edu.cn, [d]huangxin@nbu.edu.cn
**Corresponding author*

Abstract: *This study introduces a novel deep learning approach to enhance the accuracy and efficiency of diagnosis in single-photon emission computed tomography myocardial perfusion imaging (SPECT MPI). To address key limitations of current convolutional neural network (CNN)-based methods—such as insufficient information capture, difficulty in removing redundant features, and limited capacity for modeling long-range dependencies—we reconstruct the three-dimensional structure of myocardial perfusion images in a stacked format and propose a multi-branch medical transformer network. This architecture extracts comprehensive features from different anatomical views while integrating critical information, leveraging the Transformer's strength in capturing long-range dependencies to overcome traditional CNN shortcomings. Experimental results demonstrate that the proposed method consistently outperforms conventional CNN-based models across multiple evaluation metrics, achieving improved feature extraction and higher diagnostic accuracy. Comparative experiments and ablation studies further validate the effectiveness of the multi-branch Transformer architecture. The proposed multi-branch vision transformer provides a powerful tool for automated SPECT MPI diagnosis, enhancing diagnostic performance and offering potential support for clinical decision-making.*

Keywords: *Myocardial Perfusion Imaging, Tomography, Emission-Computed, Single-Photon, Vision Transformer, Coronary Artery Disease, Deep Learning*

1. Introduction

According to the 2024 Global Burden of Disease Study, coronary artery disease (CAD) affected about 315 million people worldwide in 2022, remaining the leading cause of death and disability [1,2]. Single-photon emission computed tomography myocardial perfusion imaging (SPECT MPI) is a widely used non-invasive technique for CAD diagnosis, providing three-dimensional assessment of myocardial perfusion under stress and rest conditions [3,4]. However, diagnosis still relies heavily on manual visual interpretation, which is time-consuming and dependent on clinical expertise. Consequently, computer-aided diagnosis (CAD) systems, especially deep learning-based ones, have gained importance [5].

Convolutional neural networks (CNNs) are extensively applied in MPI diagnosis due to their strong feature representation capabilities [6]. Common architectures include ResNet [7], VGG [8], and InceptionNet [9], which analyze two-dimensional MPI images for diagnostic predictions. Nevertheless, CNNs' fixed receptive fields limit their ability to capture global context and long-range dependencies.

Data security concerns, along with the high cost and labor-intensive nature of data acquisition and annotation, have hindered large-scale medical dataset construction [10,11]. This has driven interest in methods requiring less data, with stronger generalization and efficient training. Transfer learning (TL) addresses these challenges by transferring knowledge from a source domain to a target task, reducing dependence on large labeled datasets [12]. Self-supervised learning, an inductive TL approach using unlabeled data for pre-training followed by supervised fine-tuning, has shown strong potential to improve feature representation and generalization under limited labeled data [12,13].

Recent studies highlight TL's effectiveness in medical imaging. Jiao et al. [14] proposed a self-supervised method for fetal ultrasound videos, demonstrating strong transferability to downstream tasks. López et al. [15] used gender recognition as a pre-training task to build a CNN for PLN detection,

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

improving accuracy. Katamutu et al. [16] applied TL to COVID-19 detection, where a pre-trained VGG16 achieved 98% accuracy, surpassing state-of-the-art methods.

Due to scarce publicly available MPI datasets, most methods fine-tune models pre-trained on large datasets like ImageNet [17]. While stable, such approaches have limited improvement potential. MPI reports include short-axis (SA), vertical long-axis (VLA), and horizontal long-axis (HLA) slices under stress and rest (Fig. 1). Directly inputting multi-view data introduces redundancy and noise, raising computational cost and potentially reducing accuracy. Effective joint analysis across orientations is essential but often unachieved by conventional CNNs.

The Transformer architecture, originally for NLP, models long-range dependencies via self-attention [18]. Its computer vision adaptation, Vision Transformer (ViT), partitions images into patches and processes them sequentially to capture global spatial dependencies, achieving performance comparable or superior to CNNs on large datasets [19]. Murphy et al. [20] reported ViT's higher robustness to spurious correlations. Pachetti et al. [21] developed a 3D ViT for prostate cancer classification, achieving 84.6% accuracy versus 78.2% for ResNet3D [22].

In MPI, lesions may occur at multiple myocardial locations, making global contextual understanding crucial. CNNs' fixed receptive fields and ImageNet-pretrained models' mismatch with medical images limit the extraction of essential diagnostic features for reliable MPI classification.

To address the limitations of CNNs and conventional transfer learning, we propose a multi-branch vision transformer architecture for MPI diagnosis. The main contributions are:

(1) To reduce the redundancy present in existing approaches, we extract slices from the image reports and stack them into a three-dimensional format as network input, thereby restoring the volumetric information of the images.

(2) We introduce a network pre-trained on medical three-dimensional CT and two-dimensional X-ray datasets, fine-tuned on MPI data, with multiple successive Vision Transformer blocks designed to progressively extract multi-scale features from the 3D information, enhancing the network's ability to learn features at different scales.

(3) We independently process the data from each anatomical view via three separate branches and fuse their features through average pooling, effectively reducing redundancy among directions and ensuring both classification accuracy and efficiency.

