MSAL-MIR: MULTI-STAGE ADAPTIVE LOSS FOR MEDICAL IMAGE RETRIEVAL

Nguyen Van Hoang Phuc, Le Quang Nhat, Duong Manh Quan, Hoang Phuong Le, Nguyen Van Hieu*

The University of Danang - University of Science and Technology, Vietnam

*Corresponding author: nvhieuqt@dut.udn.vn

(Received: April 14, 2025; Revised: June 05, 2025; Accepted: June 19, 2025)

DOI: 10.31130/ud-jst.2025.23(9C).539E

Abstract - Efficient and accurate retrieval of medical images underpins timely diagnosis and informed clinical decisions. This work introduces a novel multi-stage training paradigm designed for medical image retrieval. In the first stage, a ConvNeXt model pretrained on ImageNet is fine-tuned using Focal Loss to address class imbalance. Building on this foundation, the feature space is refined with Triplet Margin Loss, where chosen sample triplets are used to enhance discriminative learning. Our approach further streamlines the retrieval process by applying Global Max Pooling, L2 normalization, and Principal Component Analysis (PCA) for dimensionality reduction, followed by integration with Facebook AI Similarity Search (FAISS) for efficient similarity search. Experiments on the ISIC 2017 and COVID-19 chest X-ray datasets demonstrate that the proposed method achieves significant improvements in evaluation metrics, including mean Average Precision at 5(mAP@5), Precision at 1 (P@1), and Precision at 5 (P@5)

Key words - Deep Learning; Computer Vision; Medical Image Retrieval; Healthcare Applications

1. Introduction

Significant advances in machine learning, and deep learning in particular, have transformed numerous fields over the past few decades. Convolutional Neural Networks (CNN) first introduced in the late 1970s [1], and the first successful real-world application in hand-written digit recognition appearing in 1998 [2]. Studies such as [3] and [4] applied Deep Belief Networks and Stacked Autoencoders to classify patients with Alzheimer's disease based on brain Magnetic Resonance Imaging (MRI). Another study [5] identified anatomical landmarks on the surface of the distal femur by processing three independent sets of 2D MRI slices.

Medical image retrieval has also become an important area of research. Retrieval methods have shown their potential in supporting diagnosis and treatment, helping specialists more easily identify objects in medical images. This not only saves time, but also improves accuracy in disease detection, reduces errors, and helps in clinical decision making, such as the studies by Anavi et al. [6] and Liu et al. [7], who applied their methods to X-ray image databases. Although traditional methods have achieved good performance in specific medical scenarios, they often do not fully leverage the information on the label during training. This leads to a lack of effective utilization of unlabeled data, which contributes to the reduced performance in medical image retrieval.

To solve these challenges, this study introduces the following methods:

Multi-Stage Training for Enhanced Feature Learning: Based on a pretrained ConvNeXt model, we proposed an approach that applies Focal Loss to accurately identify similar and dissimilar labels in the case of imbalanced data. Subsequently, sample pairs selected based on these labels are trained using Triplet Margin Loss, which enhances the feature space separability and optimizes the ability to discriminate between classes.

An efficient Image Retrieval system: The image retrieval system is designed so that the embedded vectors are preprocessed to enhance the retrieval efficiency.

Experiments on the evaluated datasets: Two medical datasets, ISIC 2017 for skin lesions and COVID-19 chest X-rays were used to demonstrate performance.

2. Related Works

In this section, we organize the relevant work into the following key areas:

2.1. Image retrieval systems

The basic block diagram of an image retrieval system is illustrated in Figure 1. In the retrieval process, images are fetched from large-scale databases based on feature representations extracted from the image content. Any retrieval system typically consists of two stages: the offline stage and the online stage. In the offline stage, features are extracted from large image collections (used to train the system) to build a local feature database. In the online stage, similar features are extracted from the query image, and a distance metric is computed between the features of the query image and those of the database images to assess similarity. The images with high or low similarity scores are then presented to the user as retrieval results with query labels, allowing the model to learn general semantic features through classification labels while simultaneously optimizing bedding space via metric learning to enhance instance-level discrimination.

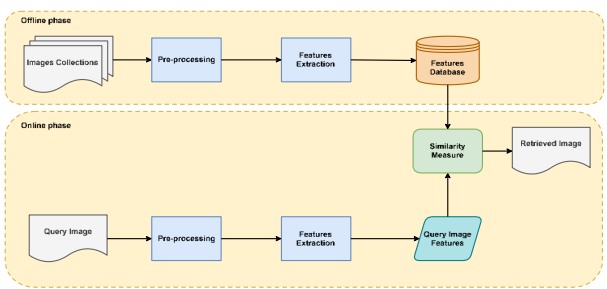


Figure 1. *Image retrieval system block diagram*

2.2. Deep Learning-Based Image Retrieval Methods

The advancement of deep learning has led to breakthroughs in automatic learning feature representations from image data. In the context of deep learning-based image retrieval, two main training approaches are commonly classified as follows:

2.2.1. Supervised Learning Using Class Labels

Convolutional neural networks (CNNs) trained for image classification - such as disease recognition or analysis of anatomical structures - typically employ the cross-entropy loss function to optimize class discrimination. Thanks to their ability to automatically extract features, CNNs have become a crucial tool in many computer vision applications.

