Group 3: Diagnosing Respiratory Disease from Chest X-Rays using Computer Vision

Ilias Arvanitakis IA2248@NYU.EDU
Eugenia Fomitcheva EDF257@NYU.EDU
Isidora Filipovic IF494@NYU.EDU
Alex Herron AH5865@NYU.EDU

1. Introduction

Respiratory diseases, including COVID-19 and pneumonia, are major public health concerns worldwide. Accurate and timely diagnosis is crucial for effective treatment and management of these diseases. While traditional diagnostic methods for respiratory diseases involve tests such as nasal swabs or sputum analysis, which are time-consuming, expensive, and can pose a risk of infection transmission to healthcare workers, chest X-rays are a non-invasive and widely available tool for diagnosing these diseases. However, their interpretation requires specialized expertise. The use of transfer learning methods, variational autoencoders, and self-supervised learning approaches has shown great potential in automating diagnosis from medical images. In this paper, we build upon existing methods found in our literature review of AI-based approaches for diagnosing respiratory diseases and compare their performance in a multi-class classification setting. We concluded that different models worked well for predicting different diagnoses, with ResNet18 and VGG16 demonstrating similar results while Ensemble VAE (E-VAE) and MAE-ViT also showed performance similarities by class. Overall, we found the most success with transfer learning methods using convolutional neural networks (CNNs), as can be seen with the success of fine-tuned ResNet18 and VGG16 in Figure 1. All code can be found at the following GitHub repository.

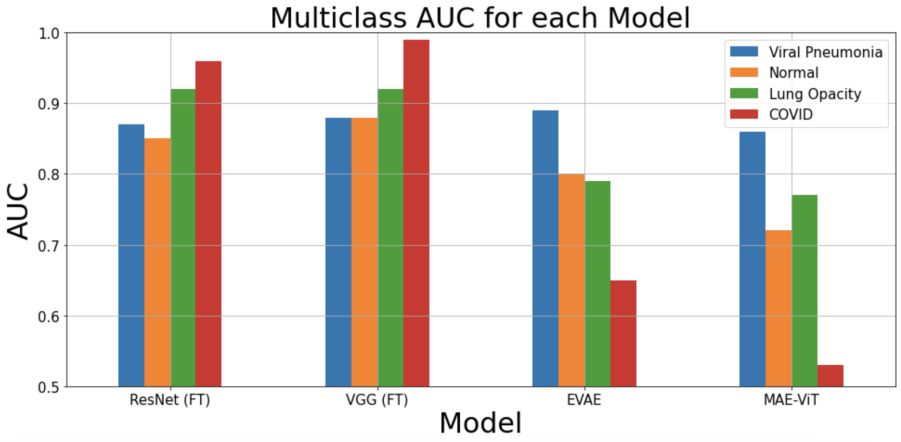

Figure 1: Comparison of Multi-class AUC for each Model

2. Related Work

The diagnosis of COVID-19 from chest X-rays has been an active area of research since the pandemic outbreak. Several studies have shown that deep learning approaches on radiography images can help diagnose COVID-19 patients in a safe, non-invasive manner and without the same risk of infection as with nasal swab collection. Following the study of Apostolopoulos and Mpesiana (2020) (which took place in

March 2020), many more researchers began investigating new approaches to predicting respiratory disease. Abiyev and Ismail (2021) used a CNN-based model for early COVID diagnosis, using a two-step approach. First, lungs were classified as healthy or unhealthy. Second, unhealthy lungs were classified as either COVID or pneumonia. This method did not involve lung opacity, only minor data augmentation to enhance the images. The notable point of this approach is that the CNN would first learn the difference between the healthy and unhealthy lungs, and then it would be learn if the patient had COVID or pneumonia.