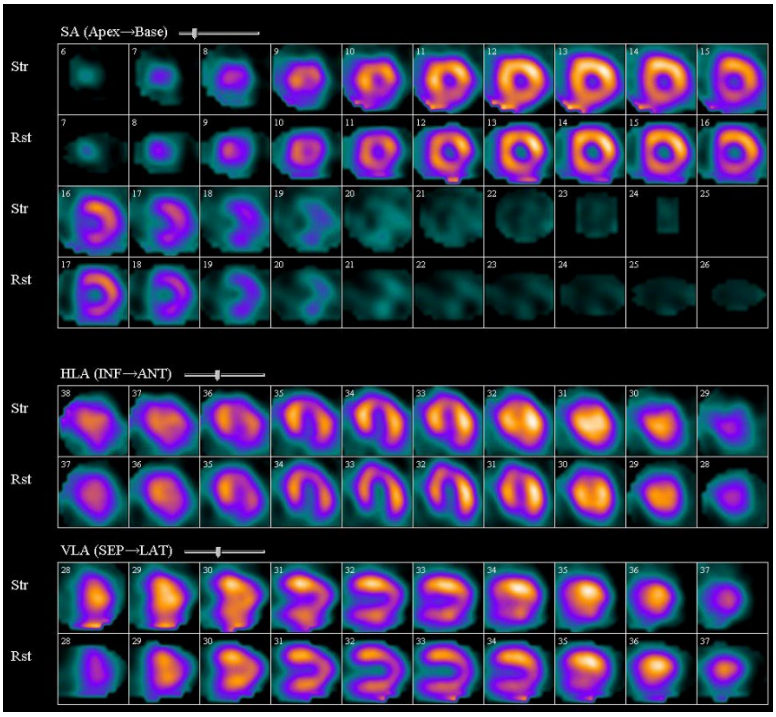

Figure 1: A complete myocardial perfusion imaging (MPI) report, including three imaging planes: short axis (SA), horizontal long axis (HLA), and vertical long axis (VLA), under two physiological states: stress (Str) and rest (Rst).

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

2. Related Work

2.1. Deep Learning Applications in SPECT Imaging

SPECT is a primary imaging modality in nuclear medicine. With rapid advances in artificial intelligence, deep learning has been widely applied to SPECT, mainly focusing on image diagnosis/classification and image quality optimization, including attenuation correction, denoising, and reconstruction.

During SPECT imaging, detectors capture photons emitted by radioactive tracers, but tissue absorption and scattering lead to signal loss and uneven intensity, causing quantitative errors and attenuation artifacts [23]. Attenuation correction (AC) compensates for these losses, improving diagnostic sensitivity and specificity. Traditional AC relies on concurrently acquired CT images, yet ~80% of devices lack this capability, and asynchronous CTs may introduce registration errors. Deep learning offers promising solutions: indirect methods predict attenuation maps (μ-maps) or pseudo-CT images for subsequent reconstruction, whereas direct methods generate AC SPECT images end-to-end [23]. Nguyen et al. [24] combined 3D-GAN and U-Net to synthesize AC images from NAC inputs, achieving optimal SSIM and NMAE metrics. Chen et al. [25] used transfer learning with U-Net and DuRDN to predict μ-maps, with DuRDN achieving a final SPECT image error of 1.11 ± 1.57%. Shanbhag et al. [26] employed a cGAN model to generate AC SPECT images without CT, raising classification accuracy from 54.6% (NAC) to 75%.

Minimizing radiation exposure and patient discomfort necessitates low-dose and fast SPECT scans, which reduce signal-to-noise ratio (SNR) and may compromise diagnostic accuracy. Traditional denoising methods rely on filtering or smoothing, often losing fine structural details. Deep learning models, however, can restore image quality and enhance diagnostic performance by learning complex mappings from low-dose to fully quantitative images [27]. Shiri et al. [28] applied ResNet to restore fully acquired images under shortened acquisition time and reduced angle, showing deep learning effectively mitigates quality loss. Zhenglin Pan et al. [29] proposed a multi-module deep learning framework for accelerated SPECT/CT planar bone imaging (2× and 3× speeds), improving visual quality and contrast agent fidelity.

Song et al. [30] developed a fully quantified 3D Res-CNN reconstruction method, achieving an NMSE of 0.153 and improved left ventricular wall resolution. Ramon et al. [31] evaluated 3D deep learning denoising at various dose levels (1/2–1/16), finding half-dose reconstruction AUC (0.799) closely matched full-dose (0.801). Wu et al. [32] proposed SCI-Net for low-dose reconstruction, leveraging structural features in the projection domain to achieve PSNR improvement from 21.95 to 33.14 and SSIM from 0.9084 to 0.9866, while reducing coefficient of variation in regions of interest.

Overall, these studies demonstrate that deep learning increasingly plays a pivotal role in improving SPECT image quality and clinical utility. Table 1 summarizes the aforementioned deep learning approaches applied to SPECT imaging.

2.2. Classification of SPECT MPI Imaging

In coronary artery disease classification, CNNs are the predominant approach. Many studies utilize state-of-the-art CNNs pre-trained on large datasets such as ImageNet, while others develop customized architectures to better extract relevant features from medical images. Kaplan Berkaya et al. [33] classified SPECT images from 192 patients to detect perfusion abnormalities (ischemia and infarction) using two models: a CNN-SVM hybrid for deep feature classification, and a knowledge-based method combining segmentation, feature extraction, and rule-based algorithms on five predefined image features. The CNN-SVM model achieved 92% accuracy, 84% sensitivity, and 100% specificity, whereas the knowledge-based model achieved 93% accuracy, 100% sensitivity, and 86% specificity.

Vincent Peter C. Magboo et al. [34] applied transfer learning by pre-training on ImageNet, freezing the main network, and fine-tuning on the SPECT-MPI dataset. Comparing backbones, VGG16 and InceptionV3 achieved 84.38% accuracy. In the same year, they proposed a hierarchical sequential neural network with three convolutional layers, a max pooling layer, and a flattening layer [35], achieving the highest accuracy of 93.75%.

Dai Kusumoto et al. [36] introduced a 3D approach by stacking slices from three directions into a ResNet34-based CNN. Features from each network were concatenated and passed through fully connected layers, achieving 88% accuracy—the first 3D classification of SPECT MPI images.

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

Table 1: The Applications of Deep Learning in SPECT Images.