Popular architectures such as ResNet [8, 9], which utilize skip connections to mitigate gradient vanishing, DenseNet [10, 11], which leverages dense connections between layers for more effective information flow, and ConvNeXt [12, 13], which incorporates improvements inspired by Transformers, have demonstrated superior performance in label-based image classification tasks. Furthermore, the application of the Vision Transformer (ViT) architecture [14] to classification tasks has shown significant potential, particularly when pretrained on large-scale datasets like ImageNet, providing general semantic features that can be fine-tuned for specific tasks. The advancement of deep learning models has opened up new directions, especially in medical image analysis and object recognition applications.

2.2.2. Query Label-Based Approach

Instead of focusing on classification, distance-based learning techniques such as Triplet Loss and Contrastive Loss are employed to learn an embedding space where samples of the same class are pulled closer together, while samples of different classes are pushed farther apart. This approach is particularly useful in image retrieval, face recognition, and data clustering tasks.

Triplet Loss [15] optimizes the model by learning from three samples: an anchor, a positive (same class), and a negative (different class). Its objective is to ensure that the distance between the anchor and the positive is smaller than the distance to the negative by at least a predefined margin. In contrast, Contrastive Loss [16, 17] It minimizes the distance between positive pairs while ensuring that negative pairs are separated by a margin.

2.2.3. Hybrid Approach Combining Classification and Query Labels

While individual training approaches each have their advantages, several recent studies have proposed combining classification labels with query labels, allowing the model to learn general semantic features through classification labels while simultaneously optimizing the embedding space via metric learning to enhance instance-level discrimination.

Several studies have proposed hybrid loss functions to simultaneously optimize both classification and representation capabilities for retrieval tasks. For example, Histogram Loss by Ustinova and Lempitsky [19] leverages the distribution of distances in the embedding space, while Multi-Similarity Loss by Wang et al. [20] exploits complex relationships among sample pairs to enhance representation learning. In addition, Center Loss by Wen et al. [21] is considered an effective approach that combines classification loss (cross-entropy) with a loss function that optimizes the distance in the embedding space. However, these methods are applied to natural image data and focus mainly on optimizing the embedding distance, with limited direct classification capability.

In this study, we propose a hybrid training approach for medical image retrieval that combines Focal Loss and Triplet Margin Loss. Focal Loss addresses class imbalance by focusing on hard samples, while Triplet Margin Loss promotes a discriminative embedding space. Designed specifically for medical data, our method enhances diagnostic support by improving retrieval accuracy and precision

2.2.4. Modern Retrieval Methods and the Application of FAISS

The significant increase in data volume and the dimensionality of representation vectors has posed considerable challenges in performing efficient retrieval in the embedding space. Modern retrieval solutions based on the Approximate Nearest Neighbor (ANN) algorithm have been developed to address this issue.

FAISS is an open-source library specifically designed for searching similar vectors in large-scale datasets. It applies techniques such as Flat Inner Product (flatIP), Product Quantization, and other vector compression strategies to optimize retrieval speed and accuracy [22].

These modern retrieval methods enable the system to perform fast searches in the embedding space, meeting the high speed and accuracy requirements in medical applications, where diagnostic time is critical.

3. Approaches

We also implement performance improvements in both stages of the image retrieval system, achieving high accuracy on the evaluate dataset and fast retrieval times.

3.1. The proposed pipeline

Medical image retrieval poses unique challenges compared to conventional image retrieval tasks, particularly due to class imbalance and the high variability of image features within the same class. To address these issues, the proposed method employs a multi-stage fine-tuning pipeline based on the ConvNeXt architecture-a vision model pretrained on ImageNet [12]. This approach is designed to adapt the model to medical image datasets while integrating advanced loss functions to improve both classification accuracy and retrieval performance.

In this study, the multi-stage training pipeline is implemented as illustrated in Figure 2, with the goal of optimizing medical image retrieval performance. The fine-tuning process is divided into three stages, each employing specific strategies, as detailed below:

Stage 1: Fine-tuning on medical data using Focal Loss;

Stage 2: Contrastive learning with a combination of Focal Loss and Triplet Margin Loss;

Stage 3: Hard negative mining and additional training.