Later, Kör et al. (2022) employed the NASNet-Mobile model, which is a SOTA CNN pretrained on 1.2 million images. Their major innovation was fine-tuning the final layers, producing higher accuracy scores. Muacevic and Adler (2021) used ground glass opacity and consolidation to train a DenseNet model to predict pneumonia. Their model was initially trained with data to predict pneumonia, and then retrained with COVID images. Although this method was not very effective, it could still be used for treatment monitoring purposes. Danilov et al. (2022) used 9 different methods to evaluate lung opacity: U-net, U-net.++, DeepLabV3, DeepLabV3+, FPN, Linknet, PSPNet, PAN, MA-Net. The process followed was as follows: (i) lung segmentation: pixel-level localization of the lungs and removal of unnecessary areas and (ii) disease segmentation: pixel-level localization of the infected area of the lungs. The most accurate model with respect to mean absolute error and root mean squared error was DeepLabV3 + for lung segmentation and MA-Net for disease segmentation. Musallam et al. (2022) used a deep Convolutional Neural Network (DCNN) called DeepChest for detection of both pneumonia and COVID. Their strategy included eliminating extraneous variables, de-noising, and increasing the size of the dataset by artificially generating images. This approach yielded strong performance and can potentially be used as a computer-aided diagnosis tool for detecting pneumonia and COVID from chest X-ray images. Addo et al. (2022) also built on the use of CNNs for medical image classification by proposing an ensemble method that concatenates pre-trained weights from two common CNN architectures (ResNet50 and VGG16) and passes them to a variational autoencoder (VAE) which learns a low-dimensional representation of the data that is used to ultimately reconstructs and classify an image.

Inspired by successes within the NLP domain, more recent efforts for image classification have adopted self-supervised learning methods for pre-training of classification models. He et al. (2021) demonstrate the ability of masked autoencoders (MAEs) to learn useful image representations by reconstructing partially masked input images. These image representations can then be used as pre-trained weights and fine-tuned during supervised training for downstream classificaton tasks. MAEs embed information globally across an image making them particularly applicable to medical imaging, where anatomical structure is important for diagnosis. Recognizing this applicability, as well as scalability and efficiency benefits of MAE, Zhou et al. (2023) demonstrate the power of MAE pre-training on various medical imaging classification tasks, including chest X-ray diagnosis, ultimately finding that Vision-Transformer (ViT) classifiers perform best when pre-trained using MAE.

3. Methods

3.1. Non-Deep-Learning Methods

For our non-deep-learning methods (composed of traditional machine learning approaches), we created a dataset class responsible for appropriately reading in the data by raw image and mask for each of four classes (normal, COVID, lung opacity, and viral pneumonia). Subsequently, we proceeded with feature engineering, where we introduced three metrics to evaluate the masked images: the size of the lung, the normalized ratio of left and right lung sizes, and the symmetry of the lungs. The equations for these features can be seen below:

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

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

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

We then sought to evaluate how effective these features would be in diagnosing patients. As this was the preliminary modeling effort, we chose to reduce our classification problem to a binary one. Specifically, we trained our models to predict whether patients had COVID or were healthy (normal). We trained 7 different models: Logistic Regression, Decision Tree Classifier, XGBoost Classifier, Naive Bayes Classifier, K-Nearest-Neighbors Classifier, Support Vector Machine, and a Random Forest Classifier. Each of these models was trained using only the three engineered features.

3.2. Convolutional Neural Networks

The seminal paper by Yosinski et al. (2014) investigated the transferability of features learned by deep neural networks, finding that the first-layer filters learned by are often simple and can be effectively transferred to other computer vision tasks while last-layer filters of CNNs are highly task-specific, requiring fine-tuning. Given the success seen with transfer learning and the challenges generally associated with obtaining large and diverse datasets of labeled medical images, we decided to invoke transfer learning with well-known CNN architectures—ResNet18 and VGG16 pre-trained on ImageNet—by comparing the scratch-train version of these models with their pre-trained, fine-tuned counterparts.

3.2.1. ResNet18

ResNet18 (He et al. (2015a)) is a deep learning model that addresses the degradation problem that often arises when training deeper neural networks. The degradation problem refers to when increasing the network depth leads to accuracy saturation and then performance degradation.

