Final Project Report

Pengwei Sui
ps3307

Zheyu Zhang
zz2980

Abstract

In the final project, we focus on the image classification task under the big topic of computer-aided diagnosis (CAD). The task is to detecing the abnormalities in the chest caused by COVID-19 based on chest radiographs. The raw radiograph data comes from the Kaggle website in the form of DICOM radiographs. Radio graphs are significantly different from the general image so the normal pre-trained model based on ImageNet may not be appropriate for this classification task. In our final project, we utilized models such as EfficientNet and SWIN transformers to construct our pipelines. The output will be finalized by taking the average of outputs from all pipelines. The final accuracy from the ensemble model reaches around 74%. This project direction has a promising future since it makes COVID-19 detection based on radiographs in minutes and offers doctors solid evidence to provide a confident diagnosis.

Contribution:
Pengwei Sui wrote section: Related work(EfficientNet), Methodology, Experiment
Zheyu Zhang wrote section: Abstract, Introduction, Related Work (Swin transformer section), and Conclusion

Source Code: https://github.com/psui3905/COMS4995-Project

1 Introduction

For much of the world, 2022 marked the beginning of the end of the COVID-19 pandemic, but research regarding COVID-19 is still going on. Caused by the virus called SARS-CoV2, COVID-19 is an infectious disease that normally leads to a variety of diseases from head or chest colds to more severe diseases like acute respiratory syndrome and Middle East respiratory syndrome. COVID-19 would result in inflammation and fluid in the lungs through pulmonary infection, which makes it possible to check the infection of COVID-19 by the information carried by the chest radiographs. Instead of checking manually, computer-aided diagnosis (CAD) can transfer the workload from medical staff to the computer. More specifically, machine learning would take the duty from there.

In this final project, we utilized ResNet50, VGG16, EfficientNet-B5, and SWIN Transformer. etc models to explore, train, and test. We acquire data from the Kaggle competition 2021 whose task is "identify and localized COVID-19 abnormalities on the chest radiographs" and we mainly focus on the study-level experiments. The raw data are the DICOM radiographs of the patient's lung and there are four classes that we need to classify them into four classes. These are "Negative for Pneumonia", "Typical Appearance", "Indeterminate Appearance" and "Atypical Appearance".

We have encountered several obstacles during our development stages. Some of them come from raw data themselves: the quantity of data in the medical-related field is normally not very adequate since each data point comes from a living patient and those data will not be acquired in easy steps. Also, the size of the image data isn't very specific. For example, each of the original DICOM (Digital Imaging and Communications in Medicine) files contains $2320 \sim 4240$ pixels. Some of the obstacles come from not well-considered model construction which may lead to training overfitting or low learning efficiency.

Near the end of our project, we made several quantitative comparisons regarding different models and various input image sizes. It turns out that the ensemble method will output the most desirable result and input image with size 512 * 512 would lead to higher accuracy.

2 Related work

In this final project, we have adapted several different classification models in the field of computer vision. Those include ResNet (Residual Neural Network), VGG (Very Deep Convolutional Network), various versions of EfficientNet, and Swin Transformer (Sifted Window Transformer). After the experiment, our final model architecture is composed of Efficient Net-B5, Efficient Net-B3, and SWIN transformer. The following section 2.1 introduce the basic of Swin transformer and next section 2.2 describe the architecture and introduction of efficient net.

2.1 SWIN Transformer

Proposed by Ze Liu, Yuton Lin .ect in August 2021, Swin (**S**hifted **win**dows) Transformer expand the applicability of Transformer to a general-purpose backbone for computer vision. [12] In the field of computer vision, CNNs (convolutional neural networks) are in the status of domination. [7] With a large data set to train the model and then fine-tune with the task-related database, CNN model such as ResNet would offer relatively desirable output. At the same time, there are also some pioneering Transformer work being done in the field of computer vision, such as ViT (vision Transformer)[3], DeiT (Data-efficient Image Transformer) [18] and their follow-ups.