Ref.	Title	Year	Task Type	Input	Output	Method	Main Result
[24]	3D Unet Generative Adversarial Network for Attenuation Correction of SPECT Images	2020	Attenuation Correction	SPECT (NAC)	SPECT(AC)	3D Unet generative adversarial network	SSIM similarity: 0.945% NMAE error: 0.034
[25]	Cross-vender, cross-tracer, and cross-protocol deep transfer learning for attenuation map generation of cardiac SPECT	2022	Attenuation Correction	SPECT (NAC)	μ-map	U-Net + Transfer Learning	μ-map error: 5.13±7.02% reconstructed image error: 1.11±1.57%
[26]	Deep learning-based attenuation correction improves diagnostic accuracy of cardiac SPECT	2023	Attenuation Correction	SPECT (NAC)	SPECT(AC)	conditional GAN	TPD AUC: 0.79 (95% CI: 0.72-0.85)
[28]	Standard SPECT myocardial perfusion estimation from half-time acquisitions using deep convolutional residual neural networks	2021	Reconstruction (fast scan)	Fast scan SPECT	Full-time SPECT	ResNet	RMSE: 6.8±2, ARE: 3.1±1.1, PSNR: 36.0±1.4;
[29]	Fast SPECT/CT planar bone imaging enabled by deep learning enhancement	2024	Reconstruction (fast scan)	Fast scan SPECT	Full-time SPECT	Handcrafted CNN	LPIPS: 0.58 FID: 0.17
[30]	Low-dose cardiac-gated SPECT studies using a residual convolutional neural network	2019	Denoising + Reconstruction (low dose)	Low Dose SPECT	Denoised Full-Dose SPECT images	3D-ResCNN	NMSE: 0.153
[31]	Improving diagnostic accuracy in low-dose SPECT myocardial perfusion imaging with convolutional denoising networks	2020	Denoising + Reconstruction (low dose)	Low-dose SPECT	Denoised Full-Dose SPECT images	Various 3D deep learning models	Best Reconstruction AUC: 0.799
[32]	Sinogram-characteristic-informed network for efficient restoration of low-dose SPECT projection data	2025	Reconstruction (Low Dose)	Low-dose SPECT	Full Dose SPECT	SCI-Net	PSNR: 21.95 → 33.14; SSIM: 0.91 → 0.99;
[33]	Classification models	2020	Classification	SPECT	Classification	Various CNNs	Best Accuracy:

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

	for SPECT myocardial perfusion imaging				result: Normal/Abnormal	+ Transfer Learning & Knowledge-based classification model	0.94
[34]	Diagnosis of coronary artery disease from myocardial perfusion imaging using convolutional neural networks	2023	Classification	SPECT	Classification result: Normal/Abnormal	Various CNNs + Transfer Learning	Best Accuracy: 84.38 Best F1-score: 90.91
[35]	SPECT-MPI for coronary artery disease: a deep learning approach	2024	Classification	SPECT	Classification result: Normal/Abnormal	Handcrafted CNN	Best Accuracy: 93.75
[36]	A deep learning-based automated diagnosis system for SPECT myocardial perfusion imaging	2024	Classification	SPECT	Classification result: Normal/Abnormal	3D ResNet	AUC: 0.91

SSIM, Structural Similarity Index Measure

NMAE, Normalized Mean Absolute Erro

TPD, Total Perfusion Deficit

AUC, Area Under the Receiver Operating Characteristic Curve

CI, Confidence Interval

RMSE, Root Mean Square Error

ARE, Absolute Relative Error

LPIPS, Learned Perceptual Image Patch Similarity

FID, Frechet Inception Distance

NMSE, normalized Mean Squared Error

PSNR, Peak Signal-to-Noise Ratio

3. Methods

3.1. Dataset

The SPECT-MPI dataset [33] comprises 192 patients, as shown in Table 2, who underwent stress/rest Tc-99m myocardial perfusion imaging (MPI) at Eskisehir Osmangazi University between December 2018 and September 2019. Stress images were acquired approximately 30 minutes after the intravenous injection of 10 mCi Tc-99m MIBI following either treadmill exercise or pharmacological stress, while rest images were obtained 30 minutes after the injection of 30 mCi Tc-99m MIBI at rest. Reconstructed slices in the short-axis (SA), horizontal long-axis (HLA), and vertical long-axis (VLA) views were extracted for analysis. Two experienced cardiologists independently reviewed all images and labeled each case as either “normal” or “abnormal.” A perfusion defect was defined as a region exhibiting significantly reduced radiotracer uptake, classified as ischemia if present only in stress images, and as infarction if present in both stress and rest images. This retrospective study was approved by the Ethics Committee of Eskisehir Osmangazi University’s Department of Nuclear Medicine.

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

Table 2: Statistical Summary of the SPECT-MPI Dataset.

Demographic Data	Value
Number of patients	192
Normal (Healthy)	42
Abnormal (Ischemia and/or Infarction)	150
Age range	26~96
Gender (male/female)	73/119

3.2. Multi-Branch Medical Transformer

Our method employs a Medical Transformer (MiT) [37] as the backbone network to extract features from SPECT myocardial perfusion images. After cropping each SPECT image, the short-axis (SA), horizontal long-axis (HLA), and vertical long-axis (VLA) slices are stacked in anatomical order to form 3D volumetric data. These volumes are concatenated along the depth dimension with stress and rest state data to produce fused input volumes as a 4D tensor with dimensions 2D×H×W. The 3D volume data undergoes upsampling and data augmentation preprocessing before being fed into three independent MiT branches, each dedicated to multi-view feature extraction. The branches share the same architecture, each containing four Transformer stages. Through multi-head self-attention, the network captures spatial dependencies, and hierarchical learnable class tokens (CLS Tokens) enable cross-layer global information aggregation. CLS Tokens along with the other sequences from each stage are passed forward to enhance contextual understanding. Finally, global mean pooling is applied to the outputs of each branch to fuse multi-view features, reduce noise, and retain shared information. The concatenated features are then fed into a fully connected layer followed by a Softmax activation for classification. The overall process can be formalized as:

$$y=\text{Softmax}\left(W_c \cdot \frac{1}{3}(f_{SA}+f_{HLA}+f_{VLA})+b_c\right) \quad (1)$$

where f_{SA}, f_{HLA}, f_{VLA}denote the feature outputs of the three branches, $W_c \in \mathbb{R}^{C\times d}$ is the weight matrix of the fully connected layer, C is the number of categories for the classification task, $b_c \in \mathbb{R}^C$is the bias term of the fully connected layer.

By means of multi-perspective collaborative modeling and hierarchical context transmission mechanism, the classification efficiency of abnormal myocardial perfusion has been significantly improved. The overall framework is shown in Figure 2.