The ResNet18 architecture consists of 18 layers, starting with a single 7×7 convolutional layer followed by a 3×3 max-pooling layer. The network is then divided into four segments, each containing two residual blocks with two convolutional layers, normalization, and a ReLU activation function. The total of these 8 residual blocks, along with the initial convolutional layer and a concluding fully connected layer, completes the 18-layer structure of ResNet18. A key feature of the architecture is the so called "skip connections" that directly link earlier layers to later ones (skipping others). Skip connections allow for the gradient to flow easily through the network and help avoid the vanishing gradient problem. These connections execute identity mappings (passing the input directly to the output without transformation), contributing no additional parameters or computational complexity, making ResNet lightweight relative to some other CNNs. ResNet18 concludes with a global average pooling layer followed by a fully connected layer and an output layer that employs a softmax activation function. This design enables effective training of the network while offering improved accuracy on large-scale and smaller datasets alike.

3.2.2. VGG16

While ResNet18 is a more manageable size (11 million parameters), VGG16 (Simonyan and Zisserman (2015)) introduces greater computational complexity (138 million parameters). Despite its size, the VGG network is relatively simple, consisting of 13 convolutional layers with varying filter sizes, followed by 3 fully connected layers (16 total). Following the first two fully connected layers, VGG16 makes use of dropout layers to prevent overfitting. Then, the final output layer uses a softmax activation function.

In practice, VGG16 makes up for some of its computational complexity with the ease of its implementation. While VGG16 does not employ the skip connections in ResNet18, it successfully relies on stacking of several convolutional layers with smaller filters to learn hierarchical features. The VGG16 model is often trained using the categorical cross-entropy loss function, which is well-suited for multi-class classification tasks.

3.3. Ensemble VAE

Variational autoencoders (VAEs) have become a popular tool due to their ability to learn meaningful, low-dimensional latent representations of data, their capacity to generate new instances with controlled variation and their stability in training. Inspired by Addo et al. (2022), an ensemble VAE achieving SOTA results on a three-class version of our classification problem, we sought to reconstruct a version of the implementation (as original code is unpublished). An illustration of our architecture is in Figure 2.

For our scratch-implemented E-VAE we leveraged our previously trained ResNet18 and VGG16 already fine-tuned on the dataset and with the final classification layer removed. The produced feature maps from these two encoders were concatenated into a single, richer feature map. The output was then reparametrized, sampling from the learned latent distribution to produce a new latent vector. The reparametrization trick, sampling the latent variables from a learned distribution parametrized by mean and variance vectors (commonly implemented with VAEs), was used to enable backpropagation during training. The resulting latent space was used for both image reconstruction and classification. The decoder of the E-VAE, which uses a series of 4 transpose convolution layers, upsampled the latent space to reconstruct the input image while a linear classification head output a prediction for the image class.

In end-to-end training, the E-VAE aims to minimize the aggregate (summed) loss function which is composed of (i) the reconstruction loss between the reconstructed and input image, (ii) the KL-divergence between the prior and learned latent distribution and (iii) the classification loss. By minimizing the aggregate loss function the E-VAE learns to produce representations of the input data, which can be used for downstream tasks beyond classification.