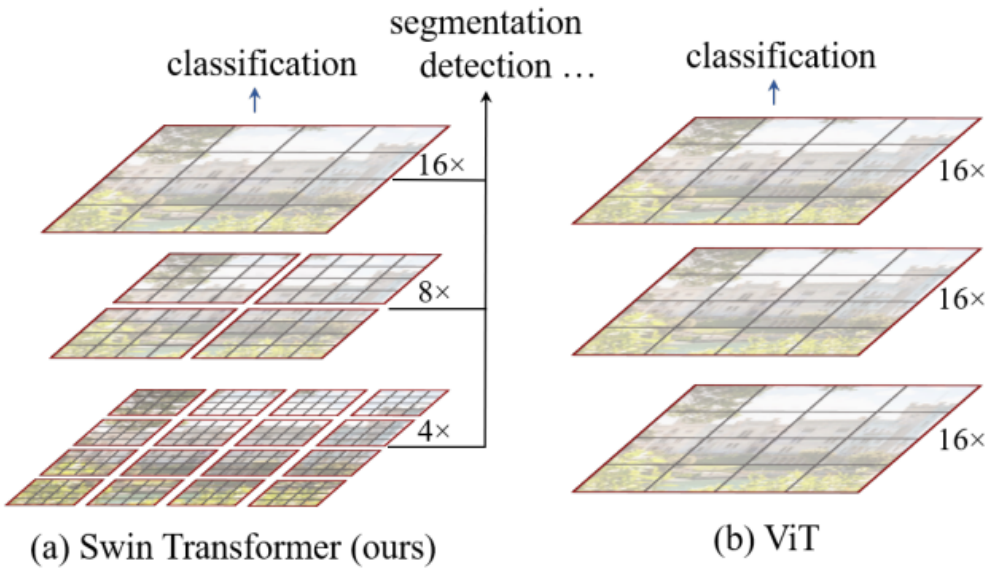


Figure 1: Feature map for Swin Transormer[12] and ViT [3].

Even though VIT is doing well in the vision field, but it has an unignorable shortage: it produces feature maps of a single low resolution and has quadratic computation complexity to input image size due to the computation of self-attention globally. [12] Also, all patches are mutually exclusive (non-overlapping), the attention mechanism can't learn from the neighbor patches. One solution is sliding window [9] based on self-attention approaches. It does cover the relation between different patches but suffers from low latency on general hardware.

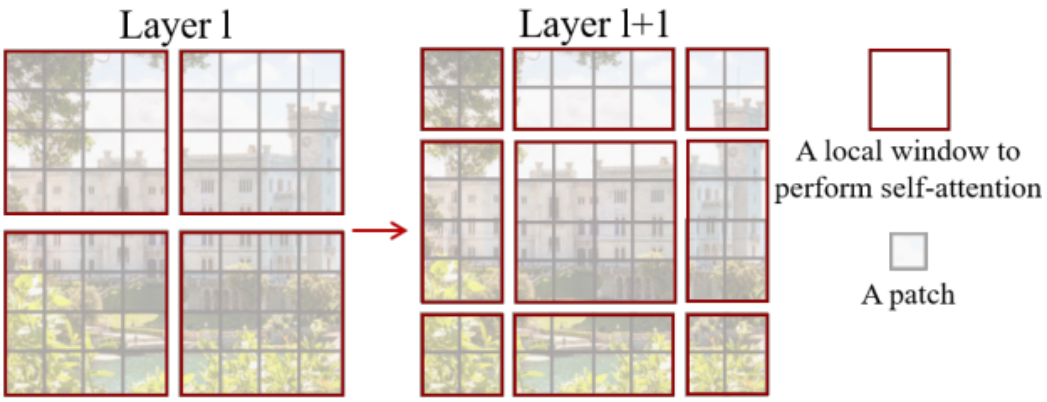


Figure 2: An illustration of the shifted window approach for computing self-attention in the proposed Swin Transformer architectures[12].

As shown in the image 2, two consecutive layers in Swin Transformer demonstrate the concept of "shifting". In Layer 1, Swin transformer uses regular window partitioning and computes self-attention in each patch. Then the window is shifted to its following layer. Each patch in the following layer covers the cross line of the nearby window in Layer 1. The self-attention computation in the following layer will add a connection of previously isolated patches. In addition, it saves plenty of computation with the comparison with the sliding window self-attention model, and therefore it doesn't suffer from the low latency. As mentioned by the author, this shifted window approach also proves beneficial for all-MLP architectures [12][17].

