**Medical Image Classification With Self-Supervised Learning**

Daniel Maged, Mazen Khaled, Youssef Alghazali, Mohamed Sherif and Mohamed Osama

Department of Ai

Nile University

***Abstract***

This project investigates the application of self-supervised learning (SSL) techniques for medical image classification, with a specific focus on chest radiographs. The main objective is to explore how various SSL models—Masked AutoEncoders (MAE), SimCLR, MoCo, BYOL and DINO—can be pre-trained on large-scale unlabeled datasets and fine-tuned to improve performance on smaller labeled datasets. By leveraging the NIH ChestX-ray14 dataset for unsupervised representation learning and the RSNA Pneumonia Detection Challenge dataset for downstream supervised classification, this study demonstrates the potential of SSL to reduce reliance on costly manual annotations while maintaining high diagnostic accuracy. Each team member implemented a different SSL technique to evaluate and compare their effectiveness for pneumonia detection. This collaborative approach provides insight into the practical benefits and trade-offs of different self-supervised paradigms in the context of medical imaging.

***Keywords*:** Self-Supervised Learning, Medical Imaging, Pneumonia Detection, ChestX-ray14, RSNA Dataset, Transfer Learning

**1.1 Background**

Medical image classification plays a vital role in assisting clinicians with early and accurate diagnosis of critical conditions, such as pneumonia. Traditional supervised learning methods require large amounts of labeled data, which is often difficult and expensive to obtain in medical domains due to the need for expert annotation. This limitation poses a significant challenge in developing scalable and generalizable models.

Recent advances in self-supervised learning (SSL) have made it possible to leverage large amounts of unlabeled data to learn high-quality representations. These techniques allow models to perform meaningful pretext tasks that do not require labels and transfer the learned knowledge to downstream tasks like disease classification. In the medical domain, this reduces the need for extensive manual labeling while maintaining competitive accuracy.

This project investigates the use of four prominent SSL methods—SimCLR, MoCo, BYOL, and MAE—applied to the NIH ChestX-ray14 dataset for pretraining. The models are then fine-tuned and evaluated on the RSNA Pneumonia Detection Challenge dataset to assess their diagnostic performance in classifying pneumonia from chest X-rays.

.

**1.2 Research Objective**

This study aims to:

* Evaluate and compare the performance of different self-supervised learning models (SimCLR, MoCo, BYOL, MAE and DINO) on chest radiograph classification tasks.
* Pre-train models using the large, unlabeled NIH ChestX-ray14 dataset and fine-tune them using the labeled RSNA Pneumonia Detection dataset.
* Analyze how each SSL method transfers learned representations to improve classification accuracy with limited supervision.
* Identify which SSL technique offers the most promising trade-off between performance and resource efficiency for medical image analysis

**2. Methodology**

**2.1MAE:**

In this approach, the Masked Autoencoder (MAE) technique was used for self-supervised pretraining. A Vision Transformer (ViT-B) model was employed as the encoder. From 60,000 unlabeled chest X-rays in the NIH dataset, 75% of the image patches were masked randomly, and the model was trained to reconstruct the missing patches using the remaining 25%. The decoder consisted of two fully connected layers, and the model was optimized with Mean Squared Error (MSE) loss over 5 epochs.

After pretraining, the decoder was replaced with a classification head for binary classification (Normal vs. Pneumonia). The pretrained encoder weights were frozen, and the classification head was fine-tuned using labeled images from the RSNA dataset. The training used the AdamW optimizer and CrossEntropyLoss.

**2.2BYOL:**

This approach used Bootstrap Your Own Latent (BYOL) for self-supervised pretraining. A ResNet-18 encoder was employed, with its final classification layer removed. The BYOL architecture included two branches:

* Online branch: with encoder, projector, and predictor
* Target branch: with frozen encoder and projector

Two augmented views of the same image were passed through both branches, and the cosine similarity between online and target features was minimized. The model was trained for 5 epochs using the NIH ChestX-ray14 dataset. The best pretraining result was achieved in epoch 3 with an average BYOL loss of 0.6977.

For fine-tuning, the encoder was frozen and a classification head (two fully connected layers) was added for pneumonia classification. The model was trained using the RSNA labeled dataset for 20 epochs with CrossEntropyLoss and Adam optimizer (lr=0.001). Preprocessing included resizing, grayscale conversion, and normalization.