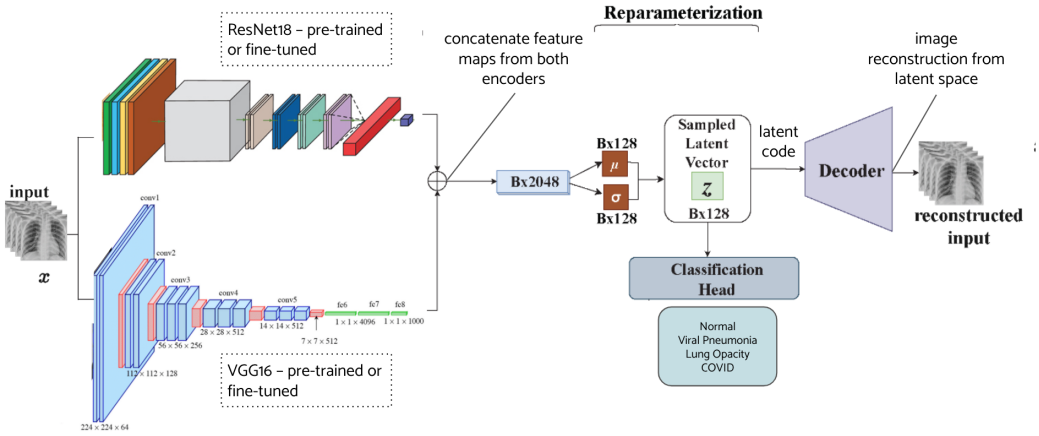

Figure 2: E-VAE Architecture with ResNet18 and VGG16 encoders

3.4. Self-Supervised Learning: MAE-ViT

The methodologies discussed above employ supervised training methods to obtain a pre-trained models on which task and data specific fine-tuning occurs. However, recent research suggests that self-supervised learning methods, which already have strong precedence in NLP tasks, can be used to learn representations of images. Transfer learning can then once again be employed by passing these learned representations and fine-tuning on downstream tasks. Following the work of Zhou et al. (2023), we implemented a Vision Transformer classifier (ViT) pre-trained using MAE. A ViT encoder, composed of a patch embedding layer, position embedding, and transformer blocks, constitutes the backbone for both pre-training and downstream tasks. MAEs are a type of neural network that use unsupervised learning to encode and decode data. The goal of a masked autoencoder is to learn a representation of the images data, which can then be used for the downstream classification task. The image is first divided into fixed-size patches

and a trainable linear projection is applied to map them to patch embeddings. A portion of the patches is randomly selected as masked. In our implementation, 75% of patches were masked. This masking rate, originally proposed by He et al. (2021), is fairly common practice and efficient as only visible patches are fed into the encoder. The patch embeddings and positional embeddings (including positional embeddings for masked patches) are inputs to the transformer encoder. The transformer block consists of alternating layers of multiheaded self-attention (MSA) and multilayer perceptron (MLP) blocks. In our implementation this block utilized an encoder with $encoder\_dim = 128$, comprises of six layers, each with MSA ($heads = 4$), normalization and a densely connected sub-layer. The self-attention layer in ViT makes it possible to embed information globally across the overall image. The lower-dimensional representation outputted by the encoder is then fed into the lightweight decoder (we use $decoder\_dim = 64$) which reconstructs the original data. A MSE loss is used for training, comparing the reconstructed image to the original. This piece constitutes the MAE as a self-supervised architecture as only the images, and no labels, are used in pre-training.

After MAE pre-training, the learned weights are transferred and fine-tuned again using the ViT architecture, albeit with adjustments for the downstream classification task. In the fine-tuning phase, no image patches are masked, but patch and positional embeddings (ie the trained weights) are passed into the encoder, just as in pre-training. Instead of a decoder, the ViT encoder output is fed into a classification head. The ViT classifier, composed of a global average pooling layer followed by a fully connected layer with a softmax activation function, is fine-tuned using the same training data, but this time in a supervised manner with the labels included. In this way, MAE-based pre-training avoids the domain discrepancy between pre-training and fine-tuning (present in transfer learning implementations using ResNet18 and VGG16) by unifying the training data of two stages.