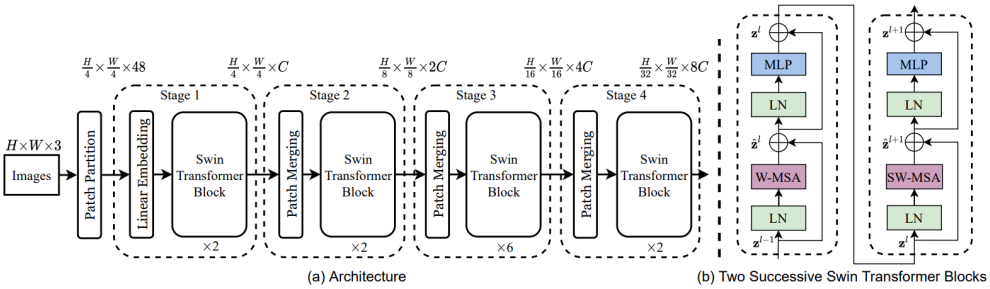


Figure 3: (a) The architecture of a Swin Transformer (b) Two successive Swin transformer Blocks[12].

Swin-B stands for the base model of Swin Transformer. Swin-T, Swin-S, Swin-L are three versions of about 0.25×, 0.5×, and 2× the original model size and computational complexity, respectively. The complexity of Swin-T and Swin-S are similar to those of ResNet-50 and ResNet-101, respectively. As concluded in the paper [12], Swin-T (architecture shows in 3 (a)), Swin-S, Swin-B reach 81.3, 83, and 83.5 accuracies respectively in regular ImageNet-1K trained models. In the ImageNet-22K pre-trained models, Swin-B would reach 85.2 accuracy.

2.2 EfficientNet

The performance of Deep Convolutional Neural Networks can be severely impacted by the model scaling, as multi-class classification on bigger images might require the model to have a larger receptive field and channels to capture the fine-grained patterns. EfficientNet proposed by [16] illustrates a new technique for rethinking the way of scaling for Convolutional Neural Networks. One of the most widely adopted ways for Convolutional Neural Network scaling is adding more layers, for instance, GPipe [11] can achieve the 84.3% top-1 accuracy on ImageNet[2] by scaling up the baseline model 4 times larger. Considering the development of a multi-task classification solution with a limited resource budget, scaling up the model architecture systematically is critical, as it is

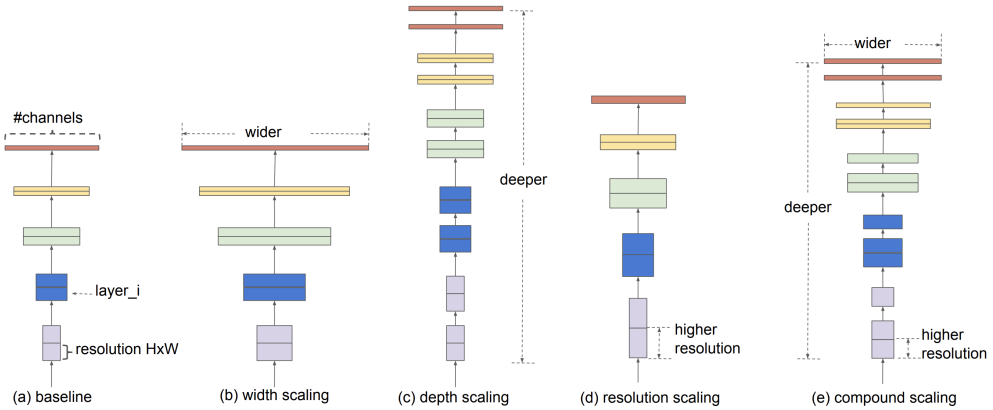


Figure 4: The model scaling methods [16]. The compound scaling (e) systematically scales up the width, depth, and resolution dimension with a fixed ratio.

strictly related to the computational demand of our solution. EfficientNet studies a principled way to scale up the depth, width, and resolution of Convolutional Neural Networks. The compound scaling method, as shown in Figure 4, introduced in this work is the key idea of EfficientNet, which balances the up-sampling of width/depth/resolution uniformly by scaling them with a constant ratio. To further improve the effectiveness, they introduce the EfficientNet architecture using the mobile inverted bottleneck convolution (MBConv) [14] that was developed by AutoML MNAS framework[8]. In this project, we will adopt EfficientNet-B7, a scaled-up version of EfficientNet-B0 (as shown in Figure 5) that demonstrated superior performance with fewer parameter new on ImageNet (Figure 6).