**2.3 DINO**

(Self-Distillation with No Labels) framework. The method relies on a student-teacher setup: the student network learns to match the output of a momentum-updated teacher network across different augmented views. The student’s weights are updated by gradient descent, while the teacher’s weights follow an exponential moving average of the student.

The model was first trained on a small sample of 10,000 chest X-ray images using the ViT-Tiny model. Later, a larger ViT-Small model was trained on a larger subset to balance capacity and efficiency. CLAHE and bilateral filters were applied as preprocessing.

Pretraining used softmax cross-entropy between the teacher and student outputs. Mixed precision training and optimized data loading improved speed and memory efficiency.

The fine-tuning phase used a new classification head trained on the RSNA dataset with frozen encoder weights. The classifier was trained using CrossEntropyLoss and the Adam optimizer.

**2.4 SimCLR**

The model development followed a two-stage approach combining self-supervised pretraining and supervised fine-tuning. In the first stage, we pretrained a ResNet-18 encoder using the SimCLR framework on unlabeled chest X-ray images. The pretraining used the NT-Xent contrastive loss with a temperature parameter set to 0.2, which helps control how strongly the model distinguishes between similar and dissimilar image features. We trained the model using the Adam optimizer with a learning rate of 3e-3 and employed cosine annealing learning rate scheduling over 5 cycles to help the model converge smoothly. To maintain training stability with mixed precision, we used gradient scaling.

For the second stage, we adapted the pretrained model for pneumonia detection by adding a simple linear classifier while keeping the encoder weights frozen. This transfer learning approach used class-weighted cross-entropy loss to handle the imbalanced dataset, with the Adam optimizer starting at a learning rate of 1e-3 and reducing it by a factor of 10 every 5 epochs. The model trained for 10 epochs total, achieving 76.52% accuracy on the test set.

**2.5 MoCo**

This study focused on binary classification of chest radiographs using the NIH Chest X-ray dataset, specifically distinguishing between Pneumonia and No Finding. The initial dataset was filtered to exclude missing image files and retain only the relevant diagnostic labels. Labels were encoded as binary classes: 1 for Pneumonia and 0 for No Finding. The dataset was divided into training, validation, and test sets using stratified sampling to preserve class distribution. To enhance model robustness, two data transformation pipelines were applied: a standard ImageNet-style normalization pipeline for supervised learning, and a MoCo-style (Momentum Contrast) augmentation pipeline for self-supervised pretraining. The MoCo augmentation strategy included random resized cropping, grayscale conversion, horizontal flipping, and color jittering, enabling the creation of two distinct augmented views of each image. A custom dataset class was implemented to support contrastive learning by returning image pairs. This approach was used to pretrain a convolutional backbone that could later be fine-tuned on the labeled data for classification.

**2.6 Final Model**

The model development followed a two-stage approach combining self-supervised pretraining with supervised fine-tuning. In the first stage, we pretrained a ResNet-18 encoder using SimCLR on chest X-ray images, employing NT-Xent contrastive loss with temperature 0.1 and achieving strong feature representations through augmentation strategies including random crops, flips, and color jitter.

For the fine-tuning phase, we adapted the pretrained model for pneumonia classification by:

1. Freezing the encoder weights to preserve learned features

2. Adding a linear classification head (512→2 dimensions)

3. Using class-weighted cross-entropy loss (weights: [0.73, 1.58]) to handle dataset imbalance

4. Training with Adam optimizer (lr=1e-3) and step learning rate scheduling (gamma=0.1 every 5 epochs)

The model was evaluated on a 60/20/20 train/val/test split of 18,136 chest X-ray images.

**3. Results**

**3.1 MAE**

* Pretraining showed consistent loss reduction, with final epoch loss = 0.0031, indicating strong feature learning from unlabeled images.
* Fine-tuning results over 5 epochs:
  + Epoch 1: Accuracy 77.27%
  + Epoch 2: Accuracy 78.00%
  + Epoch 3: Accuracy 77.81%
  + Epoch 4: Accuracy 78.74%
  + Epoch 5: Accuracy 78.94%
* Final test accuracy: 78.49%

These results demonstrate that MAE is an effective method for learning rich representations from medical images without labels, and successfully transfers to pneumonia classification tasks.

**3.2 BYOL**

During the pretraining phase, the model successfully aligned augmented views of the same image. Epoch 3 showed the best performance with a BYOL loss of 0.6977.