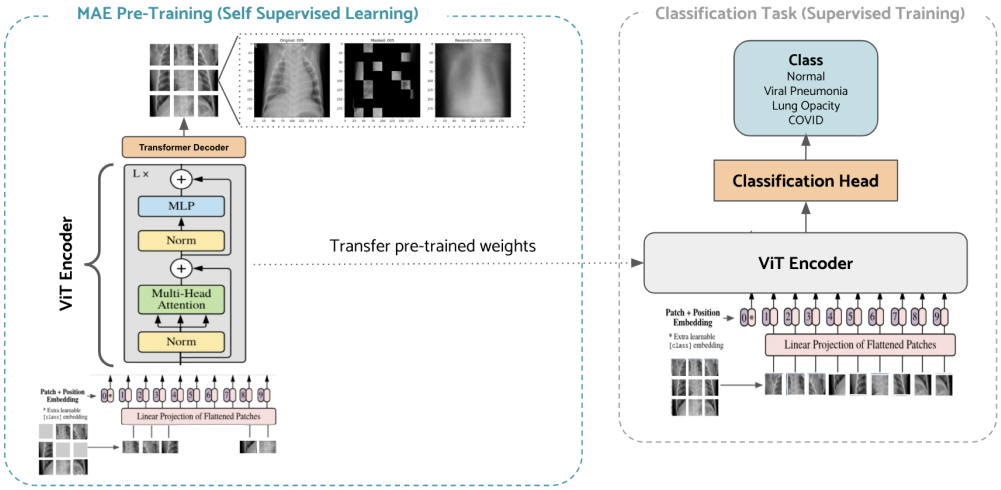

Figure 3: MAE-ViT Architecture

4. Data and Experiment Setup

4.1. Data

For this project we used data from the COVID-19 Radiography Database from Kaggle. This dataset was created by a team of researchers from Qatar University, Doha Qatar and the University of Dhaka, Bangladesh. The dataset includes images from multiple sources, including the Italian Society of Medical and Interventional Radiology COVID-19 Database, Novel Corona Virus 2019 Dataset (developed by Joseph Paul Cohen and Paul Morrison), Chest X-Ray Pneumonia Dataset (Kaggle), and Lan Dao's GitHub

which extracted images from 43 different publications. The dataset contains both X-ray images and their corresponding lung masks. The breakdown of the dataset by disease is as follows: lung opacity (6012 images/masks), COVID (3616 images/masks), viral pneumonia (1345 images/masks), and healthy patients (10,192 images/masks). The multi-source nature of our collection addresses two predominant issues identified by researchers in this field – lack of large COVID-19 X-Ray image datasets and issues with model robustness with respect to generalizing to out-of-distribution data.

4.2. Experimental Setup

As a preliminary baseline we trained and assessed the performance of 7 non-deep learning methods, using engineered features as described in Section 3.1, on the binary classification task of identifying healthy vs. COVID lungs. All models in the baseline used a 0.8/0.2 train/test split.

With the objective of outperforming these established baseline metrics and scaling to the four-class classification problem we trained and evaluated several deep learning models: ResNet18, VGG16, E-VAE, and MAE-ViT. All four approaches utilized a 0.6/0.3/0.1 train/validation/test split of the dataset and were evaluated on the created test set for multi-class AUC and accuracy metrics.

To observe the argument for transfer learning ResNet18 and VGG16 were both trained from scratch and compared to their fine-tuned counterparts on the COVID-19 Radiography Database. In the experiments for ResNet18, VGG16, and E-VAE input images were resized to 224×224 and the pixel distribution was re-standardized (mean of 0 and standard deviation of 1) to aid convergence in training. In the fine-tuning setting both CNNs were trained with early stopping at a 95% threshold for overall accuracy to prevent overfitting and used an Adam optimizer, batch size of 4 and learning rate of 0.00003. Using the validation set, the weights with best overall accuracy were saved.

For the E-VAE, images were pre-processed in the same manner as for the CNN experiments prior to feeding them into the fine-tuned and saved ResNet18 and VGG16 encoders. This model was trained end-to-end for a total of 20 epochs, using the same optimizer, batch and learning rate as for ResNet18/VGG16 and $latent\_dim = 256$.