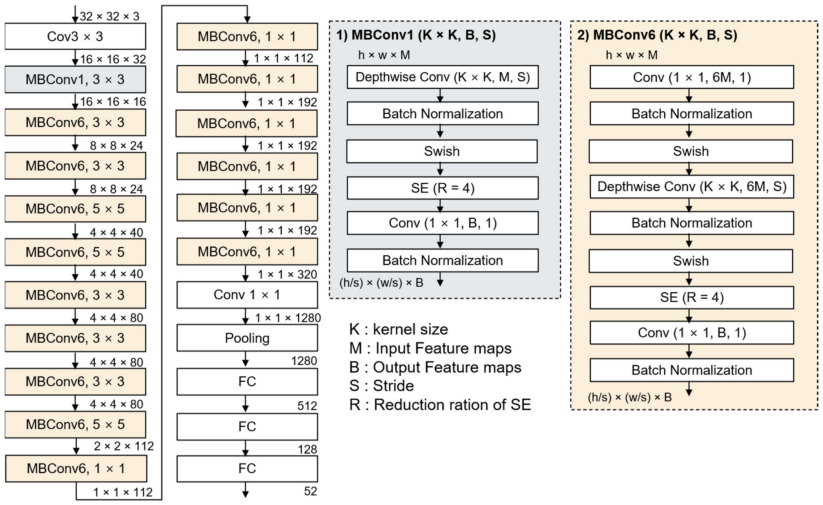
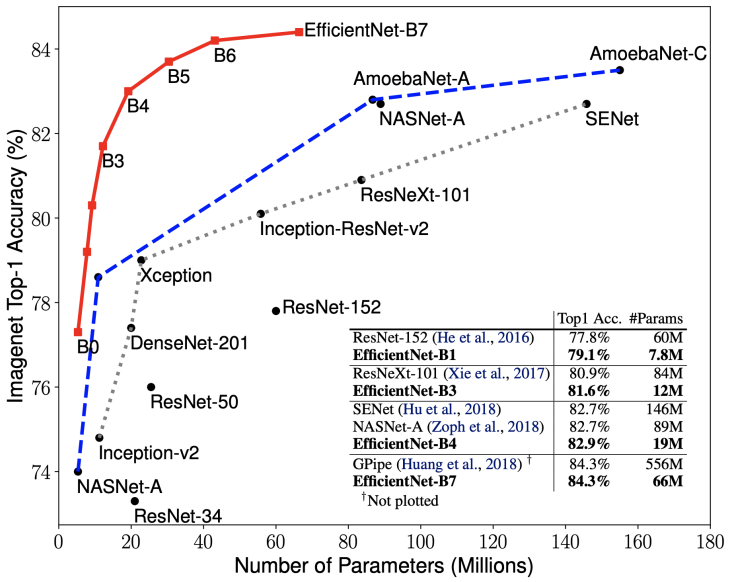


Figure 5: The architecture of EfficientNet-B0 with MBConv1 and MBConv6 illustrated by [5]

2.3 DenseNet

Densenet actually has a lot of applications in image classifications. In the paper Densely Connected Convolutional Networks(https://arxiv.org/pdf/1608.06993.pdf), the authors made extensive comparisons for ImageNet between DenseNets and ResNets. In most cases, the Densenet has a better performance. So that is why I thought of this idea.



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

†Not plotted

Figure 6: EfficientNet-B7 demonstrates superior performance (84.3% top-1 accuracy) with 8.4x lesser parameters and 6.1x faster than GPipe in 2020.

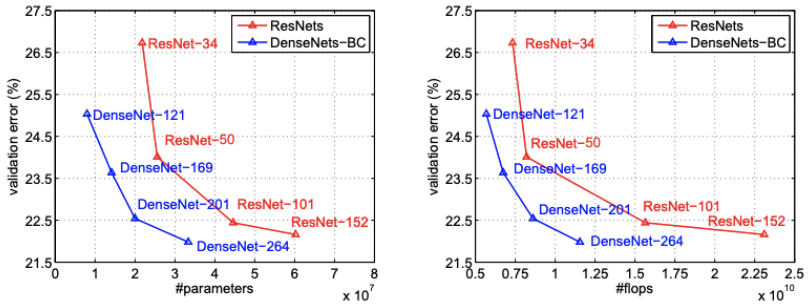


Figure 3: Comparison of the DenseNets and ResNets top-1 error rates (single-crop testing) on the ImageNet validation dataset as a function of learned parameters (*left*) and FLOPs during test-time (*right*).

