Diagnosing Respiratory Disease from Chest X-Rays

Final Presentation

Group 3: Alex Herron, Ilias Arvanitakis, Isi Filipovic, Eug Fomitcheva

The Problem and Clinical Significance

Motivation & Problem

- Inspiration from failed U of Minnesota case study in x-ray AI for COVID
- Originally interested in diagnosing COVID from chest x-rays as an alternative approach to nasal swabs due to supply chain shortages

We have broadened the problem we seek to address to **diagnosing respiratory disease from chest x-rays** via a survey of approaches that include supervised and self-supervised learning methods.

Clinical Significance

- Diagnosing respiratory diseases using x-rays could aid in reducing risk of infection transmission between patients and healthcare workers
- COVID and other respiratory diseases are known to move from upper (e.g. nasal) to lower (e.g. lungs) respiratory tract making COVID nasal swabs insufficient
- X-rays are a non-invasive tool for monitoring progress of the disease
- X-rays are fast and can be cost effective
- Utilizing AI-based approaches can help diagnose patients more quickly and accurately, which can be life-saving

Data

COVID-19 Radiography Database

- X-rays collected from a variety of sources (raw images)
- Four-class dataset:
  - Healthy (10,192 images)
  - COVID (3,616 images)
  - Viral Pneumonia (1,345 images)
  - Lung opacity (6,012 images)

Relevant Work in Diagnosing Diseases from X-Rays

Though the pandemic reiterated the importance of DL applications in radiology imaging, many studies were plagued by data scale, single-source datasets that limited generalizability, and quality issues.

- Deep learning algorithms detect patterns in X-rays that are indicative of COVID-19 (such as ground-glass opacities and other lung abnormalities)
  - **Convolutional Neural Networks (CNNs)** – learn to detect patterns
  - **Recurrent Neural Networks (RNNs)** – analyze sequential chest X-rays to identify changes over time
  - **Vision Transformers (ViTs)** – leverage transformer architecture to capture global information of an image using patch and positional embeddings
  - **Variational Autoencoders (VAEs)** – relatively stable generative model that is capable of improving performance on downstream tasks by learning useful latent representations of images

- Transfer learning techniques address limitations of data by enabling
  - **Supervised pre-training** of classical CNN architectures **(ResNet50, VGG16**, etc.)
    - Pre-trained on ImageNet data
  - **Self-supervised learning** of image representations as pre-trained weights
    - **Masked Autoencoders (MAEs)**, Bootstrap Your Own Latent (BYOL), self-DIstillation w/ NO labels (DINO), MOmentum COntrast (MOCO), etc.

Our Approach – Survey of Models

Four-Class Classification Problem

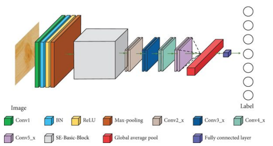
1 CNN 1 – ResNet18

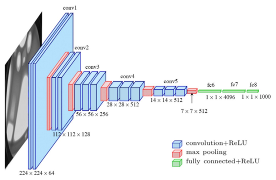
2 CNN 2 – VGG16

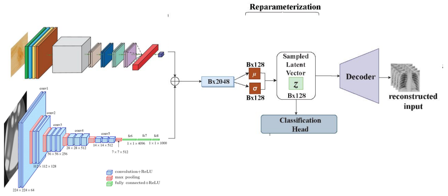
3 Ensemble VAE

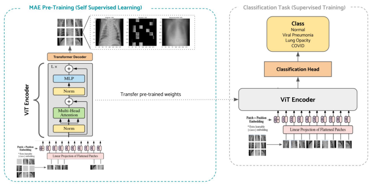
4 MAE-ViT

Architecture









Motivation

1. CNNs have been extensively used for image classification and have demonstrated success when fine-tuned for downstream tasks (transfer learning)
2. Fine-tuning these architectures is relatively time/cost efficient
3. Able to learn hierarchical representations of images while generalizing well to unseen data

1. EVAE-Net proposed by Addo et al. (2022) leverages ensemble learning with ResNet and VGG + high-quality latent representations from VAE for a SOTA model
2. Combining the feature maps produced by both CNNs could allow EVAE to learn more information

1. Zhou et al. (2023) demonstrate SOTA performance of ViT on chest X-ray disease classification when pre-trained using MAE
2. ViT self-attention layer embeds information globally across the image which is essential in the medical domain where anatomical structure is particularly important