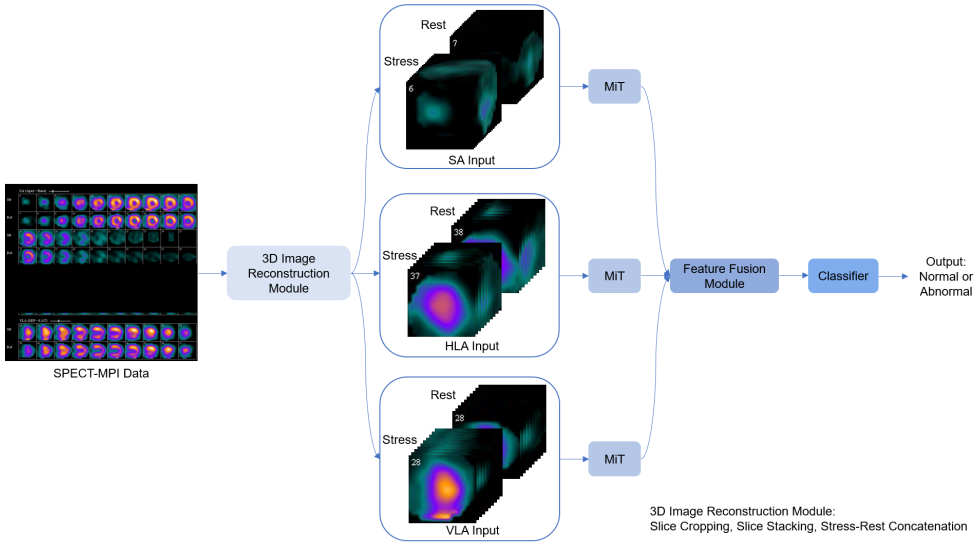

Figure 2: Architecture of the multi-branch Medical Transformer (MiT) model.

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

3.3. UniMiSS

The Universal Medical Self-Supervised (UniMiSS) framework [37] is a versatile self-supervised learning approach for medical imaging. Its core innovation, the Dimension Adaptive Embedding (SPE) module, enables unified modeling of both 2D (e.g., X-ray) and 3D (e.g., CT/MRI) images. UniMiSS adopts a student-teacher paradigm with a pyramid U-shaped Medical Transformer (MiT) [38] as backbone. The switchable SPE module dynamically performs 2D/3D embeddings, allowing the Transformer encoder-decoder to extract cross-dimensional, generalizable features, which are then projected into a contrastive feature space [39,40].

During training, the teacher network parameters are updated via an exponential moving average (EMA) of the student parameters, combined with gradient blocking to prevent model collapse [39]. A dual-granularity consistency constraint is applied: one maximizes semantic agreement between student and teacher outputs through a symmetric cross-entropy loss, and the other aligns 3D volumetric features with 2D slice representations via a body slice consistency loss, enhancing global feature representation. UniMiSS operates without manual annotations and can adaptively handle multimodal medical images, demonstrating robust and cross-dimensional generalization (Figure 3).

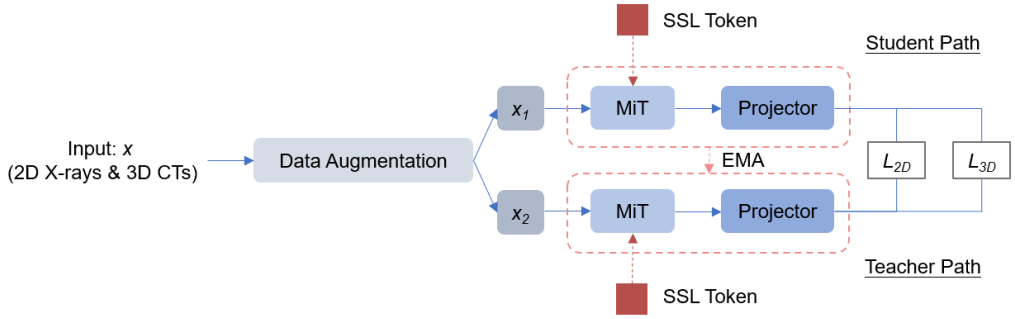

Figure 3: Overall architecture of the UniMiSS framework.

Each branch in the framework consists of a MiT backbone $F_\theta(\cdot)$ and a projector $P_\theta(\cdot)$, where the MiT extracts hierarchical features and the projector maps SSL tokens into a contrastive embedding space. The teacher's parameters μ are updated using EMA:

$$\mu \leftarrow \lambda\mu + (1-\lambda)\mu \tag{2}$$

λ is gradually increased from 0.996 to 1.0 via cosine annealing. Gradient backpropagation to the teacher is blocked to preserve feature diversity.

For 2D data, two augmented views x_1 and x_2 are generated and processed by student and teacher networks. Their outputs $f_1 = P_\theta(F_\theta(x_1; 2D))$ and $f_2 = P_\theta(F_\theta(x_2; 2D))$ are compared using a consistency loss:

$$H(f_1, f_2) = -\text{soft}\max\left(\frac{f_2 - C}{\tau_t}\right) * \log\left(\text{soft}\max\left(\frac{f_1}{\tau_s}\right)\right) \tag{3}$$

where C is the center for the teacher network's output, representing the distribution of different batches. τ_t and τ_s are the temperature parameter. The central update formula effectively prevents the excessive deviation of the teacher network output, maintains stability, and avoids model collapse [39]. The final 2D loss is symmetrized:

$$L^{2D} = E_{x\sim D^{2D}}[H(f_1, f_2) + H(f_2, f_1)] \tag{4}$$

For 3D data, the volume is processed similarly, producing volume-level features f_1, f_2 and slice-level features f_1' and f_2' averaged over all 2D slices. A cross-combination consistency loss is then applied:

$$L^{3D} = E_{x\sim D^{3D}}\begin{bmatrix} H(f_1, f_2) + H(f_1, f_2') + H(f_1', f_2) + H(f_1', f_2') \\ +H(f_2, f_1) + H(f_2, f_1') + H(f_2', f_1) + H(f_2', f_1') \end{bmatrix} \tag{5}$$

This cross-dimensional consistency encourages the model to learn coherent representations across 2D slices and 3D volumes. Slice-level features capture fine-grained local structures, while volume-level

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

alignment preserves global context, resulting in robust and generalizable 3D representations.

