Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061977

ARTICLE

Multi-Scale Vision Transformer with Dynamic Multi-Loss Function for Medical Image Retrieval and Classification

Omar Alqahtani, Mohamed Ghouse*, Asfia Sabahath, Omer Bin Hussain and Arshiya Begum

Department of Computer Science, College of Computer Science, King Khalid University, Abha, 61421, Saudi Arabia

*Corresponding Author: Mohamed Ghouse. Email: mghoth@kku.edu.sa

Received: 07 December 2024; Accepted: 04 March 2025; Published: 16 April 2025

ABSTRACT: This paper introduces a novel method for medical image retrieval and classification by integrating a multi-scale encoding mechanism with Vision Transformer (ViT) architectures and a dynamic multi-loss function. The multi-scale encoding significantly enhances the model's ability to capture both fine-grained and global features, while the dynamic loss function adapts during training to optimize classification accuracy and retrieval performance. Our approach was evaluated on the ISIC-2018 and ChestX-ray14 datasets, yielding notable improvements. Specifically, on the ISIC-2018 dataset, our method achieves an F1-Score improvement of +4.84% compared to the standard ViT, with a precision increase of +5.46% for melanoma (MEL). On the ChestX-ray14 dataset, the method delivers an F1-Score improvement of 5.3% over the conventional ViT, with precision gains of +5.0% for pneumonia (PNEU) and +5.4% for fibrosis (FIB). Experimental results demonstrate that our approach outperforms traditional CNN-based models and existing ViT variants, particularly in retrieving relevant medical cases and enhancing diagnostic accuracy. These findings highlight the potential of the proposed method for large-scale medical image analysis, offering improved tools for clinical decision-making through superior classification and case comparison.

KEYWORDS: Medical image retrieval; vision transformer; multi-scale encoding; multi-loss function; ISIC-2018; ChestX-ray14

1 Introduction

Medical image retrieval has become a vital tool in clinical diagnostics, enabling physicians to compare patient data with historical cases to improve diagnostic accuracy [1,2]. Traditional image retrieval techniques, particularly those based on Convolutional Neural Networks (CNNs), have demonstrated effectiveness by capturing local features and hierarchical spatial patterns [3–5]. However, these methods often struggle to capture the global context of images, which is crucial for medical imaging tasks that require a comprehensive analysis of both local and global structures [6,7].

Vision Transformers (ViTs) [6] have gained significant attention due to their self-attention mechanisms, which enhance context modeling and improve performance in tasks requiring a holistic understanding of data [7,8]. Recent studies have demonstrated the effectiveness of ViTs in medical imaging, particularly in handling complex and heterogeneous datasets [7,9]. Despite these advancements, challenges persist in efficiently integrating multi-scale information and optimizing models for small-object representation [10].

A fundamental challenge in medical image retrieval lies in managing the complexity and diversity of medical datasets, which often involve multi-label classifications [11]. Datasets such as ISIC-2018 [12,13] and ChestX-ray14 [1] pose unique difficulties due to the wide range of disease categories and the need for models

to generalize across varying conditions. Fig. 1 presents sample images from these datasets, highlighting their complexity and diversity. Traditional CNN-based approaches often struggle to effectively capture both global and local features, particularly in multi-label classification settings [11,14].

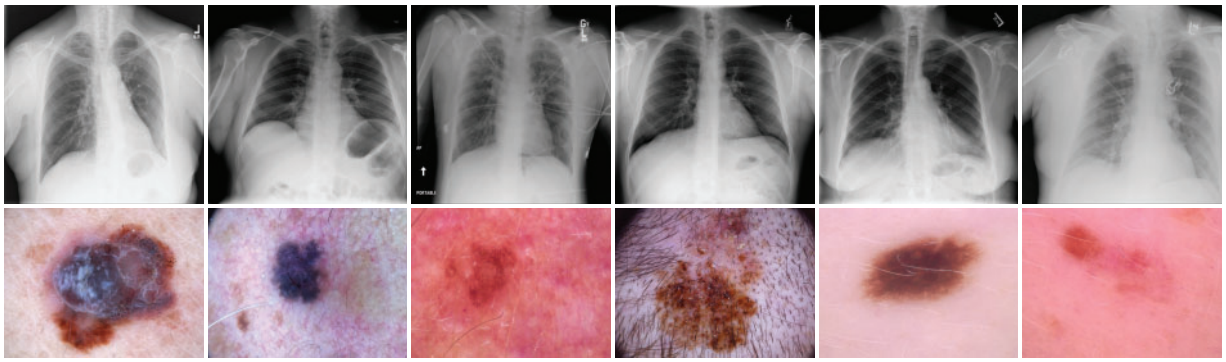


Figure 1: Sample images from the ISIC-2018 [12] and ChestX-ray14 [1] datasets. These examples illustrate the diversity of medical conditions, including skin lesions (melanoma, basal cell carcinoma) and thoracic pathologies (pneumonia, cardiomegaly), emphasizing the need for robust models capable of accurate image classification and retrieval

To address these limitations, recent research has focused on enhancing network architectures and loss functions. Multi-scale context-aware attention models [15] and improved U-Net variants [16,17] have been proposed to enhance feature extraction in medical images. Furthermore, efficient segmentation techniques integrating knowledge distillation [18] and lightweight models [19] have demonstrated promising results in skin lesion analysis.

In this paper, we present a novel approach for medical image classification and retrieval that integrates a multi-scale encoding mechanism [20] with a Vision Transformer (ViT) architecture [6]. Our method is evaluated on the ISIC-2018 [12,13] and ChestX-ray14 [1] datasets. The multi-scale encoding enhances the model's ability to capture both fine-grained and global features, leading to improved classification and retrieval performance. Additionally, we introduce a dynamic multi-loss function that adapts to different training stages, optimizing classification accuracy, feature space structuring for retrieval, and overall model robustness. By adjusting the contributions of various loss components, including cross-entropy loss [21], triplet loss [22], contrastive loss [23], and distillation loss [24], our approach achieves a balanced learning process, resulting in enhanced performance across both datasets.

Our method is tested on two widely used medical image datasets: ISIC-2018, which focuses on skin lesion classification [12,13], and ChestX-ray14, which involves the identification of various thoracic diseases [1]. The ISIC-2018 dataset presents challenging lesion categories such as melanoma, basal cell carcinoma, and benign keratosis [12], while the ChestX-ray14 dataset contains diverse pathologies, including atelectasis, edema, and pneumonia [1]. Our approach achieves higher precision, recall, and F1-scores across both datasets, significantly outperforming existing methods [7,8].

Experimental results demonstrate that the proposed approach not only achieves high classification accuracy but also excels in medical image retrieval tasks, which are critical for case-based reasoning in clinical applications. The integration of multi-scale encoding with the ViT framework [6], combined with the dynamic multi-loss function, results in a well-structured feature space that enhances both diagnostic accuracy and retrieval performance. As shown in 'Results and Analysis' Section 5, our approach surpasses traditional CNN-based models and existing ViT variants [7,8], demonstrating its potential for large-scale medical image analysis.

2 Related Work

Convolutional Neural Networks (CNNs) have long served as the backbone for image recognition and retrieval tasks, largely due to their hierarchical architectures that capture both low-level and high-level features effectively [3–5]. Pioneering architectures such as VGGNet [3], GoogLeNet [4], and ResNet [5] have substantially advanced the field by introducing innovative techniques to enhance performance and training efficiency. VGGNet underscored the significance of network depth in feature extraction, GoogLeNet optimized computational efficiency via Inception modules, and ResNet mitigated the degradation problem in deep networks through residual learning, enabling the effective training of networks with over 100 layers. Despite these advancements, CNNs inherently struggle to capture long-range dependencies and global contextual information because of their localized receptive fields.

Lightweight CNN models such as DenseNet [25], MobileNet [26], SqueezeNet [27], and Fast&Focused-Net [10] have made notable contributions to image retrieval and classification, particularly in resource-constrained settings. DenseNet improves feature reuse through dense connectivity, MobileNet lowers computational cost with depthwise separable convolutions, SqueezeNet achieves high accuracy with fewer parameters via Fire modules, and Fast&Focused-Net enhances small object encoding through a Volume-wise Dot Product (VDP) layer. Nonetheless, these traditional CNN approaches often fall short in generalizing to complex datasets characterized by high intra-class variance and inter-class similarities.

More recently, Vision Transformers (ViTs) have emerged as a powerful alternative by employing self-attention mechanisms to capture long-range dependencies and global features more effectively [6]. Unlike CNNs, ViTs treat images as sequences of patches and use attention to focus on the most relevant regions, a strategy that is particularly beneficial in medical image analysis where both local details and global context are critical [7]. However, standard ViTs typically require large-scale training datasets due to their limited inductive biases, which can lead to overfitting when applied to smaller medical datasets.

In medical imaging, ViTs have been increasingly adopted across various applications. For instance, they have been applied to tumor classification in MRI scans, where they effectively differentiate between tumor types. In segmentation tasks, models such as TransUNet have improved the delineation of anatomical structures [28,29]. Additionally, ViTs have demonstrated utility in MRI reconstruction and automated report generation in telehealth systems [28]. Nonetheless, a key limitation of ViTs in this domain is their reduced capacity to integrate multi-scale feature representations, which are crucial for segmenting small lesions and detecting subtle abnormalities.

For multi-label datasets such as ISIC-2018 [12,13] and ChestX-ray14 [1], current ViT-based approaches show promising results yet continue to face challenges including difficulties in capturing fine-grained details, high computational demands, and reliance on large annotated datasets. Some studies have attempted to overcome these issues by integrating CNN-like hierarchical feature extraction into ViTs [7] or by incorporating attention mechanisms into CNNs [14], but a comprehensive multi-scale approach remains underexplored.

Loss functions play a critical role in optimizing deep learning models for classification and retrieval tasks. Loss functions from deep metric learning, such as contrastive loss [23], triplet loss [22], and cross-entropy loss [21], enhance retrieval performance by structuring the feature space. Additionally, knowledge distillation [24] has been used to transfer insights from larger models to smaller ones, thereby improving efficiency [18]. Recently, dynamic multi-loss functions that adaptively adjust weight contributions during training have shown promise in balancing classification accuracy with retrieval performance. However, existing multi-loss strategies have yet to fully exploit training stage-dependent weight adjustments, which could further enhance model performance [7].

In medical image segmentation, architectures such as U-Net [30] and its variants, including H-DenseUNet [31], V-Net [32], and U-Net++ [30], have been widely employed for lesion segmentation. Efforts to improve these models have focused on integrating attention mechanisms [33] and using hybrid preprocessing techniques [34]. Despite these enhancements, challenges persist in efficiently capturing multi-scale contextual information. Similarly, disease classification models like CheXNet [35] have achieved performance comparable to that of radiologists in pneumonia detection, yet their dependence on fixed-scale feature extraction limits their adaptability to diverse datasets.

Our proposed approach addresses these limitations by integrating multi-scale encoding within a ViT framework and employing a dynamic multi-loss function. Unlike traditional CNNs, which struggle with long-range dependencies, or standard ViTs, which lack hierarchical feature extraction, our model explicitly incorporates multi-scale representations. Furthermore, by dynamically adjusting loss function weights during different training stages, our method optimizes both classification and retrieval performance, making it particularly effective for complex medical image datasets.

3 Proposed Method

In this work, we propose a novel approach that integrates a multi-scale Vision Transformer (ViT) architecture [6] with a dynamic multi-loss function to address the challenges inherent in medical image retrieval and classification tasks. The method leverages multi-scale image encoding to effectively capture both fine-grained and global features, while the dynamic loss scheme balances multiple learning objectives throughout the training process, enhancing both classification accuracy and retrieval performance.

3.1 Overview

Our approach begins by generating scaled versions of each input image by progressively downscaling it until the smallest side is greater than 16 pixels, similar to multi-scale techniques used in medical image analysis [15]. At each scale, the image is partitioned into patches of size 16×16 pixels, which are then flattened into one-dimensional vectors, following the methodology of ViT [6]. We augment each patch vector with a three-dimensional positional encoding that represents the scale level and the patch's spatial coordinates (X and Y), enabling the model to capture spatial relationships across different scales. This results in a sequence of enriched vectors for each image. A special classification token (`[CLS]`) is prepended to this sequence to aggregate global information, as demonstrated in transformer architectures [6].

This sequence is fed into a stack of multi-head self-attention encoder layers within the ViT architecture [6]. Each encoder layer consists of multi-head attention mechanisms and feed-forward networks that enable the model to learn long-range dependencies and global context. The output corresponding to the `[CLS]` token is extracted after the encoding layers and processed through a feed-forward layer to produce an embedding vector. This embedding is utilized for image retrieval tasks and serves as input for computing loss components like triplet loss [22] and contrastive loss [23].