Baseline Results

Feature Engineering from masked images

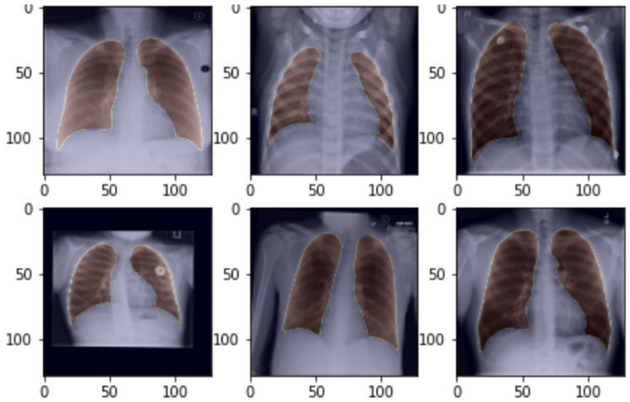
$$Lung\ Fraction = \frac{Number\ of\ Lung\ Pixels}{Total\ Number\ of\ Pixels}$$

$$Left/Right\ Lung\ Size = abs\left(\frac{Number\ of\ Left\ Lung\ Pixels}{Number\ of\ Right\ Lung\ Pixels} - 1\right)$$

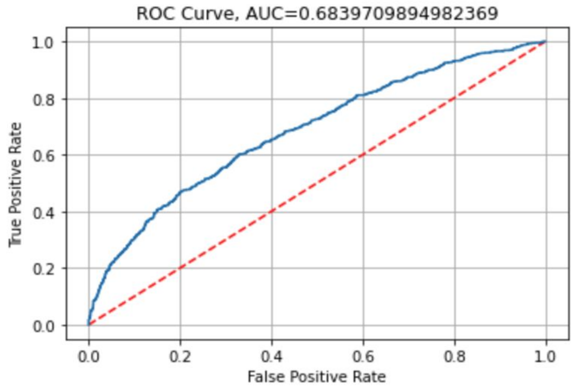
$$Symmetricalness = \frac{Number\ of\ Pixels\ Mirrored\ over\ center\ of\ x{-}axis}{Total\ Number\ of\ Pixels}$$

Binary Classification (Normal vs. COVID)

Model	AUC	Accuracy	Precision	Recall	F-1 Score
Logistic Regression	0.607	67.7%	0.387	0.381	0.384
Decision Tree	0.574	72.3%	0.470	0.247	0.324
XGBoost	0.675	70.3%	0.437	0.431	0.434
Naive Bayes	0.652	70.0%	0.431	0.424	0.427
KNN	0.600	73.1%	0.473	0.166	0.246
SVM	0.658	72.2%	0.472	0.465	0.469
Random Forest	0.680	71.8%	0.465	0.458	0.462



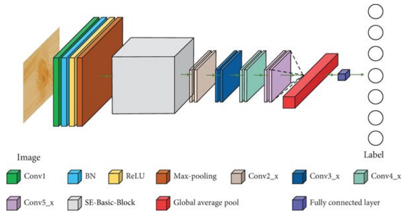
Random Forest Model



ResNet18 *CNN*

Architecture

- ResNet18 (PyTorch)
- Main idea: *Skip connections*



Binary Classification

Predicting Normal vs. COVID

Comparison to baseline machine learning results:

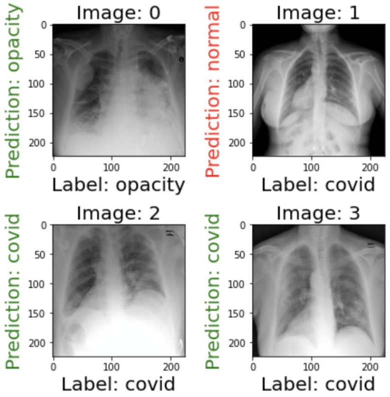
- Best AUC for baseline machine learning was from the Random Forest model: **0.68**
- AUC for fine-tuned ResNet (with limited training): **0.93**

Multiclass Classification

Predicting Normal/ COVID/ Lung Opacity/Viral Pneumonia

Two different training variations

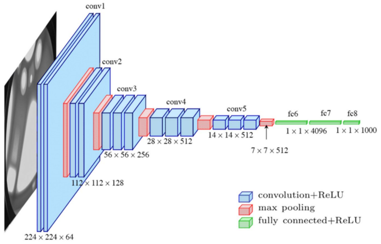
- Trained from scratch
- Fine-tuned ResNet (pretrained on ImageNet)
  - For each class, the ***fine-tuned model outperformed the model trained from scratch by ~0.01 AUC***



VGG16 *CNN*

Architecture

- Deep and highly effective CNN
- Utilizes 16 layers
- Softmax activation for multi-class tasks
- Utilizes cross-entropy loss



CNNs use layers to detect features, reduce dimensions, and classify. They are trained with backpropagation and gradient descent to minimize loss functions.