3.4. Medical Transformer

MiT is a dimension-independent network architecture that employs an encoder-decoder framework divided into four stages to progressively extract features at multiple scales. The overall structure is shown in Figure 4. Each stage consists of a Switchable Patch Embedding (SPE) module and several Transformer layers. The SPE module automatically selects the appropriate convolution strategy based on the input image's dimensionality (2D or 3D), converting the raw image into a token sequence. This adaptive module employs a learnable convolutional structure to effectively process medical images of different dimensions—particularly well-suited for modeling the continuity inherent in 3D data. It facilitates deeper exploration of spatial contextual information and cross-slice correlations within volumetric medical data. In the encoder, multi-scale feature representations are extracted through progressive downsampling. The decoder symmetrically upsamples the features and integrates the corresponding encoder stage features through skip connections (Jump Connections). This mechanism helps preserve local detail and global semantic information during upsampling, thereby improving decoding quality and detail restoration. To achieve self-supervised learning (SSL), UniMiSS introduces a learnable SSL token during the patch embedding stage [39,40]. New SSL tokens are dynamically generated and appended to the token sequence at each stage. These tokens interact with other visual tokens through the attention mechanism, effectively capturing long-range dependencies and enhancing semantic representation.

To alleviate the computational and memory burden posed by high-resolution images in the Transformer, MiT incorporates a Spatial Reduction Attention (SRA) mechanism [38]. This mechanism applies spatial downsampling to queries q, keys k, and values v before feeding them into the Multi-Head Self-Attention (MSA) module.

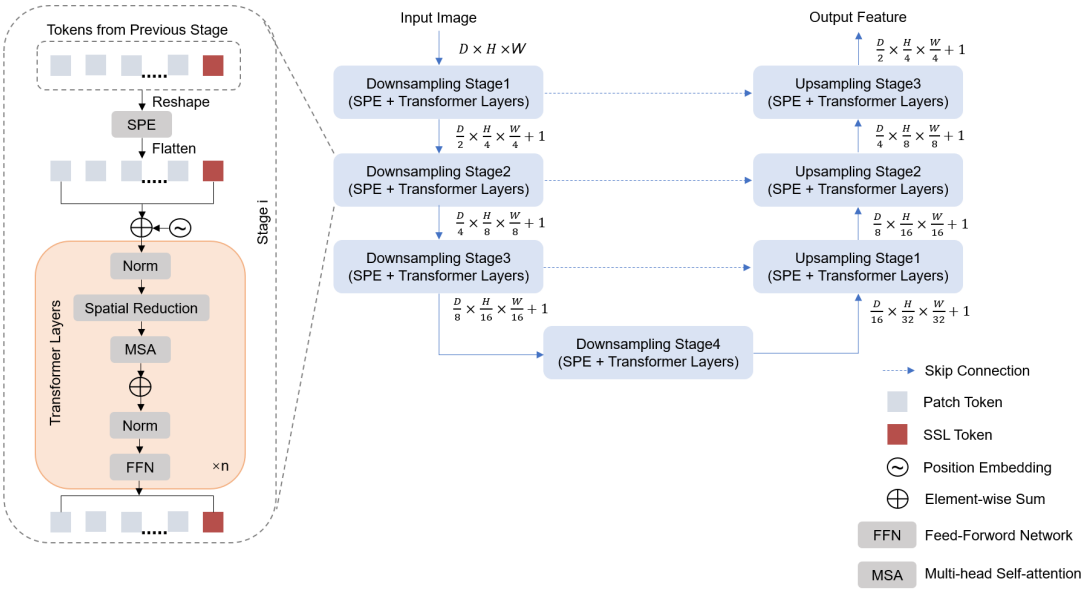

Figure 4: Overview of the MiT architecture.

3.5. Experiment Settings

For the downstream SPECT-MPI classification task, the MiT encoder pre-trained in the student pathway of UniMiSS is used as the feature extraction backbone. It processes slices from three anatomical views—horizontal long axis (HLA), vertical long axis (VLA), and short axis (SA)—with features from all views fused and passed through a fully connected layer for classification, where the output dimension matches the number of diagnostic categories.

During fine-tuning, the stacked 3D input volume is resized to 32×96×96. To address class imbalance, minority class samples are upsampled, and various data augmentation strategies are applied, including spatial transformations, noise addition, resolution degradation, mirror flipping, and color enhancement, to improve generalization.

The dataset is split 8:2 into training and validation sets. Training uses the AdamW optimizer [41] with

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

an initial learning rate of 1e-5, batch size of 16, and 200 epochs, performed on an NVIDIA RTX 3090 GPU. Model performance is assessed via ROC curves, confusion matrices, and the area under the ROC curve (AUC).

4. Results

4.1. Comparison with Existing Methods

The proposed method was compared against several representative deep learning models, including VGG16 [8], ResNet50 [7], DenseNet121 [42], and InceptionV3 [43]. All baseline models were pre-trained on ImageNet and fine-tuned on the SPECT-MPI dataset. As shown in Table 3 and Figure 5, the proposed approach achieved an AUC of 0.9600 and an F1-score of 0.9434, outperforming all baselines. The ROC curve of our model was positioned closer to the top-left corner, with a smooth and steep ascent at low false positive rates, demonstrating superior discriminative capability. Notably, the confusion matrix revealed that our method produced zero false negatives, indicating complete identification of abnormal patients, whereas baseline models exhibited varying levels of missed diagnoses.

Table 3: Classifications Performance of Different Methods.

Method	Dataset	AUC	F1-score
VGG16	SPECT-MPI	0.9200	0.9020
ResNet50		0.8457	0.8800
DenseNet121		0.9486	0.9388
InceptionV3		0.8686	0.8750
Ours		**0.9600**	**0.9434**

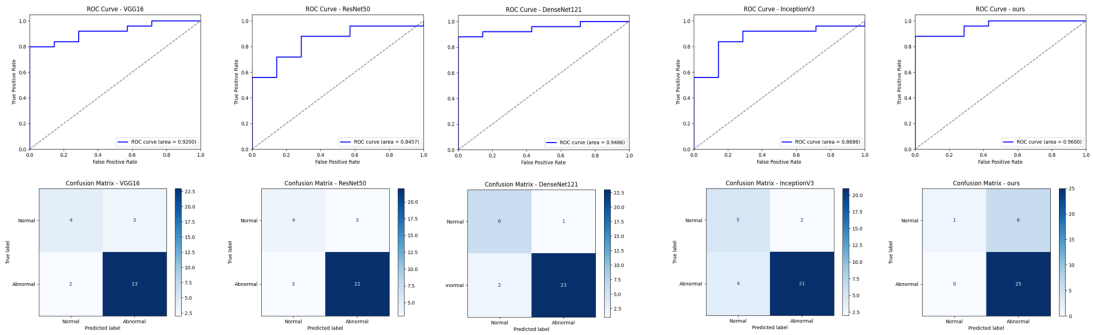

Figure 5: Comparison of classification performance across different methods (ROC Curves & confusion matrices).