An overview of the proposed architecture is depicted in Fig. 2.

For classification tasks, the embedding vector is passed through an additional feed-forward layer followed by a softmax activation function to generate class probabilities. The cross-entropy loss [21] is computed using these probabilities and the ground truth labels.

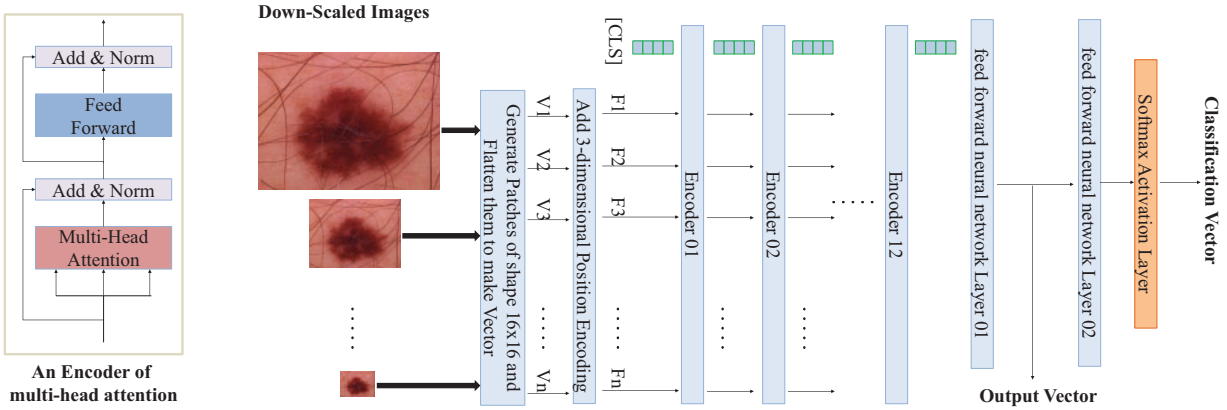


Figure 2: Overview of the proposed multi-scale Vision Transformer architecture for medical image classification and retrieval tasks. The input image is downscaled multiple times, divided into patches, and processed through a series of multi-head attention layers with positional encoding added. The final output vector is used for image retrieval and classification tasks, leveraging a dynamic multi-loss function

3.2 Dynamic Multi-Loss Function

To balance the model's performance on both classification and retrieval tasks, we employ a combined loss function that integrates cross-entropy loss (L_{CEL}), contrastive loss ($L_{\text{Contrastive}}$) [23], and triplet loss (L_{Triplet}) [22]:

$$\text{Loss} = W_{\text{CEL}} \times L_{\text{CEL}} + W_{\text{Contrastive}} \times L_{\text{Contrastive}} + W_{\text{Triplet}} \times L_{\text{Triplet}} \tag{1}$$

Here, W_{CEL}, $W_{\text{Contrastive}}$, and W_{Triplet} are weights assigned to each loss component. These weights are dynamically adjusted during different training phases to optimize the model's learning objectives. Initially, the model emphasizes learning discriminative features through higher weights on the contrastive and triplet losses, which structure the embedding space for improved retrieval [7]. As training progresses, the weight of the cross-entropy loss is increased to fine-tune classification accuracy.

3.3 Training Strategy

The training process is structured into multiple phases, each focusing on different aspects of the model's capabilities. In the early phases, the emphasis is on feature embedding and retrieval performance, facilitated by higher weights on metric learning losses and a relatively higher learning rate to encourage exploration of the feature space [7]. In subsequent phases, the focus shifts toward enhancing classification accuracy by increasing the weight of the cross-entropy loss and gradually reducing the learning rate to refine the model's predictions.

An exponential decay schedule is applied to the learning rate across the training phases to ensure smooth convergence and adapt the model to the data's complexities [22]. This dynamic adjustment of both loss weights and learning rates allows the model to balance and optimize multiple objectives effectively.

By integrating multi-scale encoding [15], a dynamic multi-loss function [7], and a phased training strategy, our proposed method effectively captures both global and local features. This leads to enhanced performance in medical image classification and retrieval tasks, particularly when dealing with complex and diverse datasets like ISIC-2018 [12] and ChestX-ray14 [1].

4 Experiments Setup

In this section, we provide a detailed description of the datasets used in our experiments, the evaluation metrics employed to assess the performance of our proposed method, and the hyperparameter settings configured during training.

4.1 Datasets

We conducted experiments using two widely recognized medical imaging datasets: ISIC-2018 [12,13] and ChestX-ray14 [1]. Fig. 1 presents sample images from both datasets, illustrating the diversity and complexity of skin lesions and chest pathologies.

4.1.1 ISIC-2018 Dataset

The International Skin Imaging Collaboration (ISIC) 2018 dataset [12] is a comprehensive collection of 10,015 high-resolution dermoscopic images designed to support automated skin cancer diagnosis. It includes seven lesion classes: Melanoma (MEL), Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratoses and Intraepithelial Carcinoma (AKIEC), Benign Keratosis (BKL), Dermatofibroma (DF), and Vascular Lesions (VASC). Released as part of the ISIC 2018 challenge [13], this dataset is widely used for benchmarking lesion segmentation, feature extraction, and disease classification algorithms. Key challenges include class imbalance, variations in image quality, and significant visual similarities between lesion types [13]. To maintain consistency with prior studies, we utilized the official training, validation, and test splits provided by the ISIC challenge [7,13].

4.1.2 ChestX-ray14 Dataset

The ChestX-ray14 dataset [1], compiled by Wang et al. [1] from the National Institutes of Health Clinical Center, is one of the largest publicly available chest X-ray databases. It consists of 112,120 frontal-view X-ray images from 30,805 unique patients, annotated with 14 common thoracic conditions: Atelectasis (ATL), Cardiomegaly (CARD), Effusion (EFF), Infiltration (INF), Mass (MAS), Nodule (NOD), Pneumonia (PNEU), Pneumothorax (PNE), Consolidation (CONS), Edema (EDE), Emphysema (EMP), Fibrosis (FIB), Pleural Thickening (PLT), and Hernia (HER). This dataset has been extensively used for training and evaluating deep learning models in chest pathology detection and classification [1,11,14,35]. Its primary challenges include multi-label annotations, high inter-class similarity, and variations in image quality [1]. To ensure fair comparison with existing methods, we followed the official training and testing splits provided by Wang et al. [1,14,35].

4.2 Evaluation Metrics

To comprehensively evaluate the performance of our proposed method on both classification and retrieval tasks, we employed a set of widely used metrics, consistent with prior studies [7,11].

For **classification tasks**, we used:

- **Precision:** The ratio of true positive predictions to the total number of positive predictions, measuring the model's accuracy in identifying relevant instances [11].
- **Recall:** The ratio of true positive predictions to the total number of actual positive instances, indicating the model's ability to capture all relevant cases [11].
- **F1-Score:** The harmonic mean of precision and recall, providing a single measure that balances both metrics [11].

- **Area Under the ROC Curve (AUC):** Measures the ability of the model to distinguish between classes across all classification thresholds, with higher values indicating better performance [11].

For **retrieval tasks**, we utilized:

- **Precision at K (P@K):** The proportion of relevant images among the top K retrieved images, assessing the quality of the retrieval results [7].
- **Recall at K (R@K):** The proportion of relevant images retrieved among the top K, relative to the total number of relevant images, indicating the retrieval system's completeness [7].
- **Mean Average Precision (mAP):** The average of precision values computed at the point of each relevant image retrieved, providing a summary of the precision-recall curve [7].

These metrics provide a comprehensive assessment of the model's performance in both accurately classifying medical images and effectively retrieving similar images for clinical reference. The use of multiple evaluation measures ensures a thorough analysis of the model's strengths and weaknesses in different aspects of medical image analysis [7].

4.3 Hyperparameter Settings

Our model was implemented using the PyTorch framework and trained on NVIDIA Tesla V100 GPUs. The training process was divided into three phases, each consisting of 70 epochs, totaling 210 epochs, similar to the training strategies employed in previous studies [7].

The key hyperparameters used in the experiments are as follows:

- **Optimizer:** We used the Adam optimizer for training, with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1e{-}8$.
- **Initial Learning Rate ($LR_{initial}$):** Set to 0.0004.
- **Learning Rate Decay:** The learning rate decayed exponentially in each phase as defined by:

$$LR_{\text{phase1}} = LR_{initial} \times \left(\frac{1}{4}\right)^{\frac{epoch}{70}} \tag{2}$$

$$LR_{\text{phase2}} = LR_{initial} \times \left(\frac{1}{16}\right)^{\frac{epoch-70}{70}} \tag{3}$$

$$LR_{\text{phase3}} = LR_{initial} \times \left(\frac{1}{64}\right)^{\frac{epoch-140}{70}} \tag{4}$$

This dynamic learning rate schedule was designed to facilitate the model's convergence and performance across different training phases [7].

- Batch Size: Set to 32 for both datasets.
- Loss Weights: The dynamic multi-loss function employed weights for cross-entropy loss (W_{CEL}), contrastive loss ($W_{\text{Contrastive}}$), and triplet loss (W_{Triplet}) as specified in Table 1, following the strategy outlined in [7].
- Image Preprocessing: Images were resized to a fixed size while maintaining aspect ratio and normalized using the mean and standard deviation of the ImageNet dataset [36].
- Data Augmentation: To enhance model generalization, random horizontal and vertical flips and rotations up to 15 degrees were applied during training, as commonly used in prior works [6,14].
- Patch Size: Each image was divided into patches of size 16×16 pixels for input into the ViT architecture [6].

- Positional Encoding: A 3-dimensional positional encoding vector representing scale, X-coordinate, and Y-coordinate was added to each patch embedding to capture spatial and scale information, following the methodology in [6].
- Regularization: Dropout with a rate of 0.1 was applied to prevent overfitting.

Table 1: Loss weight configurations for each training phase

Phase	W_{CEL}	$W_{\text{Contrastive}}$	W_{Triplet}
Phase 1	0.1	0.45	0.45
Phase 2	0.2	0.5	0.3
Phase 3	0.45	0.5	0.05

The hyperparameters were selected based on preliminary experiments and tuned to achieve optimal performance on both datasets. The dynamic adjustment of loss weights and learning rates was critical in balancing the multiple objectives of classification accuracy and retrieval effectiveness [7].

5 Results and Analysis

In this section, we present the experimental results of our proposed multi-scale Vision Transformer (ViT) architecture with a dynamic multi-loss function on the ISIC-2018 and ChestX-ray14 datasets. We compare our method with several state-of-the-art models, including traditional Convolutional Neural Networks (CNNs) and other transformer-based architectures, to demonstrate the effectiveness of our approach in both classification and retrieval tasks. We also perform ablation studies to assess the contribution of each component of our method.

5.1 Classification Performance

5.1.1 ISIC-2018 Dataset

Table 2 summarizes the classification performance of different models on the ISIC-2018 dataset. The proposed method achieves the highest overall precision, recall, and F1-score, significantly outperforming traditional CNNs like VGG16 and ResNet50, as well as the standard ViT and its variants.