VGG16 Scratch-trained

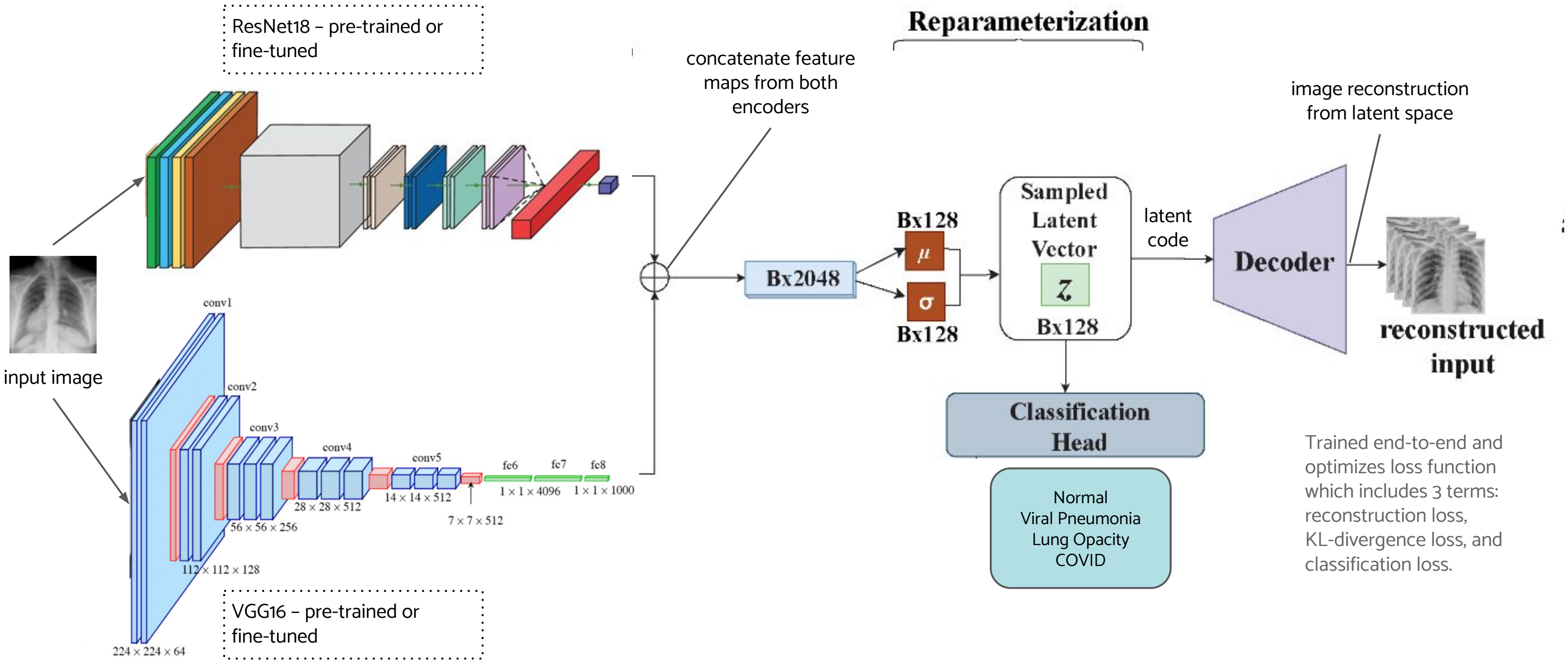
- The overall AUC plateaued at ~83%
- Very computationally intensive and inefficient