The following sections will present a detailed description of the data generation process, the loss function formulations, and the training algorithm for each stage.

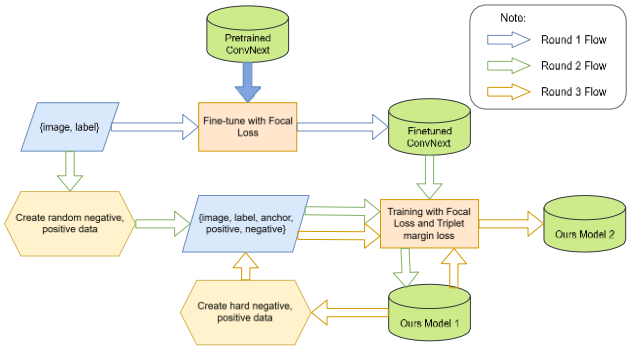


Figure 2. *Multi-stage training pipeline*

3.1.1. Stage 1: Training on medical data with Focal Loss

In the first stage of the process, the objective is to adapt a model pretrained on ImageNet to the specific characteristics of medical data, which often suffer from class imbalance due to the low prevalence of rare conditions. To address this issue, Focal Loss [23] as defined in (1) is applied to automatically adjust the weighting of samples based on the difficulty of their prediction. Specifically, Focal Loss is defined by the following formula:

$$FL(p_t) = -\alpha\,(1 - p_t)^{\gamma} log(p_t) \quad (1)$$

Where p_t is the predicted probability for the correct class, α is a parameter that balances the classes, and γ is a focusing parameter that increases the penalty for hard-to-classify samples. The term $(1 - p_t)^{\gamma}$ helps increase the contribution of the loss from misclassified samples, this formulation is flexible, allowing adjustments to the parameters α and γ.

Its flexibility in adjusting the parameters α and γ, Focal Loss improves the model's ability to detect rare conditions while maintaining strong overall accuracy. Thus, the fine-tuning stage with Focal Loss allows the model to effectively adapt to medical data, providing a solid foundation for the subsequent retrieval stages.

3.1.2. Stage 2: Contrastive learning with Focal Loss and Triplet Margin Loss

After Stage 1 establishes the basic classification performance, Stage 2 aims to enhance the embedding space to improve image retrieval effectiveness. The proposed method combines Focal Loss – ensuring classification performance – with Triplet Margin Loss (2) [15] to learn representations such that the embedding vectors of images from the same class are pulled closer together, while those from different classes are pushed farther apart.

The formula for Triplet Margin Loss is defined as follows:

$$L_{triplet} = max(d(a,p) - d(a,n) + m, 0) \quad (2)$$

where $d(a, b)$ is the distance function (commonly the Euclidean distance) between the embedding vectors of images a and b; a is the query image (anchor), p is a same-class image (positive sample), n is a different-class image (negative sample), and m is the margin value that ensures the distance between $d(a, p)$ and $d(a, n)$ reaches a certain minimum threshold.

The sample selection process plays a crucial role in triplet-based learning. Positive samples are selected from the training set as images belonging to the same disease category as the query image, while negative samples are drawn from images of different disease categories, ensuring that their similarity to the query image is sufficiently high to present a challenge to the model.

The advantage of Triplet Margin Loss lies in its ability to enhance discriminability in the embedding space: the model is encouraged to pull together the vectors of images from the same class and push apart those from different classes, thereby producing more distinctive feature representations. To balance the tasks of classification and representation learning, the overall loss function (3) is defined as a combination of Focal Loss and Triplet Margin Loss:

$$L_{total} = \lambda_1 \cdot FL + \lambda_2 \cdot L_{triplet} \quad (3)$$

where λ_1 and λ_2 are weighting parameters that control the importance of the two tasks.

3.1.3. Stage 3: Hard negative sample generation and additional training

To enhance the model's classification capability, the third stage focuses on leveraging hard negative samples-images from different classes that exhibit high similarity to the query image-to enrich the information in the embedding space. Initially, the model from Stage 2 is used to evaluate similarity scores within the training set, identifying the top 20 most similar images for each query image. Among these, those that have different labels from the query and are incorrectly predicted are selected as hard negative samples. These samples are then incorporated into the additional training process, together with randomly selected negative samples from the previous stage. If the number of hard samples exceeds a predefined threshold, they replace the randomly selected negatives. Otherwise, if hard samples are limited, the sampling strategy from the previous stage is maintained.

To prevent the model from becoming "over-focused" on difficult cases, the additional training phase is conducted with a reduced learning rate, ensuring that the model does not overfit to hard negative samples while still maintaining its ability to recognize basic patterns. The loss function in this stage is defined in (4) as follows:

$$L_{hard-negative} = \lambda_3 \cdot FL_{hard} + \lambda_4 \cdot L_{triplet-hard} \quad (4)$$

where λ_3 and λ_4 are weighting parameters adjusted to fit the characteristics of the hard negative sample set. This method significantly enhances the model's ability to handle challenging cases, thereby improving the overall effectiveness of the image retrieval system.