In the MAE-ViT implementation, inputs underwent multi-step pre-processing: images were first resized to 110×110; a margin of 20 was then added to enable random cropping to final input size of 120×120; finally, the training images were then flipped horizontally flipped to establish the final "augmented" training set. Patch size was 10×10 resulting in 144 patches per image. A masking ratio of 75% was applied. 1-D positional embeddings were used for both pre-training and fine-tuning. Pre-training and fine-tuning was implemented with a batch size of 4, a learning rate of 0.001 (reduced when plateauing with a patience of 2) and early stopping with a minimum change of 0.0001 and a patience of 4, resulting in pre-training of 30 epochs and fine-tuning of 25 epochs. Lastly, though several different freeze/training layer settings were experimented with, results reported are those obtained when only the classification head was fine-tuned for our downstream task, as this implementation achieved the best performance.

5. Results

For evaluation metrics, we used AUC, accuracy, precision, recall, and F-1 score. While these metrics can be calculated normally for binary classification tasks, for multiclass classification adjustments need to be made. For AUC and accuracy, multiclass metrics can be defined for each class by re-framing the task as binary (i.e., for COVID AUC all COVID predictions will be labeled 1 and all other classes will be labeled 0). For precision, recall, and F-1 score, we set the 'average' setting in the scikit-learn evaluation metric functions to 'macro', which calculates metrics for each class and finds their average.

For our non-deep-learning models trained for binary classification (predicting patients to either be healthy or diagnosed with COVID), the obtained results are shown in Table 1. Although we did not expect great results for these baseline models, we did observe predictive power gained from our simple feature engineering. The best performing model of the group was the Random Forest Classifier with an AUC

of 0.680, followed by the XGBoost Classifier with an AUC of 0.675. Although nowhere near SOTA, this provided an initial indication of results to be expected from more advanced deep learning approaches. Additionally, the engineered features provide some insight into what might make lungs identifiable as having COVID versus being healthy. For example, lungs that are very asymmetrical are more likely to have COVID than symmetrical lungs.

Table 1: Evaluation Metrics for Non-Deep-Learning Methods in Binary Classification

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

In the second stage of our work we used a CNN trained for binary classification (normal vs. COVID) to compare with our non-deep-learning baseline results. For this, we used pre-trained ResNet18 fine-tuned on our training data. This model reached an AUC of 0.93, far surpassing our baseline result of 0.68 (RF using engineered features from masked images). This result aligned with our expectations that convolutional neural networks should outperform non-deep-learning methods by a substantial margin.

For the multi-class classification task (identifying normal, COVID, lung opacity and viral pneumonia) we compared the performance of the previously mentioned ResNet18 with VGG16 in two settings: pre-trained + fine-tuned and scratch-trained on our data. For ResNet18, the fine-tuned model outperformed the model trained from scratch by 0.01 for overall AUC. However, scratch-trained VGG16 saw overall AUC plateau at 0.83, while the fine-tuned version reached an overall AUC of 0.93. Furthermore, it is important to note that VGG16 is heavy-weight (138 million parameters) and substantially more computationally intensive to train than ResNet18 (11 million parameters). Overall, our observations indicate that transfer learning is particularly useful for this task, both in terms of predictive power and training time/computational resources.

The results in Table 2 highlight the fact that different model architectures are better at predicting different diseases. Using multi-class AUC as our primary evaluation metric, E-VAE performs best in predicting the smallest class, viral pneumonia (AUC = 0.89), VGG16 (FT) performs best for normal and COVID (AUC = 0.88 and 0.99, respectively), and VGG16 (FT) and ResNet18 (FT) tie for best performance on lung opacity (AUC = 0.92). Additionally, ResNet18 and VGG16 demonstrated similar trends across classes while E-VAE and MAE-ViT shared some predictive similarities regarding which diseases they could best predict.

Overall, we note that E-VAE and MAE-ViT did not achieve better results than the well-known CNN architectures. These more complex architectures have several components that are subject to optimization and hyperparameter tuning. Additionally, these larger models presented computational challenges that could potentially be improved upon with greater access to computing resources and additional training time. These challenges are further expanded upon in Section 6.