Table 2: Comparison of classification performance across multiple deep learning models on the ISIC-2018 dataset. Metrics include Precision, Recall, and F1-Score for each skin lesion category. The proposed method integrates the Vision Transformer (ViT) with multi-scale and dynamic loss adjustments, achieving competitive results across most categories. **Bold** numbers indicate the best results in each category

Method	Metric	MEL	NV	BCC	AKIEC	BKL	DF	VASC	Overall
vgg16	Precision	0.5769	0.8738	0.9412	0.6519	0.5503	0.9383	0.7769	0.7585
	Recall	0.1500	0.9000	0.8000	0.8800	0.8200	0.7600	**0.9400**	0.7500
	F1-Score	0.2381	0.8867	0.8649	0.7489	0.6586	0.8398	0.8507	0.7268
squeezenet	Precision	0.7901	0.8817	0.8737	0.7436	0.7265	0.8140	0.7928	0.8032
	Recall	0.6400	0.8200	0.8300	0.8700	0.8500	0.7000	0.8800	0.7986
	F1-Score	0.7072	0.8497	0.8513	0.8018	0.7834	0.7527	0.8341	0.7972

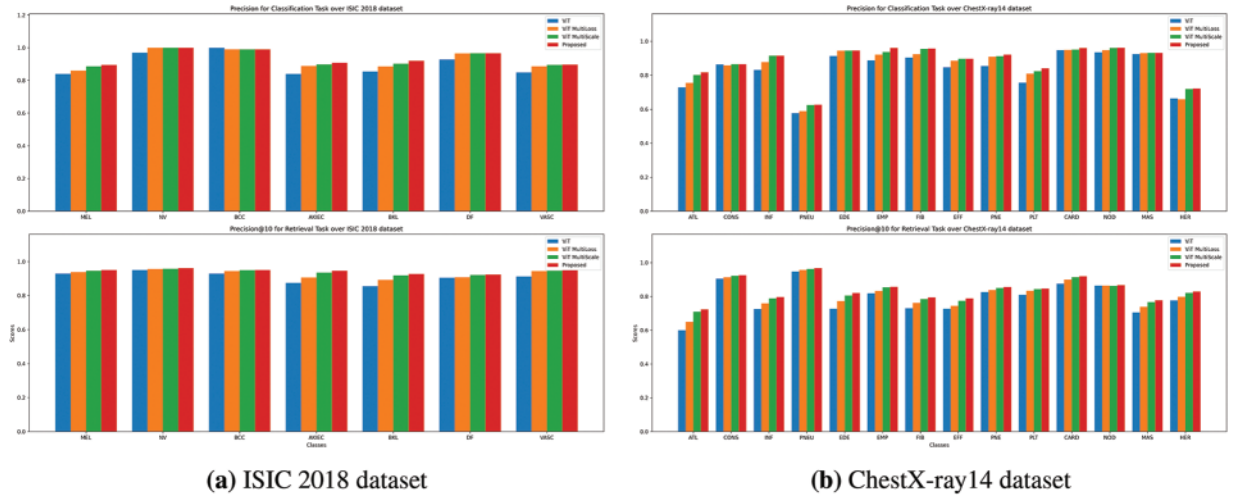
(Continued)

Table 2 (continued)

Method	Metric	MEL	NV	BCC	AKIEC	BKL	DF	VASC	Overall
mobilenet	Precision	0.8250	0.8614	0.9310	0.7360	0.7000	0.9155	0.7241	0.8133
	Recall	0.6600	0.8700	0.8100	0.9200	0.8400	0.6500	0.8400	0.7986
	F1-Score	0.7333	0.8657	0.8663	0.8178	0.7636	0.7602	0.7778	0.7978
resnet50	Precision	0.8681	0.9592	0.9667	0.8378	0.7981	0.9518	0.7317	0.8733
	Recall	0.7900	0.9400	0.8700	0.9300	0.8300	0.7900	0.9000	0.8643
	F1-Score	0.8272	0.9495	0.9158	0.8815	0.8137	0.8634	0.8072	0.8655
resnet101	Precision	0.8515	0.9588	0.9892	0.8598	0.8131	0.9419	0.8349	0.8927
	Recall	0.8600	0.9300	0.9200	0.9200	0.8700	0.8100	0.9100	0.8886
	F1-Score	0.8557	0.9442	0.9534	0.8889	0.8406	0.8710	0.8708	0.8892
googlenet	Precision	0.8462	0.9691	**1.0000**	0.8174	0.8700	0.9545	**0.9231**	0.9115
	Recall	0.8800	0.9400	0.9200	0.9400	0.8700	0.8400	**0.9600**	0.9071
	F1-Score	0.8627	0.9543	0.9583	0.8744	0.8700	0.8936	**0.9412**	0.9078
densenet121	Precision	**0.9216**	0.9700	0.9579	0.8447	0.8462	0.9205	0.8796	0.9058
	Recall	**0.9400**	**0.9700**	0.9100	0.8700	0.8800	0.8100	0.9500	0.9043
	F1-Score	**0.9307**	**0.9700**	0.9333	0.8571	0.8627	0.8617	0.9135	0.9042
ViT	Precision	0.8396	0.9691	**1.0000**	0.8393	0.8544	0.9286	0.8491	0.8971
	Recall	0.8900	0.9400	0.9200	0.9400	0.8800	0.7800	0.9000	0.8929
	F1-Score	0.8641	0.9543	0.9583	0.8868	0.8670	0.8478	0.8738	0.8932
ViT-Swin	Precision	0.8318	0.9691	**1.0000**	0.8393	0.8462	0.9277	0.8396	0.8934
	Recall	0.8900	0.9400	0.9100	0.9400	0.8800	0.7700	0.8900	0.8886
	F1-Score	0.8599	0.9543	0.9529	0.8868	0.8627	0.8415	0.8641	0.8889
Proposed	Precision	0.8942	**1.0000**	0.9896	**0.9074**	**0.9208**	**0.9663**	0.8962	**0.9392**
	Recall	0.9300	0.9600	**0.9500**	**0.9800**	**0.9300**	**0.8600**	0.9500	**0.9371**
	F1-Score	0.9118	**0.9796**	**0.9694**	**0.9423**	**0.9254**	**0.9101**	0.9223	**0.9373**

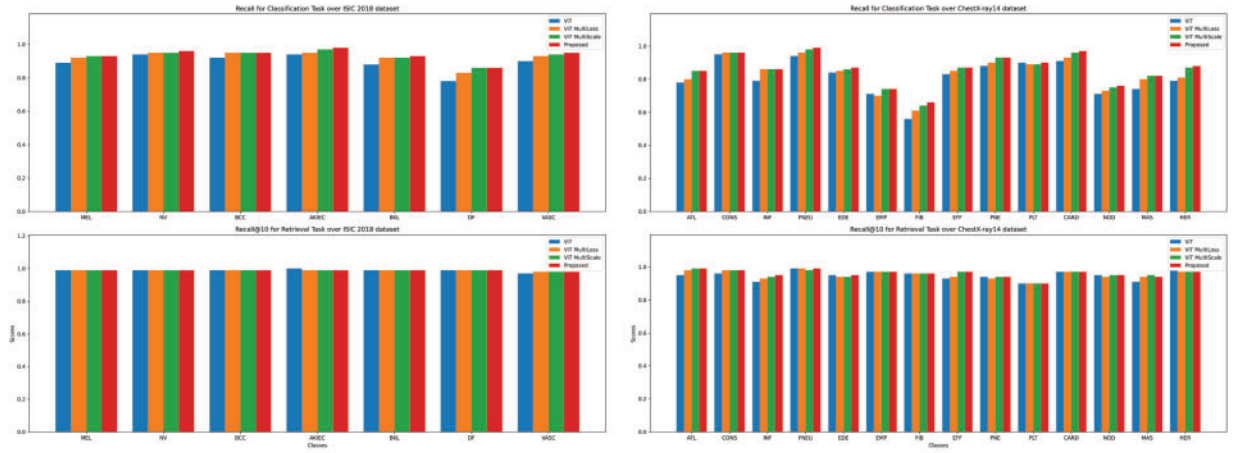
Figs. 3a, 4a, and 5a illustrate the precision, recall, and F1-score comparisons for each skin lesion category. The proposed method consistently outperforms other models across all categories, with particularly notable improvements in detecting melanoma (MEL) and basal cell carcinoma (BCC), which are critical for early diagnosis and treatment.

The normalized confusion matrix in Fig. 6a provides a visual representation of the classification accuracy for each class. The high values along the diagonal indicate that the model accurately classifies most samples, while the low off-diagonal values suggest minimal misclassifications. The Receiver Operating Characteristic (ROC) curves in Fig. 7a further demonstrate the model's strong ability to distinguish between different lesion types, with Area Under the Curve (AUC) values exceeding 0.92 for all classes.



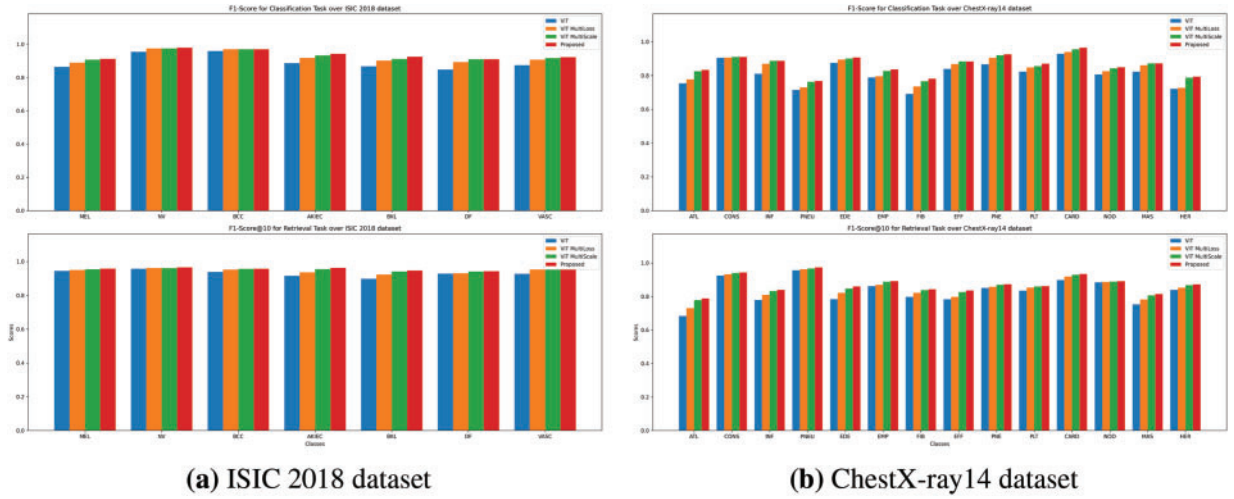
(a) ISIC 2018 dataset **(b)** ChestX-ray14 dataset

Figure 3: Precision comparison for Classification and Retrieval tasks. The precision scores are presented for both tasks (classification on top, retrieval on bottom) using ViT, ViT MultiScale, ViT MultiLoss, and the proposed method. The results highlight the superior performance of the proposed method across most lesion categories



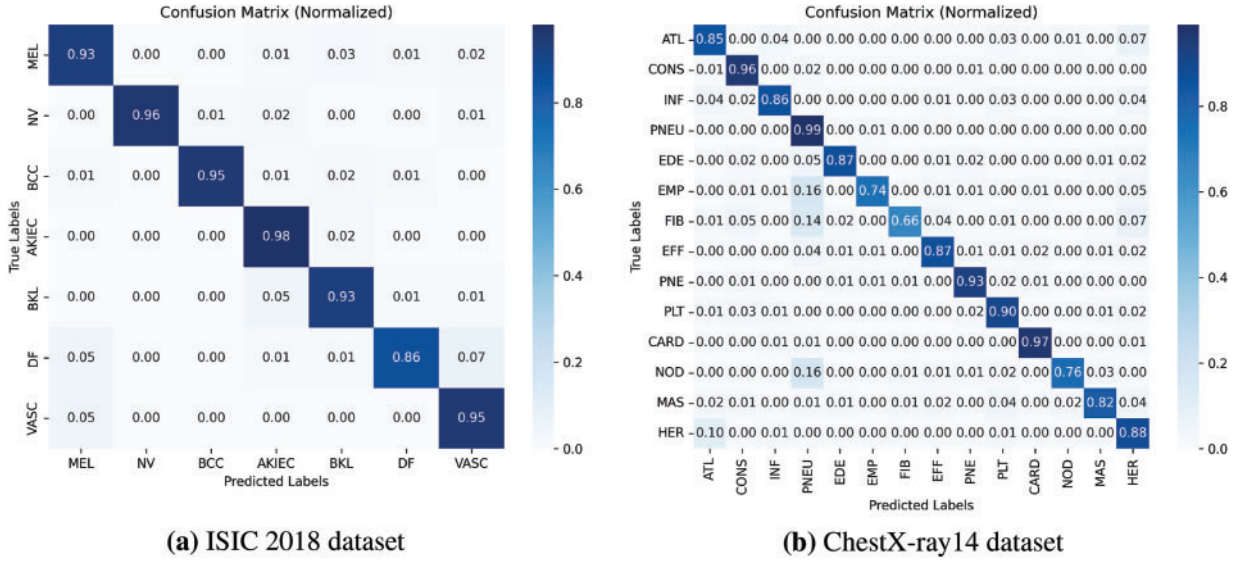
(a) ISIC 2018 dataset **(b)** ChestX-ray14 dataset

Figure 4: Recall comparison for Classification and Retrieval tasks. Recall scores for both tasks (classification on top, retrieval on bottom) are compared across the ViT, ViT MultiScale, ViT MultiLoss, and proposed method. The proposed method consistently shows improved recall, especially for challenging lesion categories