3.2. Retrieval Method

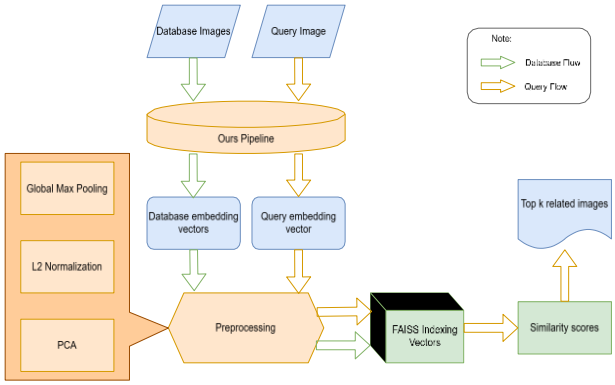


Figure 3. *Image retrieval approach*

This section proposes a novel method (Figure 3) to enhance the efficiency of medical image retrieval, with FAISS (FlatIP) serving as the central component. According to Johnson et al. [22], FAISS is an optimized and highly scalable nearest neighbor search library, particularly well-suited for large datasets in high-dimensional spaces.

The proposed method focuses on improving the pre-processing and feature representation process, inspired by previous research [15, 24]. This method introduces an embedding strategy combined with dimensionality reduction techniques to create compact and powerful representations for medical image retrieval tasks. Specifically, based on image retrieval studies. The proposed method is designed to meet the specific needs of medical image data, with the following key improvements:

- *Representation Refinement with GMP and PCA:* The embedding process incorporates Global Max Pooling followed by Principal Component Analysis to obtain lower-dimensional feature vectors. This reduces computational cost while maintaining essential information for effective image retrieval.

- *Optimization using Normalized Indexing:* The method leverages normalized embedding vectors within a FAISS FlatIP index structure. This configuration improves retrieval precision and ensures compatibility with large-scale data scenarios.

Embedding and Vector Retrieval Process: To achieve a discriminative and compact vector representation for input images, the process is systematically built through a combination of preprocessing techniques, dimensionality reduction, and retrieval using FAISS.

Step 1: Feature extraction using the pretrained ConvNeXt model. The input image is passed through the ConvNeXt model – a deep learning architecture pre-trained on ImageNet and optimized for large-scale classification tasks [12]. This model generates feature representations with high dimensionality, capturing important hierarchical spatial features crucial for medical image retrieval applications. Specifically, for each image, the model produces a feature map $E \in R^{1536x7x7}$ where:

$-$ 1536 is the number of channels in the final convolutional layer,

- 7 x 7 is the spatial size of the feature map.

The matrix E is considered as the input for the subsequent processing steps.

Step 2: Apply Global Max Pooling for dimensionality reduction. To convert the feature matrix E into a one-dimensional compact vector, the process uses the Global Max Pooling operation, as defined in (5) by the following formula:

$$v = GlobalMaxPooling(E), v \in R^{1536} \quad (5)$$

Applying Global Max Pooling allows retaining the maximum value from each channel, thereby reducing the dimensionality of the representation while preserving important features, and enhancing the stability of the representation against spatial variations.

Step 3: L2 Normalization. To enhance the comparability of embedding vectors, L2 normalization is applied to the vector v, as defined in (6):

$$\hat{v} = \frac{v}{\|v\|_2}, \|v\|_2 = 1. \quad (6)$$

This normalization ensures that the embedding vectors have a unit norm, so that similarity measures (e.g., cosine similarity) can accurately reflect the angular relationship between vectors, thereby enhancing stability and consistency during the search process.

Step 4: Dimensionality Reduction with Principal Component Analysis (PCA)

Although the embedding vector vˆ possesses strong discriminative capability, its original dimensionality of 1536 can lead to limitations in memory usage and retrieval efficiency. Therefore, Principal Component Analysis (7) (PCA) is applied to project these vectors into a lower-dimensional space:

$$\hat{v}_{PCA} = PCA(\hat{v}), \hat{v}_{PCA} \in R^{128} \quad (7)$$

In this process, PCA is trained offline on the entire dataset to identify the principal components that capture the maximum variance, with the number of components retained such that over 95% of the variance is preserved. This dimensionality reduction not only conserves computational resources and memory, but also enables the system to scale retrieval operations on large datasets while maintaining high accuracy.

Step 5: Nearest Neighbor Search with FAISS.