Models from left to right: VGG16, ResNet50, DenseNet121, InceptionV3, and the proposed method.

4.2. Comparison of Single-Axis Models and Fusion Model

Table 4 and Figure 6 summarize the diagnostic performance of single-axis and multi-axis fusion models. Among the single-axis models, the SA axis achieved the highest performance with an AUC of 0.9486 and an F1-score of 0.9434, but it still resulted in six false negatives. The HLA- and VLA-based models performed slightly worse individually but captured complementary structural cues. When integrating all three axes through fusion, the model achieved the highest AUC of 0.9600 with no false negatives, highlighting superior recall and diagnostic safety compared to single-axis approaches.

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

Table 4: Classification Performance of Single-Axis Models and Fusion Model.

Method	AUC	F1-score
HLA-Only	0.9371	0.9231
VLA-Only	0.9314	0.9259
SA-Only	0.9486	0.9434
Fusion Model (HLA+VLA+SA)	**0.9600**	**0.9434**

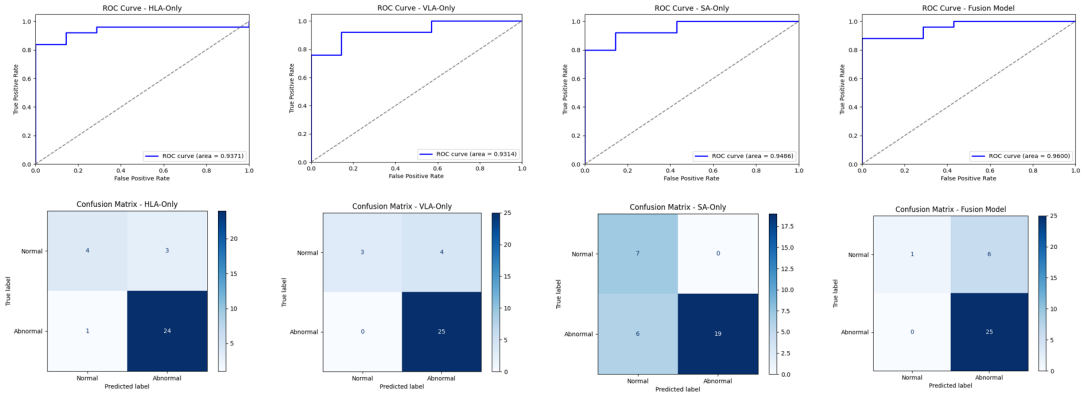

Figure 6: Comparison of classification performance between single-axis models and fusion model (ROC curves and confusion matrices).

From left to right: HLA-only, VLA-only, SA-only, and fusion model (proposed).

4.3. Comparison of Different Fusion Strategies

Three fusion strategies were compared: learnable weights, feature concatenation, and average pooling. Results in Table 5 and Figure 7 show that average pooling achieved the best overall performance (AUC = 0.9600, F1-score = 0.9434), with the fewest false positives and negatives. The concatenation strategy resulted in performance degradation (AUC = 0.9200) due to redundancy, while the learnable-weight approach showed the lowest AUC (0.8857) and signs of overfitting.

Table 5: Classification Performance of Different Fusion Methods.

Method	AUC	F1-score
Learnable weights per axis	0.8857	0.9259
Direct concatenation	0.9200	0.9259
Average pooling (proposed method)	**0.9600**	**0.9434**

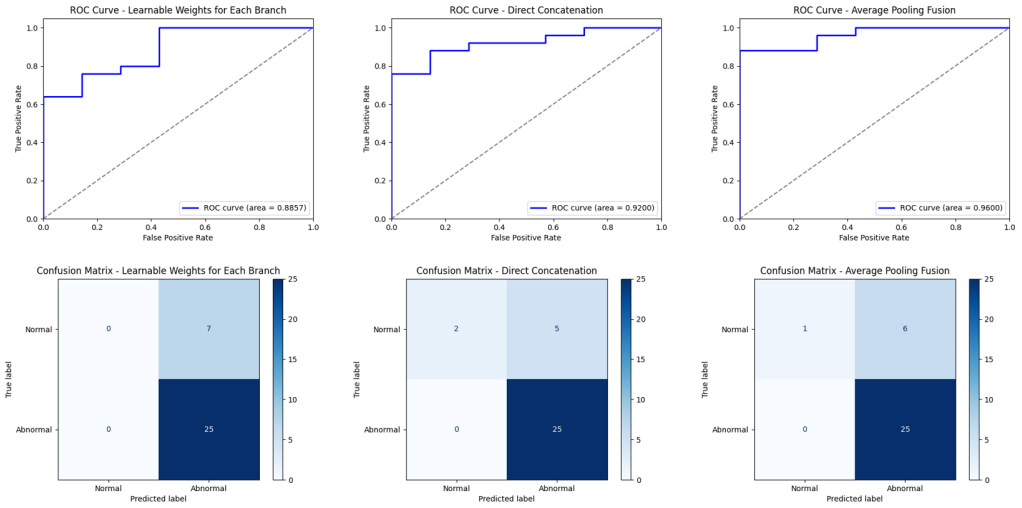

Figure 7: Comparison of different fusion strategies for classification performance (ROC curves and confusion matrices).

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

From left to right: learnable weights per axis, direct concatenation, and average pooling (proposed).

4.4. Comparison of Model Performance with and without Fine-tuning

Table 6 demonstrates the effect of fine-tuning on UniMiSS pre-trained models. Fine-tuned models significantly outperformed those without fine-tuning across all settings. For single-axis models, AUC values increased to 0.9371 (HLA), 0.9314 (VLA), and 0.9486 (SA), compared to much lower values without fine-tuning. For the fusion model, fine-tuning achieved an AUC of 0.9600 versus 0.8571 without fine-tuning. These results confirm that fine-tuning is critical to fully exploiting the representational power of the pre-trained backbone.