(a) ISIC 2018 dataset (b) ChestX-ray14 dataset

Figure 5: F1-Score comparison for Classification and Retrieval tasks. The bar charts illustrate F1-scores for both classification (top) and retrieval (bottom) tasks, comparing performance across ViT, ViT MultiScale, ViT MultiLoss, and proposed method. The proposed method achieves higher F1-Scores, indicating balanced precision and recall across categories



(a) ISIC 2018 dataset (b) ChestX-ray14 dataset

Figure 6: Normalized Confusion Matrix for classification task using the proposed method. This figure shows the classification accuracy for each skin lesion category, normalized to highlight the distribution of correct and incorrect predictions

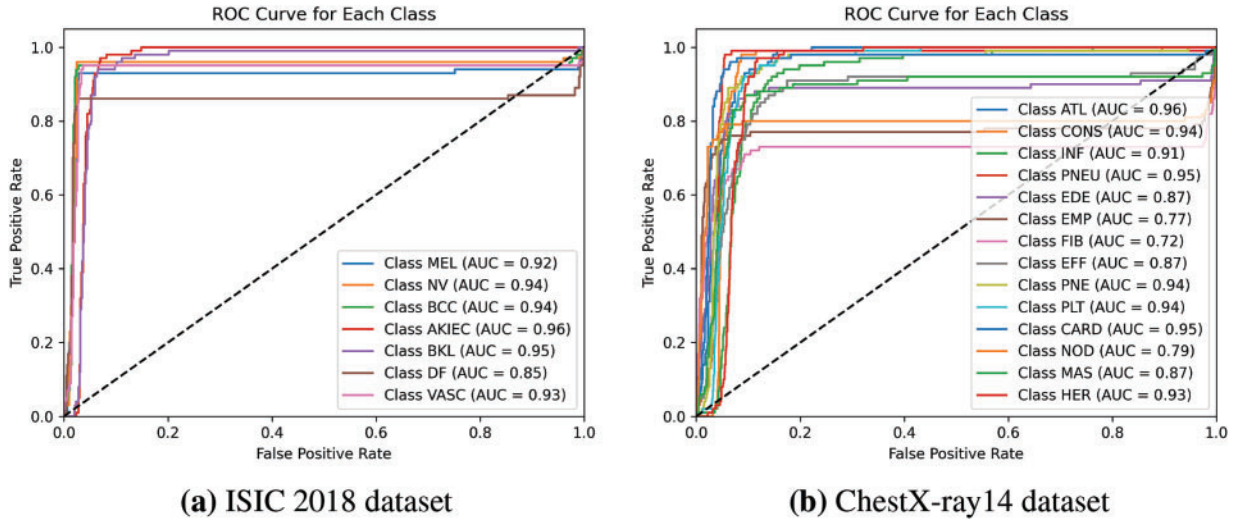


Figure 7: ROC Curve for each skin lesion class for classification task using the proposed method. The Receiver Operating Characteristic (ROC) curve and corresponding Area Under the Curve (AUC) values are plotted for each lesion category, showcasing the model's ability to distinguish between classes

5.1.2 ChestX-ray14 Dataset

Table 3 presents the classification results on the ChestX-ray14 dataset. Our proposed method achieves superior performance with higher precision, recall, and F1-score compared to other models. The improvement is particularly significant for pathologies such as cardiomegaly (CARD) and pneumonia (PNEU).

Table 3: Comparison of classification performance across various methods on the ChestX-ray14 dataset. This table includes Precision, Recall, and F1-Score for each chest pathology. The proposed ViT-based method with multi-scale encoding and dynamic loss achieves superior results in terms of overall diagnostic accuracy. **Bold** numbers indicate the best results in each category

Method	Metric	ATL	CONS	INF	PNEU	EDE	EMP	FIB	EFF	PNE	PLT	CARD	NOD	MAS	HER	Overall
vgg16	Precision	0.644	0.789	0.728	0.503	0.816	0.914	0.787	0.795	0.781	0.800	0.904	0.884	0.724	0.536	0.758
	Recall	0.560	0.900	0.670	**0.980**	0.800	0.640	0.370	0.620	0.890	0.800	0.850	0.760	0.630	0.740	0.729
	F1-Score	0.599	0.841	0.698	0.664	0.808	0.753	0.503	0.697	0.832	0.800	0.876	0.817	0.674	0.622	0.727
squeezenet	Precision	0.693	0.671	0.713	0.419	0.753	0.738	0.792	0.810	0.808	0.627	0.768	0.905	0.819	0.615	0.724
	Recall	0.610	**0.940**	0.720	0.850	0.580	0.450	0.380	0.680	0.800	0.790	0.760	0.670	0.680	0.720	0.688
	F1-Score	0.649	0.783	0.716	0.561	0.655	0.559	0.514	0.739	0.804	0.699	0.764	0.770	0.743	0.664	0.687
mobilenet	Precision	0.656	0.758	0.641	0.472	0.787	0.780	0.776	0.731	0.786	0.748	0.875	0.886	0.887	0.548	0.738
	Recall	0.610	0.940	0.660	0.940	0.630	0.710	0.450	0.570	0.810	0.800	0.770	0.700	0.630	0.690	0.708
	F1-Score	0.632	0.839	0.650	0.629	0.700	0.743	0.570	0.640	0.798	0.773	0.819	0.782	0.737	0.611	0.709
resnet50	Precision	0.618	**0.867**	0.786	0.569	0.804	0.818	0.849	0.780	0.833	0.748	0.899	0.859	0.872	0.642	0.782
	Recall	0.630	0.910	0.770	0.950	0.740	0.630	0.450	**0.780**	0.900	0.860	0.890	0.670	0.750	0.770	0.764
	F1-Score	0.624	0.888	0.778	0.712	0.771	0.712	0.588	0.780	0.865	0.800	0.894	0.753	0.806	0.700	0.762
resnet101	Precision	0.730	0.744	0.802	0.575	0.830	0.824	0.845	0.811	0.840	0.719	0.912	0.865	**0.946**	0.659	0.793
	Recall	0.730	0.930	0.770	0.880	0.730	**0.700**	0.490	0.770	0.890	0.870	0.930	0.640	0.700	0.810	0.774
	F1-Score	0.730	0.827	0.786	0.696	0.777	0.757	0.620	0.790	0.864	0.787	0.921	0.736	0.805	0.726	0.773
googlenet	Precision	0.720	0.853	0.762	0.596	0.886	0.926	0.841	0.878	0.840	0.837	0.892	0.886	0.878	0.557	0.811
	Recall	0.770	0.930	0.770	0.960	0.780	0.500	0.580	0.790	0.890	0.870	0.910	0.700	0.790	0.780	0.787
	F1-Score	0.744	0.890	0.766	0.736	0.830	0.649	0.686	0.832	0.864	0.853	0.901	0.782	0.832	0.650	0.787

(Continued)

Table 3 (continued)

Method	Metric	ATL	CONS	INF	PNEU	EDE	EMP	FIB	EFF	PNE	PLT	CARD	NOD	MAS	HER	Overall
densenet121	Precision	0.701	0.839	0.753	**0.752**	0.903	0.933	0.911	0.812	0.821	0.748	**0.968**	**0.952**	0.847	0.708	0.832
	Recall	0.750	0.940	0.730	0.970	0.840	0.700	0.720	0.820	0.870	0.890	0.910	**0.800**	0.830	0.750	0.823
	F1-Score	0.725	0.887	0.741	**0.847**	0.870	0.800	0.804	0.816	0.845	0.813	**0.938**	**0.870**	0.838	0.728	0.823
ViT	Precision	0.729	0.864	0.832	0.577	0.913	0.887	0.903	0.847	0.854	0.756	0.948	0.934	0.925	0.664	0.831
	Recall	0.780	0.950	0.790	0.940	0.840	0.710	0.560	0.830	0.880	0.900	0.910	0.710	0.740	0.790	0.809
	F1-Score	0.754	0.905	0.810	0.715	0.875	0.789	0.691	0.838	0.867	0.822	0.929	0.807	0.822	0.721	0.810
ViT-Swin	Precision	0.724	0.864	0.825	0.570	0.913	0.887	0.903	0.856	0.854	0.769	0.958	0.934	0.926	0.658	0.832
	Recall	0.760	0.950	0.800	0.940	0.840	0.710	0.560	0.830	0.880	0.900	0.910	0.710	0.750	0.790	0.809
	F1-Score	0.741	0.905	0.812	0.709	0.875	0.789	0.691	0.843	0.867	0.829	0.933	0.807	0.829	0.718	0.811
Proposed	Precision	**0.817**	0.865	**0.915**	0.627	**0.946**	**0.961**	**0.957**	**0.897**	**0.921**	**0.841**	0.960	0.962	0.932	**0.721**	**0.880**
	Recall	**0.850**	**0.960**	**0.860**	**0.990**	**0.870**	0.740	**0.660**	0.870	**0.930**	**0.900**	**0.970**	0.760	0.820	**0.880**	**0.861**
	F1-Score	**0.833**	**0.910**	**0.887**	0.767	**0.906**	**0.836**	**0.781**	**0.883**	**0.925**	**0.870**	**0.965**	0.849	0.872	**0.793**	**0.863**

Figs. 3b, 4b, and 5b display the performance metrics for each chest pathology. The proposed method shows marked improvements in detecting diseases that are challenging due to visual similarity or class imbalance, such as infiltration (INF) and edema (EDE).

The confusion matrix in Fig. 6b illustrates the classification performance across all pathologies. The ROC curves in Fig. 7b highlight the model's discriminative power, with high AUC values indicating strong classification capabilities.

5.2 Retrieval Performance

5.2.1 ISIC-2018 Dataset

Table 4 presents the retrieval performance on the ISIC-2018 dataset at different top-K levels (K = 2, 5, 10). The proposed method achieves the highest precision and recall at all levels, demonstrating its effectiveness in retrieving relevant images for clinical reference.

Table 4: Retrieval performance analysis for different methods over the ISIC-2018 dataset. The performance at top-K retrieval levels (K = 2, 5, 10) for each lesion category is presented. The best results in each category are highlighted in **bold**

Metric	@K	Method	MEL	NV	BCC	AKIEC	BKL	DF	VASC	Overall
vgg16	Precision	2	0.8950	0.9100	0.8700	0.8200	0.7700	0.9050	0.9250	0.8707
		5	0.8760	0.9040	0.8640	0.8140	0.7700	0.8840	0.9300	0.8631
		10	0.8620	0.8920	0.8450	0.8120	0.7530	0.8580	0.9320	0.8506
	Recall	2	0.9200	0.9300	0.9100	0.8900	0.8900	0.9500	0.9400	0.9186
		5	0.9400	0.9500	0.9500	0.9600	0.9200	0.9600	0.9500	0.9471
		10	0.9600	0.9500	0.9500	0.9600	0.9400	0.9800	0.9600	0.9571
	F1-Score	2	0.9033	0.9167	0.8833	0.8433	0.8100	0.9200	0.9300	0.8867
		5	0.8940	0.9159	0.8919	0.8615	0.8161	0.9097	0.9361	0.8893
		10	0.8856	0.9045	0.8742	0.8661	0.8086	0.8971	0.9388	0.8821

(Continued)

Table 4 (continued)