After dimensionality reduction, the embedding vectors $\hat{v}_{PCA}$ are indexed using FAISS with the Inner Product Space configuration. The similarity between the query vector q and a database vector xi is computed as follows:

$$Similarity(q, x_i) = q \cdot x_i. \quad (7')$$

In this configuration, the use of normalized embeddings ensures that the similarity measure effectively reflects the cosine distance between vectors. FAISS, optimized for GPU execution, enables fast indexing and retrieval even with large datasets, while also ensuring scalability as both the number of samples and the dimensionality of embedding vectors increase. This allows the system to maintain stable performance and high accuracy in large-scale similarity search tasks.

4. Experiments

4.1. Dataset

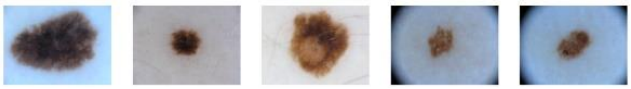
The International Skin Imaging Collaboration (ISIC) 2017 dataset consists of dermoscopic images for classifying skin lesions. We follow the same training and test split as in the X-MIR experiment, using 2,000 training images and 270 test images, annotated with three classes: melanoma, nevus, and seborrheic keratosis. Sample images from the dataset are shown in Figure 4.

Table 1. *Data statistics for the ISIC 2017 and COVID-19 chest X-ray training datasets*

Dataset	Labels	Count	Total Images
ISIC 2017	Nevus	1372	2000
	Seborrheic Ker-atosis	254	
	Melanoma	374	
COVID-19 Chest X-ray	Normal	8751	19364
	Pneumonia	5964	
	COVID-19	4649	

Table 2. *Data statistics for the ISIC 2017 and COVID-19 chest X-ray test datasets.*

Dataset	Labels	Count	Total Images
ISIC 2017	Nevus	90	270
	Seborrheic Keratosis	90	
	Melanoma	90	
COVID-19 Chest X-ray	Normal	100	300
	Pneumonia	100	
	COVID-19	100	



(a) Nevus

(b) Seborrheic Keratosis

(c) Melanoma

Figure 4. *Sample images from the ISIC 2017 dataset.*

(a) Normal

(b) Pneumonia

(c) COVID-19

Figure 5. *Sample images from the COVID-19 chest X-ray dataset*

The COVID-19 chest X-ray dataset is a comprehensive collection of labeled chest X-ray images for detecting COVID-19. This dataset differs from previous versions, and in our experiment, we use 19,364 training images and 300 test images, distributed across three classes: normal, pneumonia, and COVID-19. Sample images from this dataset are shown in Figure 5.

4.2. Data Preprocessing

In both datasets, all images are resized to 224 x 224 pixels. For training images, data augmentation methods include random horizontal flipping and rotation (±30 degrees). Finally, normalization is applied using the mean and standard deviation values of the ImageNet dataset.

4.3. Training Process

The training process is divided into 3 phases with different configuration strategies. In Phase 1, the model is trained for 10 epochs using the Focal Loss function (alpha = 1, gamma = 2) and the Adam optimizer with *learning rate 1e-4* and *weight decay 1e-5*. Furthermore, learning rate adjustment is performed using StepLR with a step size of 5 and a decay factor of 0.5, combined with the gradient clipping technique with a max norm threshold of 2.0 to stabilize the training process.

Phase 2 for 10 training epochs, during which the configuration of the Adam optimizer remains the same as in Phase 1. However, the training strategy in this phase is enhanced by combining the Focal Loss function with the Triplet Margin Loss (with $margin = 1.0$), and adding a Global Max Pooling layer to process the feature embedding, aiming to optimize the model's feature extraction capability. In this phase, the loss weights are set as $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ to balance the contributions of the Focal Loss and Triplet Loss.

In Phase 3, model fine-tuning is performed over 8 epochs with the learning rate reduced to 2e-5, while the model architecture and loss function remain the same as in Phase 2. This allows the training process to achieve a balance between learning speed and accuracy, helping to improve the overall performance of the system. Here, $\lambda_3 = 0.4$ and $\lambda_4 = 0.6$ are used to emphasize the Triplet Loss component, as this phase focuses on learning from harder examples.

Techniques to stabilize the training process include gradient clipping, weight decay, and learning rate scheduling. The model is saved at the epoch with the best performance on the validation set.

4.4. Implementation Details

All experiments conducted on PyTorch version 2.2.2 with Python 3.10.11, in an Ubuntu 22.04.3 LTS environment. Training and evaluation were performed on an NVIDIA A100 GPU (CUDA 12.6). All results were repeated three times, and the average is reported.

4.5. Baselines

We compare our proposed method against three well-established retrieval baselines:

- DELG [25]: Deep Local and Global Features model, which extracts and aggregates both local and global image descriptors for retrieval.