Table 6: Classification Performance of Fine-tuned and Non-fine-tuned Models.

Fine-tuning	Method	AUC	F1-score
Fine-tuned	HLA-Only	0.9371	0.9231
	VLA-Only	0.9314	0.9259
	SA-Only	0.9486	0.9434
	Fusion Model (HLA+VLA+SA)	**0.9600**	**0.9434**
Non-fine-tuned	HLA-Only	0.7600	0.8889
	VLA-Only	0.8229	0.8929
	SA-Only	0.6857	0.8929
	Fusion Model (HLA+VLA+SA)	0.8571	0.8889

5. Discussion

These results highlight several important findings. Firstly, compared with the traditional benchmark model based on convolutional neural networks, the proposed framework based on UniMiSS demonstrates superior diagnostic performance on the SPECT-MPI dataset. Thanks to the excellent long-distance relationship capturing ability of the Transformer architecture, the model can learn the features of different parts of the myocardium and their correlations, no longer learning the correlations between adjacent convolutional windows like traditional two-dimensional convolutional models, thus solving the limitation of fixed receptive fields in convolutional neural networks. Compared with the two-dimensional model, the three-dimensional model can learn more abundant myocardial information, providing strong assistance for diagnosis. The results of the ROC curve also show that the model is smooth and steep in the low false positive rate area, demonstrating higher robustness and clinical applicability. Moreover, the absence of false negatives in the model further highlights the advantages of this model in clinical safety and stability, which is particularly important for the diagnosis of coronary artery diseases, as missed diagnoses may lead to serious consequences.

Secondly, the analysis of the single-axis model and the fusion model emphasizes the necessity of integrating information from multiple anatomical views. The SA-Only model achieved the best results among all single-axis models, thanks to the more abundant and complete myocardial information provided by short-axis section slices. However, the results of the confusion matrix show that the SA-Only model has a higher false negative rate (FN = 6), indicating that it still needs additional perspectives to provide the missing key information. While the HLA-Only model and the VLA-Only model performed poorly, they can still provide complementary myocardial information from different perspectives, enabling the final fusion model to achieve the best effect.

Thirdly, the comparison of fusion strategies shows that simplicity and balance are the advantages. The performance of the average pooling strategy is superior to more complex strategies, such as learnable

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

weights and concatenation strategies. Average pooling can reduce the information redundancy caused by concatenation while preventing overfitting. This indicates that in the case of limited data, simple aggregation methods can produce more reliable results, which is a common challenge in medical imaging.

Finally, the fine-tuning experiment confirms the necessity of adapting the pre-trained model to the specific features of the SPECT-MPI dataset. The model without fine-tuning training showed a significant decline in performance, highlighting the importance of task-specific adaptation. This finding is consistent with previous research, that is, the pre-trained representations need to be optimized for specific domains to achieve the best diagnostic performance.

6. Conclusions

In summary, the proposed framework utilizes UniMiSS pre-training, multi-axis fusion, and average pooling to achieve robust and clinically reliable classification of myocardial perfusion images. The concept in the pre-training strategy of the UniMiSS framework enables the model to better understand the consistency between slices in the three-dimensional data. The multi-axis fusion strategy fully utilizes the key information from each axial slice, while average pooling demonstrates strong information integration ability. The model's strong generalization ability and safety indicate that it has promising prospects in the practical application of computer-aided coronary artery disease diagnosis.

References

[1] Stark B, Johnson C, Roth G. Global prevalence of coronary artery disease: an update from the Global Burden of Disease Study. J Am Coll Cardiol 2024;83(13 Suppl):2320. doi:10.1016/S0735-1097(24)04310-9.

[2] Liu X, Wu Y, Li F, Qi X, Niu L, Wu Y, et al. Global burden of early-onset ischemic heart disease, 1990 to 2019. JACC Adv 2025;4(1):101466. doi:10.1016/j.jacadv.2024.101466.

[3] Notghi A, Low CS. Myocardial perfusion scintigraphy: past, present and future. Br J Radiol 2011;84(Spec Iss 3):S229–S236. doi:10.1259/bjr/14625142.

[4] International Atomic Energy Agency. Nuclear cardiology: guidance on the implementation of SPECT myocardial perfusion imaging. IAEA Human Health Series No. 23 (Rev. 1). Vienna: IAEA; 2016.

[5] Zhang H, Qie Y. Applying deep learning to medical imaging: a review. Appl Sci 2023; 13:10521. doi:10.3390/app131810521.

[6] Yu H, Yang LT, Zhang Q, Armstrong D, Deen MJ. Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. Neurocomputing 2021; 444:92–110. doi:10.1016/j.neucom.2020.04.157.

[7] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc IEEE Conf Comput Vis Pattern Recognit; 2016. p. 770–778. doi:10.1109/CVPR.2016.90.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[9] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Proc IEEE Conf Comput Vis Pattern Recognit 2015;1–9. doi:10.1109/CVPR.2015.7298594

[10] Zhao Z, Alzubaidi L, Zhang J, Duan Y, Gu Y. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. Expert Syst Appl 2024; 242:122807. doi:10.1016/j.eswa.2023.122807

[11] Ellis RJ, Sander RM, Limon A. Twelve key challenges in medical machine learning and solutions. Intell Based Med 2022; 6:100068. doi:10.1016/j.ibmed.2022.100068

[12] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng 2010; 22:1345–1359. doi:10.1109/TKDE.2009.191

[13] Amini MR, Feofanov V, Pauletto L, Hadjadj L, Devijver É, Maximov Y. Self-training: a survey. Neurocomputing 2025; 616:128904. doi:10.1016/j.neucom.2024.128904

[14] Jiao J, Droste R, Drukker L, Papageorghiou AT, Noble JA. Self-supervised representation learning for ultrasound video. IEEE Int Symp Biomed Imaging 2020:1847–1850. doi:10.1109/ISBI45749.2020.9098666

[15] Lopes RR, Bleijendaal H, Ramos LA, Verstraelen TE, Amin AS, Wilde AA, et al. Improving electrocardiogram-based detection of rare genetic heart disease using transfer learning: An application to phospholamban p. Arg14del mutation carriers. Comput Biol Med 2021; 131:104262. doi:10.1016/j.compbiomed.2021.104262

[16] Kathamuthu ND, Subramaniam S, Le QH, Muthusamy S, Panchal H, Sundararajan SCM, et al. A deep transfer learning-based convolution neural network model for COVID-19 detection using computed tomography scan images for medical applications. Adv Eng Softw 2023; 175:103317.