Metric	@K	Method	MEL	NV	BCC	AKIEC	BKL	DF	VASC	Overall
		2	0.8900	0.9350	0.8850	0.8100	0.7850	0.8900	0.9300	0.8750
	Precision	5	0.8660	0.9260	0.8740	0.8000	0.7720	0.8880	0.9100	0.8623
		10	0.8410	0.9040	0.8610	0.8010	0.7710	0.8650	0.8810	0.8463
		2	0.9300	0.9500	0.9200	0.9300	0.8600	0.9200	0.9600	0.9243
mobilenet	Recall	5	0.9600	0.9800	0.9500	0.9800	0.9100	0.9400	0.9700	0.9557
		10	0.9800	**1.0000**	0.9700	0.9900	0.9600	0.9800	0.9700	0.9786
		2	0.9033	0.9400	0.8967	0.8500	0.8100	0.9000	0.9400	0.8914
	F1-Score	5	0.8934	0.9421	0.8949	0.8547	0.8137	0.9042	0.9285	0.8902
		10	0.8748	0.9317	0.8880	0.8582	0.8266	0.8919	0.9086	0.8828
		2	0.9550	0.9600	0.8900	0.8750	0.8400	0.9150	0.9350	0.9100
	Precision	5	0.9440	0.9320	0.8800	0.8580	0.8300	0.9060	0.9180	0.8954
		10	0.9180	0.9280	0.8710	0.8450	0.8160	0.8910	0.9070	0.8823
		2	0.9800	0.9900	0.9100	0.9700	0.8900	0.9500	0.9600	0.9500
resnet50	Recall	5	0.9800	0.9900	0.9300	0.9900	0.9700	0.9700	0.9800	0.9729
		10	0.9800	0.9900	0.9600	0.9900	0.9700	0.9900	0.9900	0.9814
		2	0.9633	0.9700	0.8967	0.9067	0.8567	0.9267	0.9433	0.9233
	F1-Score	5	0.9579	0.9462	0.8951	0.9025	0.8716	0.9218	0.9329	0.9183
		10	0.9389	0.9396	0.8907	0.8981	0.8635	0.9122	0.9242	0.9096
		2	0.9550	0.9500	0.9300	0.8500	0.8100	0.9350	0.9650	0.9136
	Precision	5	0.9340	0.9400	0.9180	0.8360	0.8000	0.9120	0.9440	0.8977
		10	0.9020	0.9300	0.9160	0.8280	0.8010	0.8790	0.9230	0.8827
		2	0.9700	0.9800	0.9400	0.9500	0.8700	0.9600	0.9700	0.9486
resnet101	Recall	5	0.9800	0.9800	0.9600	0.9800	0.9600	0.9700	0.9700	0.9714
		10	0.9800	0.9800	0.9800	0.9900	0.9900	0.9700	0.9800	0.9814
		2	0.9600	0.9600	0.9333	0.8833	0.8300	0.9433	0.9667	0.9252
	F1-Score	5	0.9485	0.9509	0.9298	0.8810	0.8497	0.9309	0.9525	0.9205
		10	0.9260	0.9407	0.9290	0.8840	0.8593	0.9067	0.9374	0.9119
		2	0.9700	0.9450	**0.9800**	0.8800	0.8200	0.9550	**0.9750**	0.9321
	Precision	5	0.9520	0.9400	0.9520	0.8620	0.8300	0.9360	0.9660	0.9197
		10	0.9320	0.9380	0.9330	0.8560	0.8320	0.9070	0.9610	0.9084
		2	0.9900	0.9500	**0.9800**	0.9200	0.8700	0.9800	**0.9900**	0.9543
googlenet	Recall	5	0.9900	0.9500	**0.9800**	0.9500	0.9400	**0.9900**	**0.9900**	0.9700
		10	0.9900	0.9600	**0.9800**	0.9700	0.9800	**0.9900**	**0.9900**	0.9800
		2	0.9767	0.9467	**0.9800**	0.8933	0.8367	0.9633	**0.9800**	0.9395
	F1-Score	5	0.9614	0.9422	**0.9613**	0.8899	0.8648	0.9506	**0.9700**	0.9343
		10	0.9457	0.9413	**0.9442**	0.8946	0.8731	0.9285	**0.9654**	0.9275

(Continued)

Table 4 (continued)

Metric	@K	Method	MEL	NV	BCC	AKIEC	BKL	DF	VASC	Overall
densenet121	Precision	2	0.9700	0.9750	0.9650	0.8650	0.8450	0.9450	0.9600	0.9321
		5	0.9540	0.9720	0.9540	0.8380	0.8340	0.9280	0.9540	0.9191
		10	0.9470	0.9710	0.9270	0.8350	0.8290	0.9010	0.9510	0.9087
	Recall	2	0.9800	0.9800	0.9800	0.9100	0.9100	0.9800	0.9600	0.9571
		5	0.9800	0.9800	**1.0000**	0.9500	0.9800	0.9800	0.9700	0.9771
		10	0.9800	0.9800	**1.0000**	1.0000	0.9900	0.9800	0.9700	0.9857
	F1-Score	2	0.9733	0.9767	0.9700	0.8800	0.8667	0.9567	0.9600	0.9405
		5	0.9631	0.9746	0.9663	0.8743	0.8785	0.9430	0.9586	0.9369
		10	0.9543	0.9728	0.9440	0.8822	0.8756	0.9244	0.9561	0.9299
ViT	Precision	2	0.9700	0.9700	0.9450	0.8700	0.8500	0.9500	0.9450	0.9286
		5	0.9440	0.9620	0.9320	0.8720	0.8620	0.9220	0.9300	0.9177
		10	0.9290	0.9500	0.9290	0.8740	0.8550	0.9050	0.9120	0.9077
	Recall	2	0.9800	0.9800	0.9700	0.9600	0.9300	0.9600	0.9600	0.9629
		5	0.9800	**0.9900**	0.9800	0.9900	0.9800	0.9900	0.9700	0.9829
		10	0.9900	**0.9900**	0.9900	1.0000	0.9900	0.9900	0.9700	0.9886
	F1-Score	2	0.9733	0.9733	0.9533	0.9000	0.8767	0.9533	0.9500	0.9400
		5	0.9556	0.9694	0.9422	0.9098	0.8976	0.9412	0.9415	0.9368
		10	0.9454	0.9578	0.9394	0.9160	0.8984	0.9291	0.9272	0.9305
ViT-Swin	Precision	2	0.9700	**0.9850**	0.9450	0.8700	0.8600	0.9500	0.9600	0.9343
		5	0.9480	0.9640	0.9320	0.8740	0.8600	0.9180	0.9360	0.9189
		10	0.9270	0.9530	0.9230	0.8740	0.8470	0.9030	0.9180	0.9064
	Recall	2	0.9800	**0.9900**	0.9700	0.9600	0.9300	0.9600	0.9600	0.9643
		5	0.9800	**0.9900**	0.9800	0.9900	0.9800	0.9900	0.9700	0.9829
		10	0.9800	**0.9900**	0.9900	1.0000	0.9900	0.9900	0.9800	0.9886
	F1-Score	2	0.9733	**0.9867**	0.9533	0.9000	0.8833	0.9533	0.9600	0.9443
		5	0.9581	0.9712	0.9432	0.9102	0.8962	0.9387	0.9475	0.9379
		10	0.9417	0.9601	0.9339	0.9162	0.8919	0.9279	0.9352	0.9296
Proposed	Precision	2	**0.9750**	0.9750	0.9700	**0.9250**	**0.9450**	**0.9650**	0.9650	**0.9600**
		5	**0.9640**	0.9640	0.9640	**0.9400**	**0.9240**	0.9380	0.9540	**0.9497**
		10	**0.9500**	0.9610	0.9500	**0.9460**	**0.9270**	0.9230	0.9510	**0.9440**
	Recall	2	**0.9800**	0.9800	0.9800	**0.9600**	**0.9700**	**0.9900**	0.9800	**0.9771**
		5	**0.9900**	0.9800	**0.9900**	**0.9900**	**0.9900**	**0.9900**	0.9800	**0.9871**
		10	**0.9900**	0.9900	**0.9900**	**0.9900**	**0.9900**	0.9900	0.9800	**0.9886**
	F1-Score	2	**0.9767**	0.9767	**0.9733**	**0.9367**	**0.9533**	**0.9733**	0.9700	**0.9657**
		5	**0.9706**	0.9679	**0.9706**	**0.9574**	**0.9446**	0.9534	0.9606	**0.9607**
		10	**0.9584**	0.9652	**0.9579**	**0.9629**	**0.9477**	0.9426	0.9573	**0.9560**

Figs. 3a and 4a illustrate the precision and recall for retrieval tasks across different models. The proposed method consistently outperforms others, particularly for melanoma (MEL) and actinic keratosis (AKIEC), which are crucial for accurate diagnosis.

5.2.2 ChestX-ray14 Dataset

The retrieval results on the ChestX-ray14 dataset are shown in Table 5. Our method achieves superior retrieval precision and recall at various top-K levels, highlighting its ability to retrieve clinically relevant images across multiple pathologies.

Table 5: Retrieval performance analysis across different techniques on the ChestX-ray14 dataset. Precision at top-K retrieval levels (K = 2, 5, 10) for each chest pathology is shown. The best results in each category are highlighted in **bold**

Method	Metric	@K	ATL	CONS	INF	PNEU	EDE	EMP	FIB	EFF	PNE	PLT	CARD	NOD	MAS	HER	Overall
vgg16	Precision	2	0.435	0.875	0.565	0.965	0.705	0.710	0.730	0.625	0.845	0.735	0.815	0.830	0.580	0.635	0.718
		5	0.432	0.858	0.576	0.964	0.714	0.674	0.720	0.608	0.844	0.746	0.818	0.800	0.582	0.632	0.712
		10	0.426	0.865	0.565	0.951	0.689	0.651	0.671	0.562	0.848	0.723	0.808	0.775	0.542	0.618	0.692
	Recall	2	0.580	0.970	0.680	0.970	0.770	0.800	0.820	0.790	0.890	0.810	0.870	0.870	0.700	0.730	0.804
		5	0.710	0.980	0.810	0.970	0.860	0.890	0.920	0.890	0.910	0.870	0.900	0.910	0.830	0.880	0.881
		10	0.820	0.980	0.850	0.970	0.880	0.930	0.940	0.960	0.920	0.900	0.920	0.950	0.870	0.930	0.916
	F1-Score	2	0.483	0.907	0.603	0.967	0.727	0.740	0.760	0.680	0.860	0.760	0.833	0.843	0.620	0.667	0.746
		5	0.506	0.899	0.640	0.967	0.751	0.729	0.779	0.685	0.866	0.784	0.842	0.835	0.647	0.705	0.759
		10	0.514	0.902	0.635	0.959	0.735	0.712	0.749	0.651	0.871	0.767	0.836	0.823	0.618	0.707	0.749
mobilenet	Precision	2	0.470	0.915	0.580	0.930	0.640	0.730	0.735	0.580	0.770	0.760	0.775	0.805	0.575	0.555	0.701
		5	0.456	0.880	0.588	0.936	0.626	0.710	0.646	0.546	0.758	0.756	0.758	0.796	0.556	0.594	0.686
		10	0.431	0.874	0.574	0.930	0.597	0.677	0.589	0.516	0.739	0.728	0.742	0.772	0.519	0.584	0.662
	Recall	2	0.650	0.960	0.720	0.950	0.760	0.830	0.860	0.720	0.810	0.810	0.860	0.860	0.660	0.710	0.797
		5	0.820	0.970	0.800	0.980	0.910	0.870	0.910	0.900	0.860	0.860	0.910	0.900	0.810	0.860	0.883
		10	0.860	0.970	0.870	0.980	0.930	0.920	0.940	0.950	0.890	0.910	0.940	0.910	0.840	0.950	0.919
	F1-Score	2	0.530	0.930	0.627	0.937	0.680	0.763	0.777	0.627	0.783	0.777	0.803	0.823	0.603	0.607	0.733
		5	0.551	0.912	0.652	0.946	0.707	0.760	0.725	0.638	0.784	0.788	0.799	0.829	0.617	0.671	0.741
		10	0.532	0.908	0.647	0.942	0.690	0.742	0.695	0.616	0.772	0.774	0.787	0.815	0.588	0.684	0.728
resnet50	Precision	2	0.555	0.875	0.735	0.965	0.730	0.805	0.765	0.675	0.830	0.835	0.845	0.855	0.650	0.675	0.771
		5	0.498	0.864	0.678	0.948	0.668	0.796	0.702	0.662	0.800	0.810	0.834	0.844	0.640	0.688	0.745
		10	0.463	0.850	0.666	0.946	0.600	0.745	0.615	0.645	0.784	0.787	0.807	0.817	0.616	0.693	0.717
	Recall	2	0.700	0.920	0.840	0.980	0.840	0.860	0.870	0.790	0.880	0.880	0.900	0.900	0.770	0.810	0.853
		5	0.840	0.950	0.890	0.980	0.910	0.910	0.910	0.840	0.900	0.900	0.950	0.940	0.830	0.960	0.908
		10	0.870	0.970	0.930	0.990	0.930	0.930	0.950	0.890	0.930	0.920	0.960	0.970	0.880	0.990	0.936
	F1-Score	2	0.603	0.890	0.770	0.970	0.767	0.823	0.800	0.713	0.847	0.850	0.863	0.870	0.690	0.720	0.798
		5	0.588	0.891	0.740	0.957	0.744	0.829	0.764	0.715	0.827	0.840	0.865	0.871	0.695	0.767	0.792
		10	0.565	0.886	0.743	0.956	0.702	0.799	0.710	0.706	0.818	0.826	0.839	0.858	0.682	0.784	0.777
resnet101	Precision	2	0.615	0.885	0.730	0.950	0.765	0.825	0.755	0.720	0.835	0.795	0.870	0.900	0.745	0.790	0.799
		5	0.586	0.882	0.694	0.930	0.708	0.778	0.748	0.708	0.826	0.798	0.868	0.860	0.680	0.760	0.773
		10	0.566	0.874	0.684	0.906	0.658	0.738	0.725	0.687	0.811	0.791	0.850	0.841	0.641	0.755	0.752
	Recall	2	0.750	0.930	0.840	0.970	0.840	0.910	0.830	0.800	0.870	0.840	0.900	0.930	0.860	0.910	0.870
		5	0.870	0.940	0.890	0.990	0.910	0.940	0.940	0.870	0.920	0.880	0.950	0.940	0.880	0.940	0.918
		10	0.910	0.940	0.890	0.990	0.930	0.940	0.970	0.890	0.930	0.910	0.970	0.950	0.910	0.970	0.936
	F1-Score	2	0.660	0.900	0.767	0.957	0.790	0.853	0.780	0.747	0.847	0.810	0.880	0.910	0.783	0.830	0.822
		5	0.660	0.901	0.743	0.947	0.767	0.827	0.801	0.757	0.848	0.825	0.890	0.882	0.731	0.815	0.814
		10	0.653	0.897	0.739	0.926	0.733	0.798	0.794	0.746	0.838	0.824	0.877	0.868	0.708	0.826	0.802
googlenet	Precision	2	0.655	0.880	0.740	0.965	0.735	0.850	0.770	0.730	0.820	0.835	0.865	0.890	0.750	0.755	0.803
		5	0.644	0.894	0.738	0.960	0.730	0.826	0.732	0.728	0.828	0.822	0.870	0.874	0.744	0.740	0.795
		10	0.618	0.880	0.730	0.956	0.706	0.807	0.688	0.697	0.831	0.818	0.865	0.854	0.729	0.737	0.780
	Recall	2	0.770	0.910	0.840	0.980	0.830	0.890	0.850	0.810	0.860	0.870	0.900	0.910	0.800	0.850	0.862
		5	0.840	0.960	0.910	0.980	0.860	0.920	0.940	0.900	0.920	0.890	0.940	0.940	0.870	0.950	0.916
		10	0.880	0.960	0.940	0.980	0.890	0.940	0.960	0.960	0.940	0.900	0.960	0.960	0.880	0.980	0.938
	F1-Score	2	0.693	0.890	0.773	0.970	0.767	0.863	0.797	0.757	0.833	0.847	0.877	0.897	0.767	0.787	0.823
		5	0.705	0.910	0.790	0.965	0.772	0.850	0.791	0.771	0.850	0.842	0.888	0.893	0.778	0.801	0.829
		10	0.690	0.900	0.790	0.961	0.759	0.840	0.763	0.758	0.855	0.841	0.887	0.879	0.764	0.804	0.821
densenet121	Precision	2	0.620	0.890	0.730	0.945	0.805	0.900	0.780	0.770	0.830	0.780	0.890	0.890	0.755	0.715	0.807
		5	0.592	0.896	0.694	0.940	0.778	0.880	0.774	0.736	0.832	0.804	0.896	0.872	0.746	0.736	0.798
		10	0.594	0.901	0.686	0.940	0.762	0.854	0.753	0.732	0.828	0.810	0.891	0.847	0.720	0.748	0.790
	Recall	2	0.730	0.960	0.790	0.960	0.900	0.940	0.850	0.850	0.870	0.840	0.930	0.940	0.820	0.830	0.872
		5	0.830	0.980	0.880	0.960	0.950	0.950	0.910	0.880	0.920	0.920	0.960	0.950	0.890	0.940	0.923
		10	0.890	**0.980**	0.910	0.970	0.960	0.960	0.940	0.930	0.930	0.930	0.970	0.960	0.900	0.980	0.944
	F1-Score	2	0.657	0.913	0.750	0.950	0.837	0.913	0.803	0.797	0.843	0.800	0.903	0.907	0.777	0.753	0.829
		5	0.661	0.925	0.750	0.945	0.828	0.900	0.817	0.771	0.855	0.836	0.915	0.890	0.783	0.799	0.834
		10	0.667	0.929	0.748	0.947	0.815	0.879	0.808	0.773	0.852	0.844	0.910	0.872	0.765	0.820	0.831