- X-MIR with DenseNet [27]: A DenseNet-based retrieval framework that leverages convolutional feature maps for matching.

None of these baseline methods employs any explicit class-imbalance mitigation techniques (e.g., weighted loss, oversampling, or under-sampling), allowing us to isolate and demonstrate the effectiveness of our imbalance-aware strategy.

4.6. Evaluation

To evaluate the effectiveness of the proposed method, we conducted comprehensive experiments on two widely used datasets: **ISIC 2017** and **COVID-19 chest X-ray**. The performance of our method is compared with the results obtained from the above baselines. Evaluation metrics include Mean Average Precision at (*mAP@5*), Precision at (*P@1*), and Precision at (*P@5*).

Mean Average Precision (mAP@k) Mean Average Precision (mAP) at k provides a general evaluation of the ranking ability of the system in the top k. First, we compute AP@k (Average Precision at k) for each query:

$$AP@k = \frac{1}{m} \sum_{j=1}^{k} P(j) \cdot rel(j) \quad (8)$$

Where:

m: Number of relevant samples for a single query.

$P(j)$: Precision at rank j, calculated as:

$$P(j) = \frac{\text{Number of relevant samples in top } j}{j}$$

$rel(j)$ Indicator function, where:

$$rel(j) = \begin{cases} 1, \text{if the sample at rank j is relevant} \\ 0, \text{otherwise} \end{cases}$$

Then, the mean Average Precision at rank k ($mAP@k$) is computed by averaging $AP@k$ over all N queries:

$$mAP@k = \frac{1}{N} \sum_{i=1}^{N} AP@k_i \quad (9)$$

Where: N: Total number of queries; $AP@k_i$: Average precision at rank k for the i-th query.

Precision at k (P@k)

Precision at rank k (P@k) measures the proportion of relevant samples among the top k retrieved results: This metric helps evaluate the accuracy of the system when users only consider a limited number of results (top k), such as the top 5 or top 10 results.

$$P@k = \frac{\text{Number of relevant samples in the top } k}{k} \quad (10)$$

Experimental results show that on ISIC 2017, the ConvNeXt model improved its accuracy from *73.7%* to *75.4%* after three training phases, while precision and recall reached 76.9% and 75.1%, respectively, in the third phase. On the COVID-19 chest X-ray dataset, the model improved accuracy from 93.3% to 94.0% and recall from 92.3% to 94%. In comparison, with ViT-B-32, the highest accuracy for ISIC 2017 (70.7%) and COVID-19 chest X-ray (89.7%) was achieved in Phase 2.

Regarding image retrieval performance, as shown in Table 5, the application of techniques such as Global Max Pooling (GMP), L2 normalization, and dimensionality reduction using PCA (Principal Component Analysis) significantly improved the model's performance. For ISIC 2017, the mAP@5 increased from 60.3% with a standard flattened vector to 71.4% when applying the full GMP + L2 + PCA processing pipeline. Similarly, the P@1 and P@5 metrics also showed noticeable improvements, reaching 77.4% and 74.7%, respectively. On the COVID-19 chest X-ray dataset, the improvement was even more remarkable, with mAP@5 reaching 93.5% and P@1 reaching 94.3%.

Table 3. *Classification performance of ConvNeXt over three rounds on ISIC 2017 and COVID-19 datasets*

Dataset	Metric	Round 1	Round 2	Round 3
ISIC 2017	Accuracy	73.7	75.2	**75.4**
	Precision	74.3	76.4	**76.9**
	Recall	72.2	74.5	**75.1**
COVID-19 Chest X-ray	Accuracy	93.3	93.0	**94.0**
	Precision	94.9	94.7	**95.1**
	Recall	92.3	92.0	**93.0**

Table 4. *Classification performance of ViT-B-32 over three rounds on ISIC 2017 and COVID-19 datasets*

Dataset	Metric	Round 1	Round 2	Round 3
ISIC 2017	Accuracy	68.5	70.7	70.2
	Precision	69.1	71.3	70.8
	Recall	67.4	69.9	69.3
COVID-19 Chest X-ray	Accuracy	88.2	89.7	89.5
	Precision	89.1	90.4	90.1
	Recall	87.5	88.9	88.6

Table 5. *Retrieval performance of ConvNeXt under different index preprocessing steps*

Dataset	Metric	Flat Features Vector	GMP (Max Pooling)	GMP+ L2 Normalization	GMP + L2 + PCA
ISIC 2017	mAP@5	60.3	63.9	68.3	**71.4**
	P@1	63.8	71.5	72.6	**77.4**
	P@5	66.4	68.5	72.3	**74.7**
COVID-19 Chest X-ray	mAP@5	81.4	88.3	89.7	**93.5**
	P@1	85.3	91.0	93.3	**94.3**
	P@5	84.2	90.9	91.3	**94.1**