3 Methodology

For the original SIIM-FISABIO-RSNA Covid-19 detection challenge, the competition aimed to identify and localize abnormalities in real chest radiographs of patients, as shown in Figure 7. In this work, we focus on multi-label classification (a study-level task on the SIIM-FISABIO-RSNA dataset). In this section, we first briefly formulate the problem, which is described in Sec. 3.1. Then, we introduce our early-stage CNN-based solution, a Multi-modal Convolutional Neural Network architecture in Sec. 3.2. After that, we elaborate our final proposed method, a hybrid ensemble solution with EfficientNet and SWIN transformer for categorizing the chest radiographs of patients in Sec. 3.3.

3.1 Problem Formulation

Considering given chest radiographs x_i, the goal at the study level can be formulated as a multi-class image classification task, which categorizes the X-ray photograph as negative for pneumonia or typical, indeterminate, or atypical for COVID-19. The format for a given label y_i^m is a confidence score $y_i^m \in [0, 1]$ of the true class, for $m \in$ {"negative", "typical", "indeterminate", "atypical"}, as shown in Figure 8. Suppose we use $\mathcal{N}(y_i^m \mid x_i; \phi)$ to represent the deep convolutional model, where ϕ is the model parameters, we can formulate our goal as an optimization problem, which is to maximize

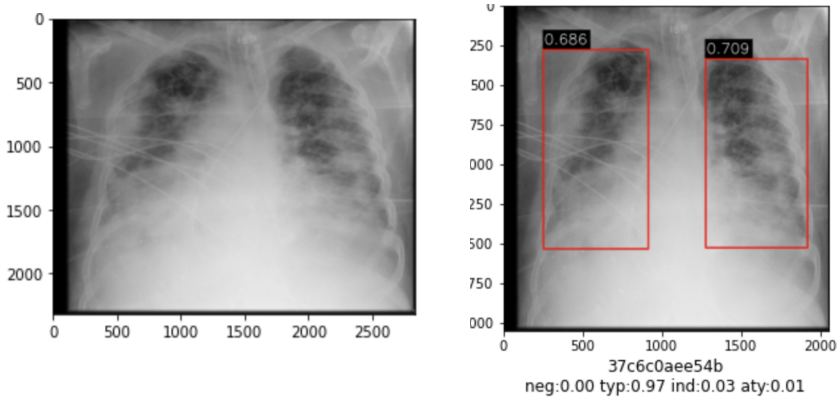


Figure 7: In the original SIIM-FISABIO-RSNA Covid-19 competition, given chest radiographs, the task is to do both multi-label classification at the study level and object detection at the image level.

the model accuracy give for given radiographs:

$$\max_{\phi} \text{Accuracy}(\mathcal{N}(y_i^m \mid x_i; \phi)), \text{ where } x_i \in X, y_i^m \in Y \tag{1}$$

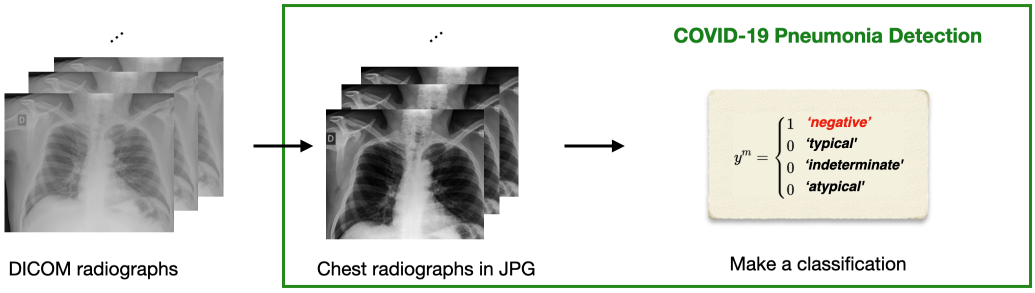


Figure 8: the study-level task is to predict one of the label from 'negative', 'typical', 'indeterminate', or 'atypical' for each DICOM radiograph.