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

doi:10.1016/j.advengsoft.2022.103317
[17] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. Proc IEEE Conf Comput Vis Pattern Recognit 2009; 248–255. doi:10.1109/CVPR.2009.5206848
[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[19] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[20] Murphy ZR, Venkatesh K, Sulam J, Yi PH. Visual transformers and convolutional neural networks for disease classification on radiographs: a comparison of performance, sample efficiency, and hidden stratification. Radiol Artif Intell 2022; 4(6):e220012. doi:10.1148/ryai.220012.
[21] Pachetti E, Colantonio S, Pascali MA. On the effectiveness of 3D vision transformers for the prediction of prostate cancer aggressiveness. In: Proc Int Conf Image Anal Process; 2022. p. 317–328. doi:10.1007/978-3-031-13324-4_27.
[22] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proc IEEE Conf Comput Vis Pattern Recognit; 2018. p. 6546–6555. doi:10.1109/CVPR.2018.00684
[23] Chen X, Liu C. Deep-learning-based methods of attenuation correction for SPECT and PET. J Nucl Cardiol 2023;30(5):1859-1878.
[24] Nguyen TT, Chi TN, Hoang MD, Thai HN, Duc TN. 3D Unet generative adversarial network for attenuation correction of SPECT images. In: 2020 4th international conference on recent advances in signal processing, telecommunications & computing (SigTelCom). Hanoi, Vietnam: IEEE, 2020:93–97.
[25] Chen X, Hendrik Pretorius P, Zhou B, Liu H, Johnson K, Liu YH, King MA, Liu C. Cross-vender, cross-tracer, and cross-protocol deep transfer learning for attenuation map generation of cardiac SPECT. J Nucl Cardiol 2022;29(6):3379-3391.
[26] Shanbhag AD, Miller RJH, Pieszko K, Lemley M, Kavanagh P, Feher A, et al. Deep learning-based attenuation correction improves diagnostic accuracy of cardiac SPECT. J Nucl Med 2023;64(3):472-478.
[27] Apostolopoulos ID, Papandrianos NI, Feleki A, Moustakidis S, Papageorgiou EI. Deep learning-enhanced nuclear medicine SPECT imaging applied to cardiac studies. EJNMMI Phys 2023;10(1):6.
[28] Shiri I, AmirMozafari Sabet K, Arabi H, Pourkeshavarz M, Teimourian B, Ay MR, et al. Standard SPECT myocardial perfusion estimation from half-time acquisitions using deep convolutional residual neural networks. J Nucl Cardiol 2021;28(6):2761-2779.
[29] Pan Z, Qi N, Meng Q, Pan B, Feng T, Zhao J, Gong NJ. Fast SPECT/CT planar bone imaging enabled by deep learning enhancement. Med Phys 2024;51(8):5414–5426.
[30] Song C, Yang Y, Wernick MN, Pretorius PH, King MA. Low-dose cardiac-gated SPECT studies using a residual convolutional neural network. In: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). Venice, Italy: IEEE, 2019:653–656.
[31] Ramon AJ, Yang Y, Pretorius PH, Johnson KL, King MA, Wernick MN. Improving diagnostic accuracy in low-dose SPECT myocardial perfusion imaging with convolutional denoising networks. IEEE Trans Med Imaging 2020; 39:2893–2903.
[32] Wu R, Liu H, Lai P, Yuan W, Li H, Jiang Y. Sinogram-characteristic-informed network for efficient restoration of low-dose SPECT projection data. Med Phys 2025;52(1):414–432.
[33] Berkaya SK, Sivrikoz IA, Gunal S. Classification models for SPECT myocardial perfusion imaging. Comput Biol Med 2020; 123:103893. doi:10.1016/j.compbiomed.2020.103893
[34] Magboo VPC, Magboo MSA. Diagnosis of coronary artery disease from myocardial perfusion imaging using convolutional neural networks. Procedia Comput Sci 2023; 218:810-817. doi:10.1016/j.procs.2023.01.061
[35] Magboo VPC, Magboo MSA. SPECT-MPI for coronary artery disease: a deep learning approach. Acta Med Philipp 2024 May 15;58(8):67-75. doi: 10.47895/amp.vi0.7582. PMID: 38812768; PMCID: PMC11132284.
[36] Kusumoto D, Akiyama T, Hashimoto M, et al. A deep learning-based automated diagnosis system for SPECT myocardial perfusion imaging. Sci Rep 2024; 14:13583. doi: 10.1038/s41598-024-64445-2.
[37] Xie Y, Zhang J, Xia Y, Wu Q. UniMiSS: Universal medical self-supervised learning via breaking dimensionality barrier. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, eds. Computer Vision – ECCV 2022. Lecture Notes in Computer Science, vol 13681. Cham: Springer; 2022. p. 555–572. doi:10.1007/978-3-031-19803-8_33.
[38] Wang W, Xie E, Li X, et al. Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In: Proc IEEE/CVF Int Conf Comput Vis; 2021. p. 568–578. doi:10.1109/ICCV48922.2021.00062.
[39] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers. In: Proc IEEE/CVF Int Conf Comput Vis; 2021. p. 9650–9660. doi:10.1109/ICCV48922.2021.00951.

ISSN 2616-5775 Vol. 8, Issue 11: 38-51, DOI: 10.25236/AJCIS.2025.081105

[40] Chen X, Xie S, He K. An empirical study of training self-supervised vision transformers. In: Proc IEEE/CVF Int Conf Comput Vis; 2021. p. 9640–9649. doi:10.1109/ICCV48922.2021.00950.
[41] Loschilov I, Hutter F. Fixing weight decay regularization in Adam. In: Proc Int Conf Learn Representations; 2018.
[42] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc IEEE Conf Comput Vis Pattern Recognit; 2017. p. 2261–2269. doi:10.1109/CVPR.2017.243.
[43] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception architecture for computer vision. Proc IEEE Conf Comput Vis Pattern Recognit 2016; 2818–2826.