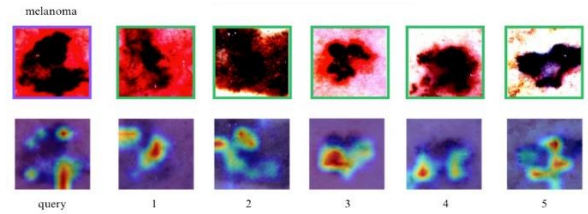
Table 6. *Method comparison on ISIC 2017 and COVID-19 datasets*

Dataset	Metric	X-MIR	DELG	ViT-B-32 (Ours)	ConvNeXt (Ours)
ISIC 017	mAP@5	61.6	58.7	64.2	**71.4**
	P@1	69.6	66.8	67.8	**77.4**
	P@5	69.2	61.8	65.5	**74.7**
COVID-19 Chest X-ray	mAP@5	89.4	79.3	90.1	**93.5**
	P@1	90.8	83.7	92.2	**94.3**
	P@5	90.3	80.4	89.0	**94.1**

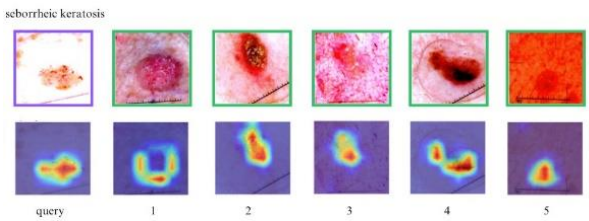
Table 6 compares the performance of the proposed method with other state-of-the-art models such as X-MIR, ViT-B-32, and DELG on the same two datasets. The results demonstrate that our method outperforms all others across every metric. Specifically, on ISIC 2017, mAP@5 reaches 71.4%, significantly higher than ViT-B-32 (64.2%) and DELG (58.7%). At the same time, P@1

and P@5 also achieve the highest values of 77.4% and 74.7%, respectively. Similarly, on the COVID-19 chest X-ray dataset, our method continues to lead with a mAP@5 of 93.5%, surpassing ViT-B-32 (90.1%) and X-MIR (89.4%).

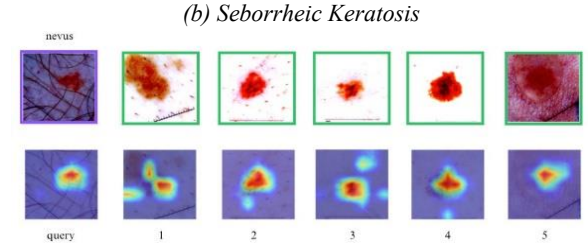
4.7. Grad-CAM Visualization and Inference Results



(a) Melanoma

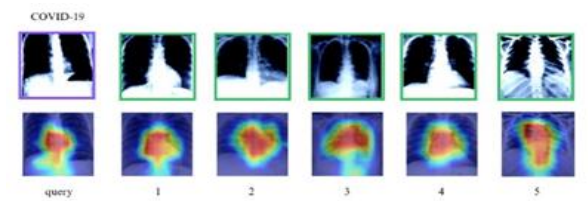


(b) Seborrheic Keratosis

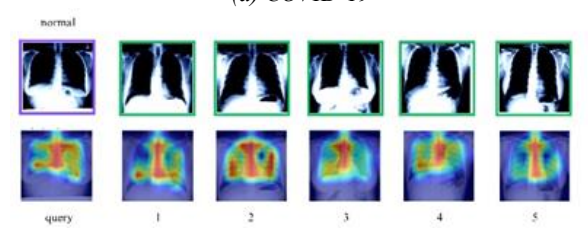


(c) Nevus

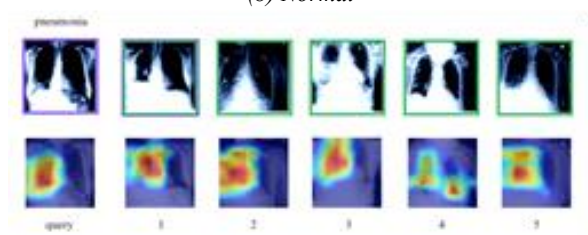
Figure 6. *Examples of inference results and Grad-CAM visualizations on the ISIC 2017 dataset*



(a) COVID-19



(b) Normal



(c) Pneumonia

Figure 7. *Examples of inference results and Grad-CAM visualizations on the COVID-19 chest X-ray dataset*

We demonstrate the interpretability of model using Grad-CAM. Figures 6 and 7, we present both the original input images and their corresponding Grad-CAM for representative cases of melanoma, seborrheic keratosis, and nevus from the ISIC 2017 dataset, as well as COVID-19 and normal chest X-ray images from the COVID-19 dataset.