3.2 Early Stage Solution: A Multi-modal Convolutional Neural Network Architecture

Inspired by [13], we discovered that a series of classic deep convolutional networks powered by transfer learning manifests reasonable performance on similar tasks about COVID-19 radiograph detection in previous. Several most used deep convolutional networks are VGG [15]; ResNet [6] and DensNet [10]. Hence, in the first stage of our solution, we set up a Multi-model Convolutional Neural Network architecture, consisting of a ResNet-50V2 classification pipeline and a VGG-16 classification pipeline, as shown in Figure 9. Both pipelines adopted the domain knowledge that was pretrained on ImageNet[2]. Before training, the chest radiograph is pre-processed and resized to 224 x 224 pixels in JPG format. Two identical copies of the same argumented chest radiographs are passed to two classification pipelines separately. The Multi-model ConvNet Architecture is designed by concatenating the extracted features of VGG16 and ResNet-50V2 and feeding forward to a fully connected layer before connecting to softmax.

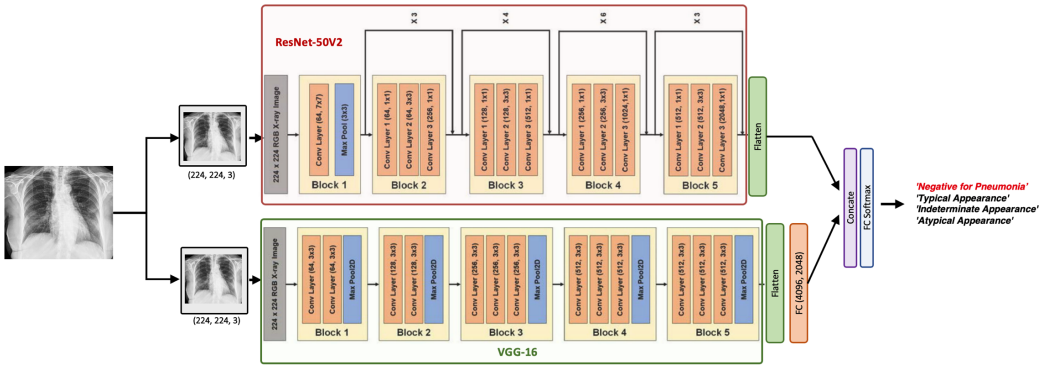


Figure 9: Our early stage approach: A combination approach consisted of VGG16 and ResNet-50V2

However, disturbed by insufficient feature extraction capacity, the early-stage solution does not illustrate a reliable and desirable performance. This observation inspired our team to keep improving our model solution and motivated us to conduct our final approach, which is discussed in the following section.

3.3 Our Final Approach: A Hybrid Ensemble Solution

During the training of our early-stage solution, one of the improvement directions that we discovered is to maintain a reasonably large input feature size and scale up the deep convolutional neural networks to capture more fine-grained features. The main reason is owed to the average size of raw DCM radiographs, which is around 2240x2240 pixels to 4240x4240 pixels. Resizing the chest X-ray imagery to a roughly small resolution (e.g. 224x224 pixels) can make it suffer quality loss and affect the feature extraction of models. Motivated by this, we rethink our model architecture and conduct our final approach: a hybrid ensemble solution based on the Swin transformer and EfficientNet. As shown in Figure 10, the overview of the approach can be regarded as a combination of three separate pipelines: EfficientNet-B3, EfficientNet-B5, and Swin transformer. We inherit the combination methodology that we adopted in our early-stage solution, which averages the confidential score laid by three different model pipelines and feeds it into a fully connected softmax layer to make the final classification. Leveraged by domain knowledge learned from the transfer learning, every model adopted in each pipeline is pretrained on the ImageNet and CheXpert datasets. Inspired by the work proposed by [1], we also use pseudo-random test generation, where retain the image that meets the condition: negative prediction smaller than 0.2 and maximum prediction of the rest class greater than 0.8. The prediction score of these radiographs on the test set is used as a soft-label ground truth in the following training.[19]