(Continued)

Table 5 (continued)

Method	Metric	@K	ATL	CONS	INF	PNEU	EDE	EMP	FIB	EFF	PNE	PLT	CARD	NOD	MAS	HER	Overall
ViT	Precision	2	0.620	0.900	0.745	0.970	0.810	0.910	0.840	0.780	0.830	0.830	0.910	0.910	0.760	0.800	0.830
		5	0.618	0.902	0.738	0.954	0.782	0.862	0.798	0.754	0.832	0.818	0.894	0.900	0.728	0.778	0.811
		10	0.600	0.905	0.725	0.947	0.727	0.818	0.730	0.727	0.825	0.809	0.875	0.863	0.705	0.776	0.788
	Recall	2	0.730	0.950	0.860	0.980	0.870	0.930	0.910	0.890	0.870	0.870	0.940	0.920	0.810	0.910	0.889
		5	0.880	0.960	0.900	0.990	0.890	0.960	0.930	0.910	0.920	0.870	0.970	0.930	0.870	0.960	0.924
		10	0.950	0.960	0.910	0.990	0.950	0.970	0.960	0.930	0.940	0.900	0.970	0.950	0.910	0.980	0.948
	F1-Score	2	0.657	0.917	0.783	0.973	0.830	0.917	0.863	0.817	0.843	0.843	0.920	0.913	0.777	0.837	0.849
		5	0.690	0.923	0.790	0.963	0.817	0.893	0.843	0.800	0.854	0.838	0.916	0.908	0.771	0.835	0.846
		10	0.684	0.926	0.780	0.957	0.784	0.862	0.797	0.783	0.851	0.835	0.898	0.885	0.754	0.839	0.831
ViT-Swin	Precision	2	0.630	0.900	0.750	0.970	0.815	0.905	0.835	0.765	0.830	0.835	0.910	0.910	0.765	0.805	0.830
		5	0.612	0.900	0.742	0.954	0.788	0.864	0.804	0.748	0.832	0.816	0.898	0.902	0.732	0.782	0.812
		10	0.602	0.906	0.732	0.945	0.736	0.814	0.728	0.721	0.827	0.804	0.876	0.864	0.706	0.779	0.789
	Recall	2	0.740	0.950	0.860	0.980	0.870	0.940	0.900	0.860	0.860	0.870	0.950	0.920	0.820	0.910	0.888
		5	0.880	0.960	0.900	0.990	0.910	0.950	0.940	0.910	0.920	0.870	0.980	0.930	0.860	0.960	0.926
		10	0.940	0.960	0.910	0.990	0.960	0.970	0.950	0.940	0.940	0.900	0.980	0.950	0.900	0.990	0.949
	F1-Score	2	0.667	0.917	0.787	0.973	0.833	0.917	0.857	0.797	0.840	0.847	0.923	0.913	0.783	0.840	0.850
		5	0.683	0.921	0.795	0.963	0.825	0.892	0.849	0.795	0.854	0.836	0.922	0.909	0.773	0.836	0.847
		10	0.686	0.927	0.785	0.955	0.794	0.858	0.795	0.778	0.852	0.832	0.900	0.885	0.754	0.842	0.832
Proposed	Precision	2	**0.775**	**0.925**	**0.810**	**0.975**	**0.865**	**0.925**	**0.880**	**0.830**	**0.870**	**0.870**	**0.950**	**0.915**	**0.800**	**0.870**	**0.876**
		5	**0.758**	**0.926**	**0.804**	**0.968**	**0.854**	**0.904**	**0.842**	**0.798**	**0.858**	**0.854**	**0.940**	**0.902**	**0.792**	**0.864**	**0.862**
		10	**0.723**	**0.925**	**0.796**	**0.967**	**0.820**	**0.857**	**0.794**	**0.788**	**0.855**	**0.846**	**0.919**	**0.867**	**0.777**	**0.829**	**0.840**
	Recall	2	**0.850**	**0.950**	**0.870**	**0.980**	**0.890**	**0.940**	**0.930**	**0.890**	**0.910**	**0.880**	**0.970**	**0.920**	**0.870**	**0.920**	**0.912**
		5	**0.950**	**0.980**	**0.920**	**0.990**	**0.940**	**0.960**	**0.940**	**0.920**	**0.930**	**0.880**	**0.970**	**0.940**	**0.920**	**0.960**	**0.943**
		10	**0.990**	**0.980**	**0.950**	**0.990**	**0.950**	**0.970**	**0.960**	**0.970**	**0.940**	**0.900**	**0.970**	**0.950**	**0.940**	**0.970**	**0.959**
	F1-Score	2	**0.800**	**0.933**	**0.830**	**0.977**	**0.873**	**0.930**	**0.897**	**0.850**	**0.883**	**0.873**	**0.957**	**0.917**	**0.823**	**0.887**	**0.888**
		5	**0.813**	**0.943**	**0.840**	**0.973**	**0.879**	**0.919**	**0.876**	**0.835**	**0.877**	**0.865**	**0.951**	**0.914**	**0.825**	**0.897**	**0.886**
		10	**0.789**	**0.943**	**0.840**	**0.973**	**0.860**	**0.892**	**0.844**	**0.836**	**0.873**	**0.862**	**0.934**	**0.892**	**0.815**	**0.873**	**0.873**

As depicted in Figs. 3b and 4b, the proposed method shows enhanced retrieval performance for diseases such as pneumothorax (PNEU) and cardiomegaly (CARD), which are critical for patient care.

5.3 Ablation Studies

To evaluate the contribution of each component of our proposed method, we conducted ablation studies by comparing the performance of the baseline ViT model, ViT with multi-scale encoding (ViT MultiScale), ViT with the dynamic multi-loss function (ViT MultiLoss), and the full proposed method.

5.3.1 Impact of Multi-Scale Encoding

Tables 6 and 7 show that incorporating multi-scale encoding into the ViT architecture (ViT MultiScale) leads to improved classification and retrieval performance compared to the baseline ViT model. This enhancement is attributed to the model's ability to capture fine-grained details and global context by processing images at multiple scales.

Table 6: Ablation study of the proposed method over the ISIC-2018 dataset for medical image classification and retrieval tasks. The table examines the impact of multi-scale encoding and multi-loss adjustments on classification and retrieval performance. Metrics include Precision, Recall, and F1-Score, evaluated across all lesion categories. The best results in each category are highlighted in **bold**