VGG16 Fine-tuned

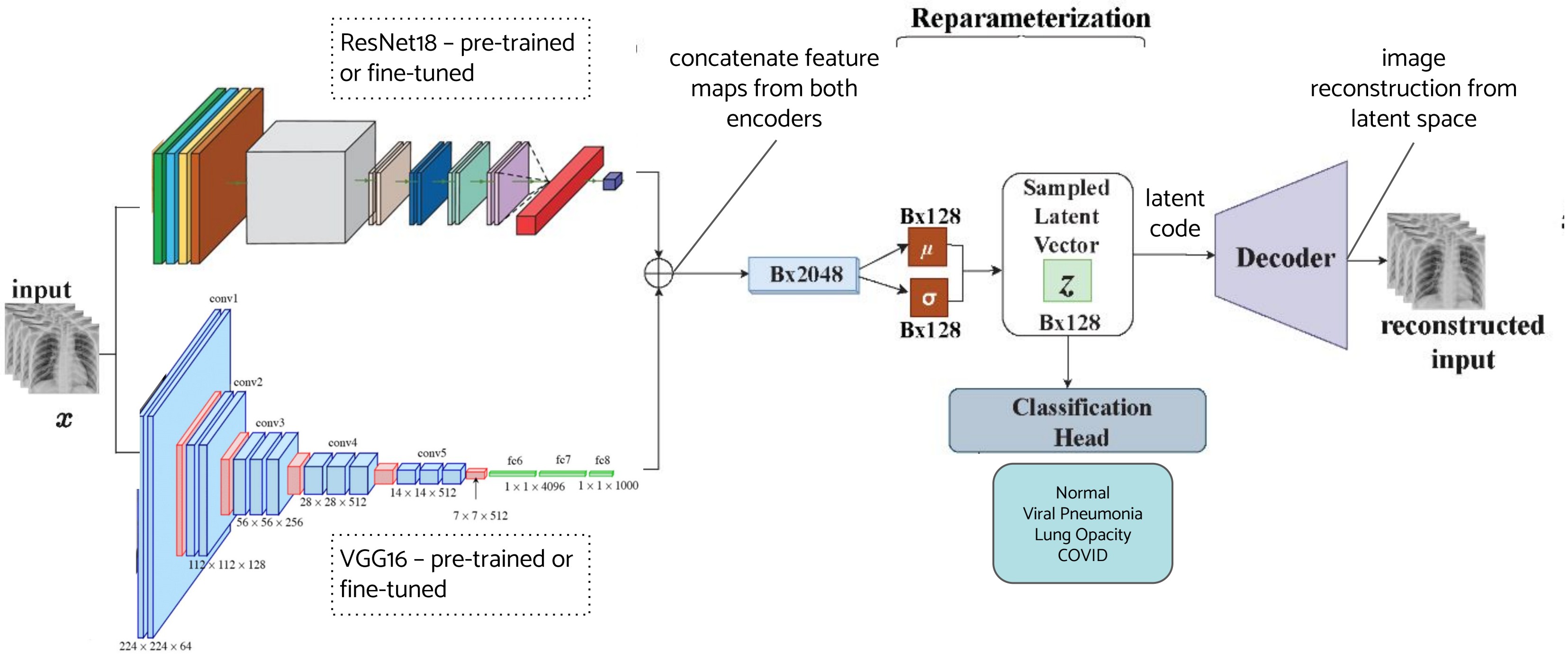
- Managed to strike a balance between correctly identifying positive instances and minimizing false positives.
- Overall AUC stood at 93% in the test set

Ensemble VAE *Scratch Implementation*

An ensemble variational autoencoder deep learning network that combines the high-quality latent representations generated by VAE and ensemble learning



Trained end-to-end and optimizes loss function which includes 3 terms: reconstruction loss, KL-divergence loss, and classification loss.