During fine-tuning, the model achieved a final validation accuracy of 76.96% after 20 epochs. This indicates effective feature transfer from BYOL pretraining, supporting the method’s suitability for medical image classification when labeled data is limited.

**3.3 DINO**

On the 10,000-image sample with ViT-Tiny, the model achieved:

* Pretraining DINO loss: 3.9 (expected due to high uncertainty in early epochs)
* Fine-tuning test accuracy: 78.8%

This performance slightly improved over baseline and shows the potential of DINO in medical imaging. However, larger-scale pretraining is needed to unlock full performance.

The pipeline, architecture, and training setup were confirmed functional. The main limitation was computational constraints, which limited the model’s ability to reduce loss and learn stronger representations during pretraining. Nonetheless, the results indicate that the DINO framework is promising for medical self-supervised tasks with more training time.

**3.4 SimCLR**

The results show that this combination of contrastive pretraining with a carefully chosen temperature parameter followed by lightweight supervised fine-tuning can effectively learn from medical imaging data, even when labeled examples are limited. The temperature value of 0.2 proved particularly effective for this chest X-ray dataset, helping the model learn discriminative features while maintaining stable training. Future work could explore adjusting this temperature parameter or experimenting with different learning rate schedules to potentially improve performance further.

**3.5 MoCo**

The use of MoCo-style augmentations enabled effective self-supervised pretraining by generating diverse image views for contrastive learning. This approach helped the model learn generalizable feature representations from unlabeled data, which is especially valuable in medical imaging contexts where annotations are limited. The stratified splitting preserved class balance across all dataset subsets, and the use of ImageNet normalization facilitated compatibility with pretrained models such as ResNet. The integration of contrastive learning before supervised fine-tuning provided a strong starting point for classification tasks, potentially improving downstream performance. While detailed evaluation metrics were not provided here, the pipeline is structured to support robust training and evaluation of models in clinical image classification scenarios.

**3.6 Final Model**

Key results showed:

- Training accuracy stabilized around 79.5% after 20 epochs

- Validation accuracy peaked at 80.46% in epoch 3

- Final test accuracy reached 80.22% with loss of 0.4279

The training process benefited from:

- Larger batch size (128) enabled by GPU acceleration

- Progress monitoring through tqdm integration

- Regular validation checks every epoch

- Learning rate reduction at epoch 5 and 10

This implementation demonstrates how self-supervised pretraining can effectively bootstrap medical image classification, with the frozen encoder approach providing:

1. Faster fine-tuning by only training the classifier head

2. Prevention of catastrophic forgetting of pretrained features

3. Efficient use of limited labeled medical data

The final model shows strong performance on pneumonia detection while maintaining computational efficiency through transfer learning principles.

**Comparison Table With Related Work**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | |  | | --- | |  |  |  | | --- | | **SSL Type** | | Architecture | Pretraining Dataset | Fine-tune Dataset | Accuracy (Ours) | Accuracy (Their Work) | Notes |
| MAE | Generative | |  | | --- | |  |  |  | | --- | | ViT-B/16 (MAE) | | NIH ChestX-ray14 (60k) | RSNA Pneumonia | |  | | --- | | **78.49%** |  |  | | --- | |  | | Not reported | |  | | --- | |  |  |  | | --- | | ViT-based; not implemented in their review, only mentioned as promising | |
| SimCLR | Contrastive | ResNet-18 | NIH ChestX-ray14 (60k) | RSNA Pneumonia | |  | | --- | | 76.52% |  |  | | --- | |  | | 76.4–79.6% | Matches accuracy of SimCLR in some reviewed papers (e.g., Azizi et al.) |
| BYOL | Predictive | ResNet-18 | NIH ChestX-ray14 (60k) | RSNA Pneumonia | |  | | --- | | 76.96% |  |  | | --- | |  | | 77.0–80.2% | Comparable to literature using BYOL in radiology |
| DINO | Self-distillation | ViT-Tiny | NIH ChestX-ray14 (10k) | RSNA Pneumonia | |  | | --- | | **78.8%** |  |  | | --- | |  | | 78.5–81.0% | |  | | --- | |  |  |  | | --- | | Consistent with Payer et al. ViT results on X-ray | |
| Final Model | Combined + Fine-tuned | ResNet / VIT | NIH ChestX-ray14 | RSNA Pneumonia | |  | | --- | | **80.22%** |  |  | | --- | |  | | - | |  | | --- | |  |  |  | | --- | | Best performance in our study; fine-tuned from SimCLR-encoded weights | |