6. Discussion and Limitations

The objective of this study was to evaluate the performance of several deep learning models for the multi-class classification of chest X-rays into viral pneumonia, lung opacity, COVID-19 and normal classes. We

Table 2: Evaluation Metrics for Deep-Learning Methods in Multi-class Classification

Inference on Test Set	**ResNet (FT)**	**VGG (FT)**	**EVAE**	**MAE-ViT**
Multiclass AUC				
Viral Pneumonia	0.87	0.88	**0.89**	0.86
Normal	0.85	**0.88**	0.80	0.72
Lung Opacity	**0.92**	**0.92**	0.79	0.77
COVID	0.96	**0.99**	0.65	0.53
Multiclass Accuracy				
Viral Pneumonia	0.90	0.93	**0.97**	0.95
Normal	0.89	**0.91**	0.80	0.72
Lung 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

compared the performance of ResNet18, VGG16, E-VAE, and MAE-ViT models trained and evaluated on the COVID-19 Radiography Database.

As alluded to in Section 5, while we found our fine-tuned CNN architectures, and particularly VGG16, to outperform other models, we believe that steps could be taken to further improve the performance of E-VAE and MAE-ViT. From a data pre-processing perspective, as the dataset used features class imbalances, data augmentation and/or sampling could aid the performance of our models. We found E-VAE to be a promising approach for the multi-class classification problem (achieving high scores for certain classes) but due to implementation choices made on the basis of time and our original inspiration for the architecture (Addo et al. (2022)), performance could have been impacted. Further exploration in the form of ablation studies on the latent space dimension and depth of decoder, as examples, could help us better understand how to optimize the VAE itself. Finally, to optimize the E-VAE architecture, hyperparameter tuning and re-weighting of components of the loss function (the trade-off between reconstruction loss, KL-divergence, and classification loss) could be explored.

We've also identified several areas for improving the MAE-ViT implementation. Due to memory constraints, our input images were scaled down to size of 120×120, resulting in a loss of granularity. As other works on X-ray image classification demonstrated greater performance using images of input size 256×256 (with patch size 16), we believe increasing our input size may yield better results. Implementing sine-cosine position embeddings (as opposed to 1D position embeddings used in our implementation) may lead to better performance as experimental evidence suggests this achieves better reconstruction in pre-training (Zhou et al. (2023)). Pre-training on a greater number of epochs, as demonstrated by Zhou et al. (2023) who pre-trained their MAE for 800 epochs, may also lead to better reconstruction, resulting in higher downstream performance. Lastly, experimenting with freezing and unfreezing various layers in the downstream task fine-tuning may also achieve better results.

Regardless of potential model improvements, all four deep learning approaches suffer from interpretability limitations. For example, while these models are able to learn image representations they cannot indicate exact regions of interest in a chest x-ray, which could be critical for them to be used in conjunction with radiologists. Future work in this domain may explore attention mechanisms, which can identify significant regions, or utilize visualization techniques like class activation mapping to emphasize relevant information, thereby simplifying the task of diagnosis for radiologists.

7. Contributions

- Ilias – VGG modeling
- Eugenia – Scratch implementation of E-VAE
- Isidora – Implementation of MAE-ViT
- Alex – Non-deep-learning modeling, ResNet/VGG modeling

All team members contributed equally to the creation of final presentations as well as relevant sections in this report corresponding to their modeling focuses as outlined above.

References

URL https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database.

Rahib H. Abiyev and Abdullahi Ismail. Advanced learning methods in statistical signal processing and applications. 2021. URL https://www.hindawi.com/journals/mpe/2021/3281135/.

Daniel Addo, Shijie Zhou, Jehoiada Kofi Jackson, Grace Ugochi Nneji, Happy Nkanta Monday, Kwabena Sarpong, Rutherford Agbeshi Patamia, Favour Ekong, and Christyn Akosua Owusu-Agyei. Evae-net: An ensemble variational autoencoder deep learning network for covid-19 classification based on chest x-ray images. *Diagnostics*, 2022. URL https://www.mdpi.com/2075-4418/12/11/2569.