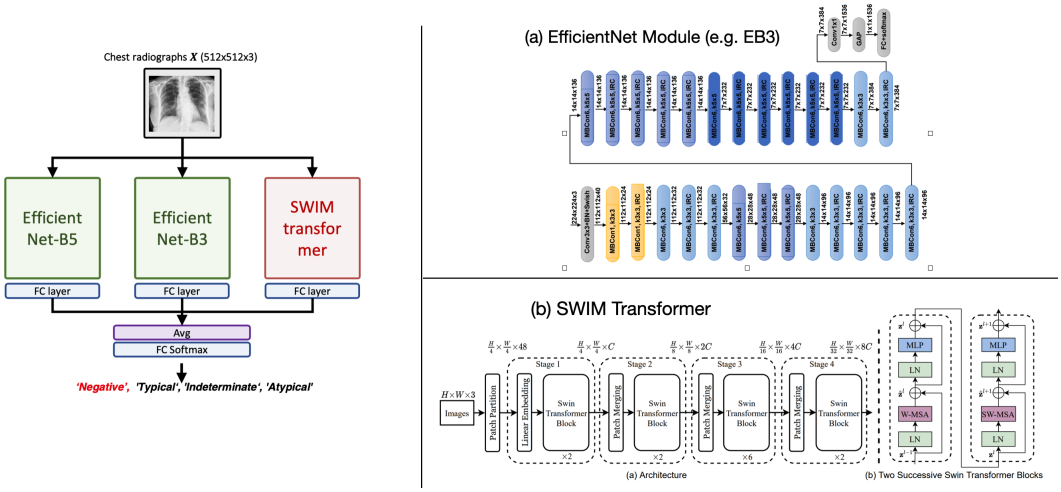


Figure 10: Our final approach: A hybrid ensemble solution based on the Swin transformer and EfficientNet.

4 Experiment

In this section, we evaluate our proposed hybrid ensemble solution under various settings on the SIIM COVID-19 dataset. The data distribution of the SIIM dataset and the preprocessing techniques we employ will also be elaborated on separately. Then, extra qualitative analysis is conducted in this section to evaluate the effectiveness of each proposed component in this study.

4.1 SIIM-FISABIO-RSNA COVID-19 Dataset

SIIM-FISABIO-RSNA COVID-19 dataset contains around 6000+ high-quality chest radiographs among four different classes in Digital Imaging and Communications in Medicine (DICOM) format. Carefully reviewing the training set in SIIM dataset, we discovered that the data distribution for each class is variant: 48.8% labeled as typical, 27.7% labeled as negative, 17.7% labeled as indeterminate, and 6.6% labeled as atypical.

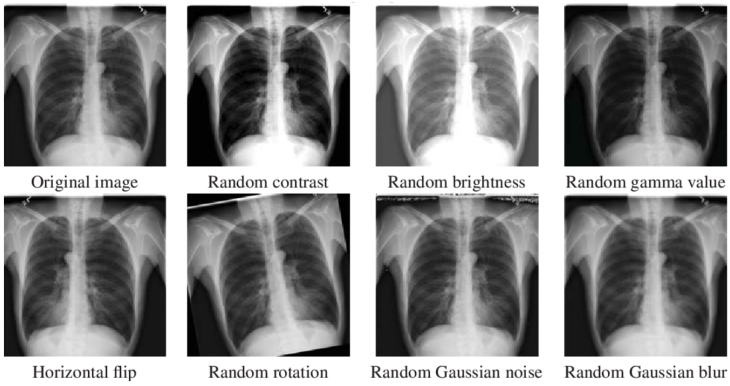


Figure 11: The data argumentation methods we adopted during the training of our final approach.

4.2 Pre-processing

For the data preprocessing, we follow the previous preprocessing standard adopted by [4]. We first transform all raw radiographs from DCM format to JPG format, then remove 107 duplicated chest X-ray imagery that was identified by competitors in the SIIM COVID Kaggle competition last year. As the data partitioning, we split the training, validation, and test sets with a ratio of 90-5-5%. After

that, as shown in Figure 11, we map the radiographs through a series of data augmentations to each partition, including HorizontalFlip, VerticalFlip, ShiftScaleRotate, RandomResizedCrop, Blur, IAASharpen, IAAEmboss, RandomBrightnessContrast, and Cutout.

4.3 Results

The three tables below show our experiment results under various methods, as shown in Table 123. Our first series of experiment focus on finding the best size of the input image. As clearly indicated, EfficientNet-B3 works best with the input image of size 512×512, and the accuracy reaches 0.682. The smaller size may lead to the loss of information and the larger size may blur the model's learning focus.

Methods	***SIIM COVID-19*** Acc.
EfficientNet-B3 (128x128)	0.591
EfficientNet-B3 (256x256)	0.665
EfficientNet-B3 (512x512)	**0.682**
EfficientNet-B3 (1024x1024)	0.680

Table 1: Result of EfficientNet-B3 under various input image size

Our second series of experiments focuses on the best optimizer for the model. We have tried SGD (Stochastic Gradient Descent) optimizer, Adam (Adaptive Moment Estimation) optimizer, and MADGRAD (Momentumized, Adaptive, Dual Averaged Gradient) optimizer. It turns out MADGRAD helped the model reach the best accuracy with our test data and the accuracy reached 0.68.