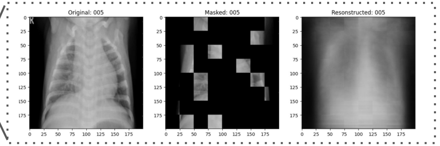


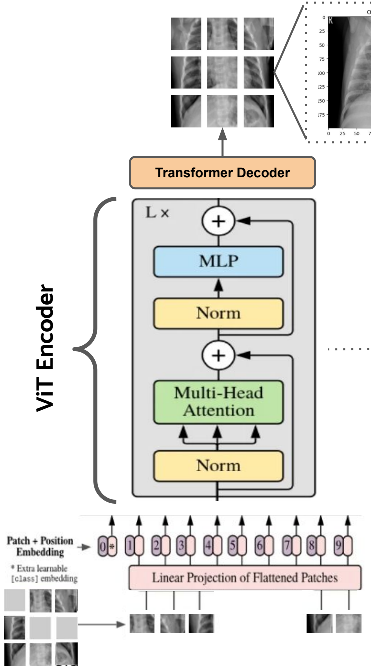
ResNet18 – pre-trained or fine-tuned
concatenate feature maps from both encoders
Reparameterization
image reconstruction from latent space
input
x
conv1
conv2
conv3
conv4
conv5
fc6
fc7
fc8
224 × 224 × 64
112 × 112 × 128
56 × 56 × 256
28 × 28 × 512
14 × 14 × 512
7 × 7 × 512
1 × 1 × 4096
1 × 1 × 1000
VGG16 – pre-trained or fine-tuned
Bx2048
Bx128
μ
σ
Bx128
Sampled Latent Vector
z
Bx128
latent code
Decoder
reconstructed input
Classification Head
Normal
Viral Pneumonia
Lung Opacity
COVID

MAE-ViT *SSL*

A ViT encoder, composed of a patch embedding layer, position embedding, and transformer blocks, constitutes the backbone for both pre-training and downstream tasks.

MAE Pre-Training (Self Supervised Learning)