Medical Image Classification Task										
Metric		**Method**	**MEL**	**NV**	**BCC**	**AKIEC**	**BKL**	**DF**	**VASC**	**Overall**
Precision		ViT	0.8396	0.9691	**1.0000**	0.8393	0.8544	0.9286	0.8491	0.8971
		ViT MultiLoss	0.8598	**1.0000**	0.9896	0.8879	0.8846	0.9651	0.8857	0.9247
		ViT MultiScale	0.8857	**1.0000**	0.9896	0.8981	0.9020	**0.9663**	0.8952	0.9338
		Proposed	**0.8942**	**1.0000**	0.9896	**0.9074**	**0.9208**	**0.9663**	**0.8962**	**0.9392**
Recall		ViT	0.8900	0.9400	0.9200	0.9400	0.8800	0.7800	0.9000	0.8929
		ViT MultiLoss	0.9200	0.9500	0.9500	0.9500	0.9200	0.8300	0.9300	0.9214
		ViT MultiScale	**0.9300**	0.9500	0.9500	0.9700	0.9200	**0.8600**	0.9400	0.9314
		Proposed	**0.9300**	**0.9600**	0.9500	**0.9800**	**0.9300**	**0.8600**	**0.9500**	**0.9371**
F1-Score		ViT	0.8641	0.9543	0.9583	0.8868	0.8670	0.8478	0.8738	0.8932
		ViT MultiLoss	0.8889	0.9744	0.9694	0.9179	0.9020	0.8925	0.9073	0.9218
		ViT MultiScale	0.9073	0.9744	0.9694	0.9327	0.9109	**0.9101**	0.9171	0.9317
		Proposed	**0.9118**	**0.9796**	0.9694	**0.9423**	**0.9254**	**0.9101**	**0.9223**	**0.9373**
Medical Image Retrieval Task										
Metric	**@K**	**Method**	**MEL**	**NV**	**BCC**	**AKIEC**	**BKL**	**DF**	**VASC**	**Overall**
Precision	2	ViT	0.9700	0.9700	0.9450	0.8700	0.8500	0.9500	0.9450	0.9286
		ViT MultiLoss	0.9650	0.9700	0.9700	0.9000	0.8850	0.9550	**0.9800**	0.9464
		ViT MultiScale	0.9650	**0.9750**	0.9700	0.9150	0.9350	**0.9650**	**0.9750**	0.9571
		Proposed	**0.9750**	**0.9750**	0.9700	**0.9250**	**0.9450**	**0.9650**	0.9650	**0.9600**
	5	ViT	0.9440	0.9620	0.9320	0.8720	0.8620	0.9220	0.9300	0.9177
		ViT MultiLoss	0.9500	0.9600	0.9540	0.9020	0.8960	0.9260	0.9560	0.9349
		ViT MultiScale	0.9640	0.9600	0.9620	0.9320	0.9200	**0.9400**	0.9540	0.9474
		Proposed	**0.9640**	**0.9640**	**0.9640**	**0.9400**	**0.9240**	0.9380	0.9540	**0.9497**
	10	ViT	0.9290	0.9500	0.9290	0.8740	0.8550	0.9050	0.9120	0.9077
		ViT MultiLoss	0.9380	0.9560	0.9440	0.9060	0.8920	0.9080	0.9440	0.9269
		ViT MultiScale	0.9460	0.9570	0.9490	0.9350	0.9180	0.9210	0.9460	0.9389
		Proposed	**0.9500**	**0.9610**	**0.9500**	**0.9460**	**0.9270**	**0.9230**	**0.9510**	**0.9440**
Recall	2	ViT	0.9800	0.9800	0.9700	0.9600	0.9300	0.9600	0.9600	0.9629
		ViT MultiLoss	0.9800	0.9800	**0.9900**	0.9600	0.9400	0.9800	0.9800	0.9729
		ViT MultiScale	0.9800	0.9800	0.9800	0.9600	0.9600	**0.9900**	0.9800	0.9757
		Proposed	0.9800	0.9800	0.9800	0.9600	**0.9700**	**0.9900**	0.9800	**0.9771**
	5	ViT	0.9800	**0.9900**	0.9800	**0.9900**	0.9800	0.9900	0.9700	0.9829
		ViT MultiLoss	0.9800	0.9800	**0.9900**	0.9800	0.9800	0.9900	0.9800	0.9829
		ViT MultiScale	**0.9900**	0.9800	**0.9900**	**0.9900**	0.9800	0.9900	0.9800	**0.9857**
		Proposed	**0.9900**	0.9800	**0.9900**	**0.9900**	**0.9900**	0.9900	0.9800	**0.9871**

(Continued)

Table 6 (continued)

Medical Image Retrieval Task									
Metric	**@K**	**Method**	**MEL**	**NV**	**BCC**	**AKIEC**	**BKL**	**DF**	**VASC**	**Overall**
F1-Score	10	ViT	**0.9900**	**0.9900**	**0.9900**	**1.0000**	**0.9900**	0.9900	0.9700	**0.9886**
		ViT MultiLoss	**0.9900**	**0.9900**	**0.9900**	0.9900	**0.9900**	0.9900	0.9800	**0.9886**
		ViT MultiScale	**0.9900**	**0.9900**	**0.9900**	0.9900	**0.9900**	0.9900	0.9800	**0.9886**
		Proposed	**0.9900**	**0.9900**	**0.9900**	0.9900	**0.9900**	0.9900	0.9800	**0.9886**
	2	ViT	0.9733	0.9733	0.9533	0.9000	0.8767	0.9533	0.9500	0.9400
		ViT MultiLoss	0.9700	0.9733	**0.9767**	0.9200	0.9033	0.9633	**0.9800**	0.9552
		ViT MultiScale	0.9700	**0.9767**	0.9733	0.9300	0.9433	**0.9733**	0.9767	0.9633
		Proposed	**0.9767**	**0.9767**	0.9733	**0.9367**	**0.9533**	**0.9733**	0.9700	**0.9657**
	5	ViT	0.9556	0.9694	0.9422	0.9098	0.8976	0.9412	0.9415	0.9368
		ViT MultiLoss	0.9596	0.9654	0.9634	0.9271	0.9227	0.9438	0.9646	0.9495
		ViT MultiScale	0.9706	0.9648	0.9694	0.9509	0.9399	**0.9548**	0.9612	0.9588
		Proposed	**0.9706**	**0.9679**	**0.9706**	**0.9574**	**0.9446**	0.9534	**0.9606**	**0.9607**
	10	ViT	0.9454	0.9578	0.9394	0.9160	0.8984	0.9291	0.9272	0.9305
		ViT MultiLoss	0.9506	0.9620	0.9527	0.9364	0.9239	0.9313	0.9531	0.9443
		ViT MultiScale	0.9551	0.9616	0.9572	0.9554	0.9415	0.9413	0.9536	0.9522
		Proposed	**0.9584**	**0.9652**	**0.9579**	**0.9629**	**0.9477**	**0.9426**	**0.9573**	**0.9560**

5.3.2 Impact of Dynamic Multi-Loss Function

The results also indicate that applying the dynamic multi-loss function (ViT MultiLoss) improves performance over the baseline by better balancing the learning objectives of classification and retrieval tasks. The adaptive weighting of loss components allows the model to focus on different aspects during training, enhancing overall robustness.

5.3.3 Combined Effect

The full proposed method, which integrates both multi-scale encoding and the dynamic multi-loss function, achieves the highest performance in both datasets. This demonstrates the synergistic effect of combining these two components, leading to significant improvements over individual enhancements.

Table 7: Ablation study of the proposed method over the ChestX-ray14 dataset for medical image classification and retrieval tasks. The study evaluates the effect of multi-scale encoding and dynamic loss adjustments in enhancing diagnostic accuracy. Performance is measured in terms of Precision, Recall, and F1-Score. **Bold** numbers indicate the best performance in each category

Medical Image Classification Task																
Metric		**Method**	**ATL**	**CONS**	**INF**	**PNEU**	**EDE**	**EMP**	**FIB**	**EFF**	**PNE**	**PLT**	**CARD**	**NOD**	**MAS**	**HER**	**Overall**
Precision		ViT	0.7290	0.8636	0.8316	0.5767	0.9130	0.8875	0.9032	0.8469	0.8544	0.7563	0.9479	0.9342	0.9250	0.6639	0.8309
		ViT MultiLoss	0.7547	0.8571	0.8776	0.5890	0.9444	0.9211	0.9242	0.8854	0.9091	0.8091	0.9490	0.9481	0.9302	0.6585	0.8541
		ViT MultiScale	0.8019	**0.8649**	**0.9149**	0.6242	0.9451	0.9367	0.9552	**0.8969**	0.9118	0.8241	0.9505	0.9615	0.9318	0.7190	0.8742
		Proposed	**0.8173**	**0.8649**	**0.9149**	**0.6266**	**0.9457**	**0.9610**	**0.9565**	**0.8969**	**0.9208**	**0.8411**	**0.9604**	**0.9620**	**0.9318**	**0.7213**	**0.8801**
Recall		ViT	0.7800	0.9500	0.7900	0.9400	0.8400	0.7100	0.5600	0.8300	0.8800	0.9000	0.9100	0.7100	0.7400	0.7900	0.8093
		ViT MultiLoss	0.8000	**0.9600**	0.8600	0.9600	0.8500	0.7000	0.6100	0.8500	0.9000	0.8900	0.9300	0.7300	0.8000	0.8100	0.8321
		ViT MultiScale	**0.8500**	**0.9600**	0.8600	0.9800	0.8600	**0.7400**	0.6400	**0.8700**	**0.9300**	0.8900	0.9600	0.7500	0.8200	0.8700	0.8557
		Proposed	**0.8500**	**0.9600**	**0.8600**	**0.9900**	**0.8700**	**0.7400**	**0.6600**	**0.8700**	**0.9300**	**0.9000**	**0.9700**	**0.7600**	**0.8200**	**0.8800**	**0.8614**
F1-Score		ViT	0.7536	0.9048	0.8103	0.7148	0.8750	0.7889	0.6914	0.8384	0.8670	0.8219	0.9286	0.8068	0.8222	0.7215	0.8104
		ViT MultiLoss	0.7767	0.9057	0.8687	0.7300	0.8947	0.7955	0.7349	0.8673	0.9045	0.8476	0.9394	0.8249	0.8602	0.7265	0.8340
		ViT MultiScale	0.8252	0.9100	**0.8866**	0.7626	0.9005	0.8268	0.7665	0.8832	0.9208	0.8558	0.9552	0.8427	0.8723	0.7873	0.8568
		Proposed	**0.8333**	**0.9100**	**0.8866**	**0.7674**	**0.9062**	**0.8362**	**0.7811**	**0.8832**	**0.9254**	**0.8696**	**0.9652**	**0.8492**	**0.8723**	**0.7928**	**0.8627**
Medical Image Retrieval Task																	
Metric	**@K**	**Method**	**ATL**	**CONS**	**INF**	**PNEU**	**EDE**	**EMP**	**FIB**	**EFF**	**PNE**	**PLT**	**CARD**	**NOD**	**MAS**	**HER**	**Overall**
Precision	2	ViT	0.6200	0.9000	0.7450	0.9700	0.8100	0.9100	0.8400	0.7800	0.8300	0.8300	0.9100	0.9100	0.7600	0.8000	0.8296
		ViT MultiLoss	0.6900	0.9250	0.7800	0.9700	0.8450	0.8950	0.8450	0.7650	0.8500	0.8500	0.9350	0.9100	0.7700	0.8350	0.8475
		ViT MultiScale	0.7650	0.9250	0.8100	0.9700	0.8550	0.9250	0.8800	0.8100	0.8700	0.8650	0.9450	0.9100	0.7900	0.8700	0.8707
		Proposed	**0.7750**	**0.9250**	**0.8100**	**0.9750**	**0.8650**	**0.9250**	**0.8800**	**0.8300**	**0.8700**	**0.8700**	**0.9500**	**0.9150**	**0.8000**	**0.8700**	**0.8757**
	5	ViT	0.6180	0.9020	0.7380	0.9540	0.7820	0.8620	0.7980	0.7540	0.8320	0.8180	0.8940	0.9000	0.7280	0.7780	0.8113
		ViT MultiLoss	0.6780	0.9180	0.7660	0.9620	0.8120	0.8820	0.8200	0.7680	0.8440	0.8420	0.9200	0.8920	0.7680	0.8120	0.8346
		ViT MultiScale	0.7380	0.9220	0.7980	0.9620	0.8480	0.9000	0.8380	0.7900	0.8580	0.8500	0.9320	0.8920	0.7900	0.8500	0.8549
		Proposed	**0.7580**	**0.9260**	**0.8040**	**0.9680**	**0.8540**	**0.9040**	**0.8420**	**0.7980**	**0.8580**	**0.8540**	**0.9400**	**0.9020**	**0.7920**	**0.8640**	**0.8617**
	10	ViT	0.6000	0.9050	0.7250	0.9470	0.7270	0.8180	0.7300	0.7270	0.8250	0.8090	0.8750	0.8630	0.7050	0.7760	0.7880
		ViT MultiLoss	0.6490	0.9130	0.7590	0.9570	0.7720	0.8320	0.7620	0.7440	0.8380	0.8330	0.8990	0.8630	0.7380	0.7980	0.8112
		ViT MultiScale	0.7090	0.9220	0.7880	0.9620	0.8050	0.8540	0.7840	0.7740	0.8500	0.8430	0.9140	0.8620	0.7660	0.8210	0.8324
		Proposed	**0.7230**	**0.9250**	**0.7960**	**0.9670**	**0.8200**	**0.8570**	**0.7940**	**0.7880**	**0.8550**	**0.8460**	**0.9190**	**0.8670**	**0.7770**	**0.8290**	**0.8402**

(Continued)

Table 7 (continued)