Methods	***SIIM COVID-19*** Acc.
EfficientNet-B3 (512x512) + SGD	0.591
EfficientNet-B3 (512x512) + Adam	0.665
EfficientNet-B3 (512x512) + MADGRAD	**0.680**

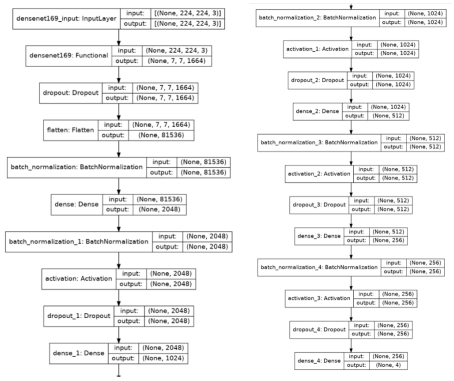
Table 2: Result of EfficientNet-B3(512x512) under various optimizers

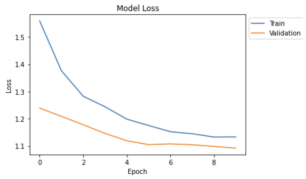
Our final series of experiments focus on the performance of the different model. We have adopted a neural network that combined ResNet-50 with VGG-16, EfficientNet-B3 model, EfficientNet-B5 model, and SWIN (Shifted Window) Transformer. Also, we utilized an ensemble model to aggregate the power of different models. In the end, the ensemble model gives out our best accuracy and that's 0.739.

Methods	***SIIM COVID-19*** Acc.
ResNet-50 + VGG-16	0.495
EfficientNet-B3	0.682
EfficientNet-B5	0.717
SWIN Transformer	0.694
Ensemble	**0.739**

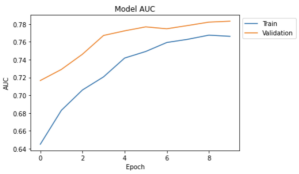
Table 3: Our proposed hybrid ensemble solution integrated with EfficientNet-B3, EfficientNet-B5, and Swin transformer manifest superior performance among the other alternatives, under the unified training setting.

For densenet we tried different fine-tune techniques, such as different densenet versions, unfreeze the layers, use "normal" as the kernel initializer, data rotations, use sigmoid as activation function, half lr and double epochs. But adding two more layers and making it deeper seems to be most useful.





add more layers



5 Conclusion

This final project report presents our project motivation, methodology, experiment process, and final result. Our final model construction is composed of three pipelines which contain Efficient Net-B5, Efficient Net-B3, and Swin transformer respectively. This model takes a chest radiograph image and predicts the patient's COVID-19 illness condition within four classes: Negative, Typical, Indeterminate, and Atypical. Based on our accuracy data in the experiment, the accuracy could reach around 74%, which could be a beneficial aid for the Doctor to make the inspection decision.

References

[1] GitHub - dungnb1333/SIIM-COVID19-Detection: 1st place solution for SIIM-FISABIO-RSNA COVID-19 Detection Challenge — github.com. `https://github.com/dungnb1333/SIIM-COVID19-Detection`. [Accessed 02-Nov-2022].

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2010.11929*, 2010.

[4] dungnb1333. Dungnb1333/siim-covid19-detection: 1st place solution for siim-fisabio-rsna covid-19 detection challenge.

[5] Sumyung Gang, Ndayishimiye Fabrice, Daewon Chung, and Joonjae Lee. Character recognition of components mounted on printed circuit board using deep learning. *Sensors*, 21:2921, 04 2021.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *CoRR*, abs/1908.00709, 2019.

[9] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3464–3473, 2019.

[10] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[11] Yanping Huang, Yonglong Cheng, Dehao Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, and Zhifeng Chen. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *CoRR*, abs/1811.06965, 2018.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[13] Mohammad Rahimzadeh and Abolfazl Attar. A modified deep convolutional neural network for detecting covid-19 and pneumonia from chest x-ray images based on the concatenation of xception and resnet50v2. *Informatics in Medicine Unlocked*, 19:100360, 2020.

[14] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[16] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019.

[17] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

[18] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[19] Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, Marisa Caparrós, Germán González, and Jose María Salinas. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients, 2020.