5. Conclusions

In this paper, we propose a multi-stage training pipeline designed to enhance the performance of medical image retrieval systems. By leveraging the ConvNeXt model, fine-tuned on medical data with advanced loss functions such as Focal Loss and Triplet Margin Loss, our approach achieves higher accuracy and retrieval performance compared to traditional methods. Furthermore, the integration of Principal Component Analysis (PCA) for dimensionality reduction and the use of FAISS for efficient similarity search further enhance the retrieval process, making it both scalable and highly effective.

Our experiments demonstrate the proposed method in addressing the unique challenges of medical image datasets, such as class imbalance and high intra-class variability. The combination of these techniques enables accurate and efficient retrieval of relevant medical images-an essential aspect for real-world healthcare applications.

REFERENCES

[1] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”, *Biological Cybernetics*, vol. 36, pp. 193–202, 1980, doi:. 10.1007/BF00344251.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition”, *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, doi: 10.1109/5.726791.

[3] T. Brosch and R. Tam, for the Alzheimer’s Disease Neuroimaging Initiative, “Manifold learning of brain MRIs by deep learning”, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, Berlin, Germany: Springer, 2013, vol. 8150, pp. 629–636, doi: 10.1007/978-3-642-40763-5_78.

[4] S. M. Plis *et al*., “Deep learning for neuroimaging: a validation study”, *Frontiers in Neuroscience*, vol. 8, p. 229, 2014, doi: 10.3389/fnins.2014.00229.

[5] D. Yang, S. Zhang, Z. Yan, C. Tan, K. Li, and D. Metaxas, “Automated anatomical landmark detection on distal femur surface using convolutional neural network”, in *Proc. IEEE 12th Int. Symp. Biomedical Imaging (ISBI)*, Brooklyn, NY, USA, 2015, pp. 17–21, doi: 10.1109/ISBI.2015.7163806.

[6] Y. Anavi, I. Kogan, E. Gelbart, O. Geva, and H. Greenspan, “Visualizing and enhancing a deep learning framework using patients’ age and gender for chest x-ray image retrieval”, in *Proc. SPIE Medical Imaging: Computer-Aided Diagnosis*, 2016, vol. 9785, p. 978510, doi: 10.1117/12.2217587.

[7] X. Liu, H. R. Tizhoosh, and J. Kofman, “Generating binary tags for fast medical image retrieval based on convolutional nets and Radon Transform”, in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Vancouver, Canada, 2016, pp. 2872–2878, doi: 10.1109/IJCNN.2016.7727562.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition”, *arXiv preprint arXiv:* 1512.03385, 2016

[9] N. Tajbakhsh *et al*., “Convolutional neural networks for medical image analysis: Full training or fine tuning?” *IEEE Trans. Medical Imaging*, vol. 35, no. 5, pp. 1299-1312, 2016.

[10] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), arXiv preprint arXiv*:1608.06993, 2017.

[11] H. C. Shin et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning", *IEEE Trans. Medical Imaging*, vol. 35, no. 5, pp. 1285-1298, 2016

[12] Z. Liu *et al*., "A ConvNet for the 2020s", *arXiv preprint arXiv:*2201.03545, 2022

[13] A. Esteva *et al*., "Dermatologist-level classification of skin cancer with deep neural networks", *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[14] A. Dosovitskiy *et al*., "An image is worth 16x16 words: Transformers for image recognition at scale", *arXiv preprint arXiv:* 2010.11929., 2020.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering", *arXiv preprint arXiv:*1503.03832, 2015.

[16] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping", in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1735–1742.

[17] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep ranking for image similarity learning", *arXiv preprint arXiv:*1405.0301, 2014.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection", *arXiv preprint arXiv:* 1708.02002, 2017.

[19] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss", *arXiv preprint arXiv:*1611.00822., 2016.

[20] X. Wang *et al*., "Multi-similarity loss with general pair weighting for deep metric learning", *arXiv preprint arXiv:* 1904.06657, 2019.

[21] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition", in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2016, pp. 499-515.

[22] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs", *arXiv preprint arXiv*: 1702.08734, 2017.

[23] T.-Y. Lin *et al*., "Focal loss for dense object detection", *arXiv preprint arXiv:* 1708.02002, 2017.

[24] J. Deng *et al.,* "ArcFace: Additive angular margin loss for deep face recognition", in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] B. Cao, A. Araujo, and J. Sim, "DELG: Deep local and global features for image retrieval", *arXiv preprint arXiv:* 2001.05027, 2020.

[26] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale", *arXiv preprint arXiv:*2010.11929., 2020.

[27] B. Hu, B. Vasu, and A. Hoogs, "*X-MIR: Explainable medical image retrieval*", 2*022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV),* 2022. Pp. 440-450.