Medical Image Retrieval Task																	
Metric	**@K**	**Method**	**ATL**	**CONS**	**INF**	**PNEU**	**EDE**	**EMP**	**FIB**	**EFF**	**PNE**	**PLT**	**CARD**	**NOD**	**MAS**	**HER**	**Overall**
Recall	2	ViT	0.7300	0.9500	0.8600	0.9800	0.8700	0.9300	0.9100	0.8900	0.8700	0.8700	0.9400	0.9200	0.8100	0.9100	0.8886
		ViT MultiLoss	0.7700	0.9500	0.8600	0.9900	0.8700	0.9300	0.9300	0.8600	0.8900	0.8800	0.9500	0.9300	0.8400	0.9100	0.8971
		ViT MultiScale	0.8300	0.9500	0.8700	0.9800	0.8800	0.9400	0.9300	0.8800	0.9100	0.8800	0.9700	0.9200	0.8500	0.9200	0.9079
		Proposed	**0.8500**	**0.9500**	**0.8700**	**0.9800**	**0.8900**	**0.9400**	**0.9300**	**0.8900**	**0.9100**	**0.8800**	**0.9700**	**0.9200**	**0.8700**	**0.9200**	**0.9121**
	5	ViT	0.8800	0.9600	0.9000	0.9900	0.8900	0.9600	0.9300	0.9100	0.9200	0.8700	0.9700	0.9300	0.8700	0.9600	0.9243
		ViT MultiLoss	0.9100	0.9600	0.9000	0.9900	0.9200	0.9500	0.9400	0.9100	0.9200	0.8800	0.9700	0.9300	0.9100	0.9600	0.9321
		ViT MultiScale	0.9400	0.9700	0.9200	0.9800	0.9400	0.9700	0.9400	0.9200	0.9300	0.8800	0.9700	0.9300	0.9200	0.9600	0.9407
		Proposed	**0.9500**	**0.9800**	**0.9200**	**0.9900**	**0.9400**	**0.9600**	**0.9400**	**0.9200**	**0.9300**	**0.8800**	**0.9700**	**0.9400**	**0.9200**	**0.9600**	**0.9429**
	10	ViT	0.9500	0.9600	0.9100	0.9900	0.9500	0.9700	0.9600	0.9300	0.9400	0.9000	0.9700	0.9500	0.9100	0.9800	0.9479
		ViT MultiLoss	0.9800	0.9800	0.9300	0.9900	0.9400	0.9700	0.9600	0.9400	0.9300	0.9000	0.9700	0.9400	0.9400	0.9700	0.9529
		ViT MultiScale	0.9900	0.9800	0.9400	0.9800	0.9400	0.9700	0.9600	0.9700	0.9400	0.9000	0.9700	0.9500	**0.9500**	0.9700	0.9579
		Proposed	**0.9900**	**0.9800**	**0.9500**	**0.9900**	**0.9500**	**0.9700**	**0.9600**	**0.9700**	**0.9400**	**0.9000**	**0.9700**	**0.9500**	**0.9400**	**0.9700**	**0.9593**
F1-Score	2	ViT	0.6567	0.9167	0.7833	0.9733	0.8300	0.9167	0.8633	0.8167	0.8433	0.8433	0.9200	0.9133	0.7767	0.8367	0.8493
		ViT MultiLoss	0.7167	0.9333	0.8067	0.9767	0.8533	0.9067	0.8733	0.7967	0.8633	0.8600	0.9400	0.9167	0.7933	0.8600	0.8640
		ViT MultiScale	0.7867	0.9333	0.8300	0.9733	0.8633	0.9300	0.8967	0.8333	0.8833	0.8700	0.9533	0.9133	0.8100	0.8867	0.8831
		Proposed	**0.8000**	**0.9333**	**0.8300**	**0.9767**	**0.8733**	**0.9300**	**0.8967**	**0.8500**	**0.8833**	**0.8733**	**0.9567**	**0.9167**	**0.8233**	**0.8867**	**0.8879**
	5	ViT	0.6899	0.9228	0.7896	0.9631	0.8174	0.8929	0.8429	0.8002	0.8544	0.8377	0.9158	0.9077	0.7711	0.8346	0.8457
		ViT MultiLoss	0.7425	0.9335	0.8084	0.9681	0.8446	0.9025	0.8610	0.8115	0.8626	0.8558	0.9358	0.9035	0.8056	0.8595	0.8639
		ViT MultiScale	0.7948	0.9389	0.8363	0.9665	0.8749	0.9206	0.8739	0.8287	0.8767	0.8625	0.9442	0.9048	0.8231	0.8883	0.8810
		Proposed	**0.8129**	**0.9434**	**0.8403**	**0.9731**	**0.8788**	**0.9194**	**0.8758**	**0.8351**	**0.8767**	**0.8650**	**0.9506**	**0.9135**	**0.8245**	**0.8970**	**0.8861**
	10	ViT	0.6836	0.9255	0.7797	0.9570	0.7842	0.8620	0.7972	0.7831	0.8508	0.8354	0.8979	0.8847	0.7536	0.8393	0.8310
		ViT MultiLoss	0.7297	0.9326	0.8096	0.9626	0.8216	0.8694	0.8216	0.7979	0.8578	0.8530	0.9189	0.8868	0.7815	0.8526	0.8497
		ViT MultiScale	0.7786	0.9406	0.8329	0.9671	0.8480	0.8881	0.8378	0.8268	0.8703	0.8597	0.9300	0.8886	0.8065	0.8671	0.8673
		Proposed	**0.7886**	**0.9433**	**0.8398**	**0.9732**	**0.8597**	**0.8925**	**0.8438**	**0.8360**	**0.8734**	**0.8618**	**0.9344**	**0.8918**	**0.8150**	**0.8728**	**0.8733**

5.4 Discussion

The experimental results validate the effectiveness of our proposed method in medical image classification and retrieval tasks. The integration of multi-scale encoding enables the model to capture important features at different resolutions, which is particularly beneficial for medical images where lesions and pathologies may vary greatly in size and appearance.

The dynamic multi-loss function allows the model to balance multiple learning objectives, optimizing for both classification accuracy and retrieval effectiveness. By adjusting the loss weights during training, the model can adapt to the complexities of the data, improving generalization and robustness.

The superior performance of our method over traditional CNNs and existing transformer-based models underscores the potential of combining multi-scale processing with advanced training strategies. This approach addresses the challenges posed by complex medical datasets, such as class imbalance and high inter-class similarity.

Overall, the proposed method demonstrates significant advancements in medical image analysis, offering improved tools for clinicians in diagnosis and decision-making processes.

6 Conclusion and Future Work

This paper presents a novel multi-scale Vision Transformer (ViT) architecture with a dynamic multi-loss function for medical image classification and retrieval. By integrating multi-scale encoding, our approach effectively captures both fine-grained and global features, while the dynamic loss function adaptively balances multiple learning objectives.

Extensive experiments on the ISIC-2018 and ChestX-ray14 datasets demonstrate that our method consistently outperforms existing CNN-based and transformer-based models. The results confirm improved classification accuracy, precision, and retrieval performance, making our approach a valuable tool for medical image analysis.

Despite its effectiveness, the model's performance is influenced by dataset diversity and computational complexity. Future research could focus on extending the approach to other imaging modalities, improving optimization strategies, and enhancing model interpretability for real-world clinical adoption.

Our work contributes to advancing deep learning techniques in medical imaging, with the potential to assist healthcare professionals in more accurate diagnoses and improved patient care.

Acknowledgement: The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through small group research under grant number RGP1/278/45.

Funding Statement: This research was funded by the Deanship of Research and Graduate Studies at King Khalid University through small group research under grant number RGP1/278/45.

Author Contributions: The authors confirm their contribution to the paper as follows: Study conception and design: Omar Alqhatani, Mohamed Ghouse. Data collection: Asfia Sabahath, Omer Bin Hussain. Analysis and interpretation of results: Omar Alqhatani, Mohamed Ghouse, Asfia Sabahath. Draft manuscript preparation: Mohamed Ghouse, Arshiya Begum. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All datasets used in this study are publicly available and properly cited in the manuscript. Further details and links can be provided upon request.

Ethics Approval: This study did not involve any human or animal subjects. Hence, ethical approval is not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. arXiv:1705.02315. 2017.
2. Harley AW, Ufkes A, Derpanis KG. Evaluation of deep convolutional nets for document image classification and retrieval. arXiv:1502.07058. 2015.
3. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2015.
4. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. arXiv:1409.4842. 2014.
5. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV, USA: IEEE. p. 770–8.
6. 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. In: International Conference on Learning Representations; 2021; Vienna, Austria. [cited 2025 Jan 20]. Available from: https://arXiv.org/abs/2010.11929.
7. Pundhir A, Sagar S, Singh P, Raman B. Echoes of images: multi-loss network for image retrieval in vision transformers. Med Biolo Eng Comput. 2024;62(7):2037–58. doi:10.1007/s11517-024-03055-6.
8. Ashraf SMN, Mamun MA, Abdullah HM, Alam MGR. SynthEnsemble: a fusion of CNN, vision transformer, and hybrid models for multi-label chest X-ray classification. In: 2023 26th International Conference on Computer and Information Technology (ICCIT); 2023. p. 1–6.
9. Shen T, Li X. Automatic polyp image segmentation and cancer prediction based on deep learning. Front Oncol. 2023;12:1087438. doi:10.3389/fonc.2022.1087438.
10. Ali T, Roy PP, Saini R. Fast&Focused-Net: enhancing small object encoding with VDP layer in deep neural networks. IEEE Access. 2024;12:130603–16. doi:10.1109/ACCESS.2024.3447888.
11. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest X-ray classification. Sci Rep. 2019;9(1):6381. doi:10.1038/s41598-019-42294-8.
12. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data. 2018;5(1):180161. doi:10.1038/sdata.2018.161.
13. Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D, et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). arXiv:1902.03368. 2019.
14. Innat M, Hossain MF, Mader K, Kouzani AZ. A convolutional attention mapping deep neural network for classification and localization of cardiomegaly on chest X-rays. Sci Rep. 2023;13(1):6247. doi:10.1038/s41598-023-32611-7.
15. Alam MS, Wang D, Liao Q, Sowmya A. A multi-scale context aware attention model for medical image segmentation. IEEE J Biomed Health Inform. 2023;27(8):3731–9. doi:10.1109/JBHI.2022.3227540.
16. Nampalle KB, Pundhir A, Jupudi PR, Raman B. Towards improved U-Net for efficient skin lesion segmentation. Multimed Tools Appl. 2024;83(28):71665–82. doi:10.1007/s11042-024-18334-5.
17. Liang S, Tian S, Yu L, Kang X. Improved U-Net based on contour attention for efficient segmentation of skin lesion. Multimed Tools Appl. 2023;83(11):33371–91. doi:10.1007/s11042-023-16759-y.
18. Zhang Z, Lu B. Efficient skin lesion segmentation with boundary distillation. Med Biol Eng Comput. 2024;62(9):2703–16. doi:10.1007/s11517-024-03095-y.
19. Hao S, Wu H, Jiang Y, Ji Z, Zhao L, Liu L, et al. GSCEU-Net: an end-to-end lightweight skin lesion segmentation model with feature fusion based on U-Net enhancements. Information. 2023;14(9):486. doi:10.3390/info14090486.
20. Ali T, Siddiqui MFH, Shahab S, Roy PP. GMIF: a gated multiscale input feature fusion scheme for scene text detection. IEEE Access. 2022;10:93992–4006.
21. Rubinstein RY. The cross-entropy method for combinatorial and continuous optimization. Methodol Comput Appl Probab. 1999;1(2):127–90.

22. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015; Boston, MA, USA. p. 815–23.
23. Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06); 2006; New York City, NY, USA. vol. 2, p. 1735–42.
24. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531. 2015.
25. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; Honolulu, HI, USA: IEEE. p. 2261–9.
26. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. 2017.
27. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv:1602.07360. 2016.
28. Al-hammuri K, Gebali F, Kanan A, Thirumarai Chelvan I. Vision transformer architecture and applications in digital health: a tutorial and survey. Visual Comput Indust, Biomed Art. 2023;6(1):1–14. doi:10.1186/s42492-023-00140-9.
29. Halder A, Gharami S, Sadhu P, Singh PK, Wozniak M, Ijaz MF. Implementing vision transformer for classifying 2D biomedical images. Sci Rep. 2024;14(1):12567. doi:10.1038/s41598-024-63094-9.
30. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. UNet++: a nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. (DLMIA 2018, ML-CDS 2018); 2018; Granada, Spain. p. 3–11.
31. Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans Med Imag. 2018;37(12):2663–74. doi:10.1109/TMI.2018.2845918.
32. Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV); 2016; Stanford, CA, USA: IEEE. p. 565–71.
33. Ramadan R, Aly S. DGCU–Net: a new dual gradient-color deep convolutional neural network for efficient skin lesion segmentation. Biomed Signal Process Control. 2022;77(1):103829. doi:10.1016/j.bspc.2022.103829.
34. Malik S, Akram T, Ashraf I, Rafiullah M, Ullah M, Tanveer J. A hybrid preprocessor DE-ABC for efficient skin-lesion segmentation with improved contrast. Diagnostics. 2022;12(11):2625. doi:10.3390/diagnostics12112625.
35. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225. 2017.
36. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009; Miami, FL, USA: IEEE. p. 248–55.