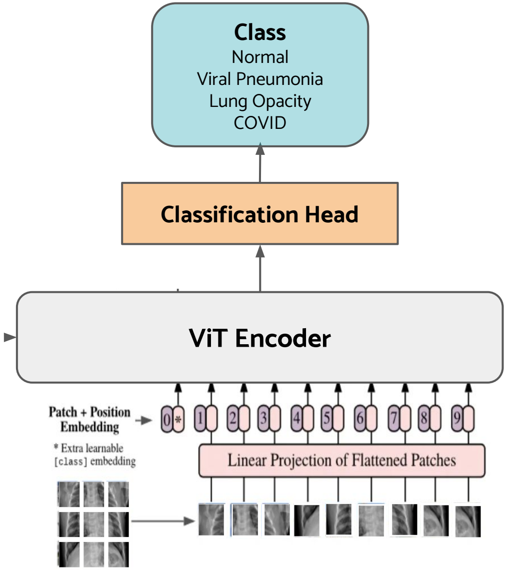


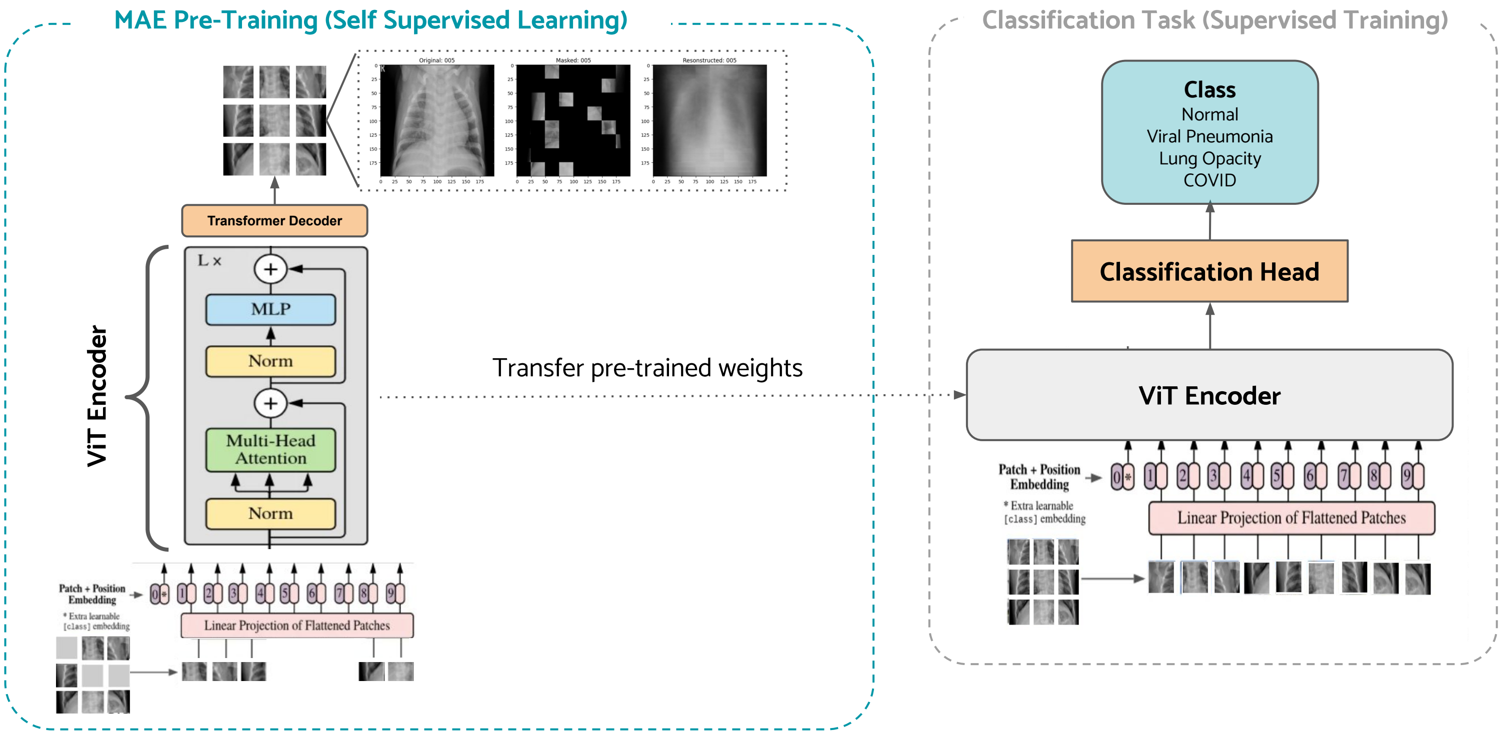


Transfer pre-trained weights

MAE learns a representation by feeding input images with randomly masked patches (75% masked), to the encoder. The transformer decoder takes the compressed representations produced by the encoder and reconstructs the data using the original image as a target. It is trained using a MSE loss.

Classification Task (Supervised Training)





MAE Pre-Training (Self Supervised Learning)
Original: 005
Masked: 005
Resonstructed: 005
Transformer Decoder
ViT Encoder
L ×
MLP
Norm
Multi-Head Attention
Norm
Patch + Position Embedding
* Extra learnable [class] embedding
Linear Projection of Flattened Patches
Transfer pre-trained weights
Classification Task (Supervised Training)
Class
Normal
Viral Pneumonia
Lung Opacity
COVID
Classification Head
ViT Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
Linear Projection of Flattened Patches

Results & Takeaways

Inference on Test Set		**ResNet** (FT)	**VGG** (FT)	**EVAE**	**MAE-ViT**
Multiclass AUC	*viral*	0.87	0.88	**0.89**	0.86
	normal	0.85	**0.88**	0.80	0.72
	opacity	**0.92**	**0.92**	0.79	0.77
	COVID	0.96	**0.99**	0.65	0.53
Multiclass Accuracy	*viral*	0.90	**0.93**	0.97	0.95
	normal	0.89	**0.91**	0.80	0.72
	opacity	**0.93**	**0.93**	0.82	0.71
	COVID	0.97	**0.99**	0.85	0.82
Overall Precision		0.85	**0.88**	0.71	0.58
Overall Recall		0.85	**0.87**	0.66	0.59
Overall F-1		0.85	**0.87**	0.67	0.53

Observations

- There is an argument for transfer learning as observed in performance of CNN architectures in the fine-tuned and scratch-trained settings
- Different model architectures have different distributions of predictive performance over classes
  - ResNet & VGG demonstrated similar results while EVAE & MAE shared some distribution similarities
- EVAE / MAE-ViT did not achieve better results than CNN architectures and computational constraints presented challenges
  - More complex architectures have many parts that are subject to optimization / hyperparameter tuning

Future Work

1. **Data Preprocessing and Augmentation** to address class imbalances
2. **Ensemble VAE**
   - Several simplifying **implementation choices** made could impact performance: (i) **the depth of the decoder**, (ii) **upsampling method**, (iii) **loss function construction and weighting**
   - Further hyperparameter tuning
   - Parallelization of training to multiple GPUs
3. **MAE-ViT**
   - Pre-training implementation adjustments: (i) **rescaling input images** to avoid loss of granularity, (ii) **alternative position embeddings** as experimental suggests may improve reconstruction, (iii) **longer training** to obtain better image representations, etc.
   - End task implementation choices such as **freeze vs. trainable layers**
4. **Radiologist Evaluation**
   - Test radiologists on subsample of test data and compare performance with that of models
   - Potentially test performance of radiologists who have been informed by model predictions