Ioannis D. Apostolopoulos and Tzani A. Mpesiana. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine volume*, 2020. URL https://link.springer.com/article/10.1007/s13246-020-00865-4.

Sankhadeep Chatterjee, Soumyajit Maity, Mayukh Bhattacharjee, Soumen Banerjee, Asit Kumar Das, and Weiping Ding. Variational autoencoder based imbalanced covid-19 detection using chest x-ray images. *New generation computing*, 2022. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9676807/.

Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 2019. URL https://doi.org/10.1016/j.media.2019.101539.

Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul. Can ai help in screening viral and covid-19 pneumonia? *Institute for Electrical and Electronic Engineers*, 2020. URL https://ieeexplore.ieee.org/document/9144185.

Viacheslav V. Danilov, Diana Litmanovich, Alex Proutski, Alexander Kirpich, Dato Nefaridze, Alex Karpovsky, and Yuriy Gankin. Automatic scoring of covid-19 severity in x-ray imaging based on a novel deep learning workflow. 2022. URL https://www.nature.com/articles/s41598-022-15013-z.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020. URL https://openreview.net/forum?id=YicbFdNTTy.

Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *Nature*, 2021. URL https://www.nature.com/articles/s41746-020-00376-2.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Microsoft Research, and Jian Sun. Deep residual learning for image recognition. 2015a. URL https://arxiv.org/pdf/1512.03385.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2015b. URL https://arxiv.org/abs/1512.03385.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable vision learners. *cs.CV*, 2021. URL //arxiv.org/pdf/2111.06377.

Hakan Kör, Hasan Erbay, and Ahmet Haşim Yurttakal. Diagnosing and differentiating viral pneumonia and covid-19 using x-ray images. 2022. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9042669/.

Romany F. Mansour, José Escorcia-Gutierrez, Margarita Gamarra, Deepak Gupta, Oscar Castillo, and Sachin Kumar. Unsupervised deep learning based variational autoencoder model for covid-19 diagnosis and classification. *Pattern Recognition Letters*, 2021. URL https://www.sciencedirect.com/science/article/pii/S016786552100310X.

Alexander Muacevic and John R Adler. Predicting covid-19 pneumonia severity on chest x-ray with deep learning. 2021. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7451075/.

Ahmed Salem Musallam, Ahmed Sobhy Sherif, and Mohamed K. Hussein. Efficient framework for detecting covid-19 and pneumonia from chest x-ray using deep convolutional network. 2022. URL https://www.sciencedirect.com/science/article/pii/S1110866522000020.

Soumya Ranjan Nayak, Deepak Ranjan Nayak, Utkarsh Sinha, Vaibhav Arora, and Ram Bilas Pachori. Application of deep learning techniques for detection of covid-19 cases using chest x-ray images: A comprehensive study. *Biomedical Signal Processing and Control*, 2021. URL https://www.sciencedirect.com/science/article/pii/S1746809420304717.

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M. Zughaier, Muhammad Salman Khan, and Muhammad E.H. Chowdhury. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in Biology and Medicine*, 2021. URL https://www.sciencedirect.com/science/article/pii/S001048252100113X?via%3Dihub.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. 2017. URL https://arxiv.org/abs/1711.05225.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. URL https://arxiv.org/pdf/1409.1556.pdf.

Yazan Qiblawey Anas Tahir Serkan Kiranyaz Saad Bin Abul Kashem Mohammad Tariqul Islam Somaya Al Maadeed Susu M. Zughaier Muhammad Salman Khan Muhammad E.H. Chowdhury Tawsifur Rahman, Amith Khandakar. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. 2021. URL https://www.sciencedirect.com/science/article/pii/S001048252100113X?via%3Dihub.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *NIPS Foundation*, 2014. URL https://arxiv.org/pdf/1409.1556.pdf.

Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image classification and segmentation. 2023. URL https://arxiv.org/pdf/2203.05573.pdf.