Vision Transformer based COVID-19 Detection using Chest X-rays

Koushik Sivarama Krishnan
Email: koushik.nov01@gmail.com

Karthik Sivarama Krishnan
Email: ks7585@rit.edu

Abstract - **COVID-19 is a global pandemic, and detecting them is a momentous task for medical professionals today due to its rapid mutations. Current methods of examining chest X-rays and CT scan requires profound knowledge and are time consuming, which suggests that it shrinks the precious time of medical practitioners when people's lives are at stake. This study tries to assist this process by achieving state-of-the-art performance in classifying chest X-rays by fine-tuning Vision Transformer(ViT). The proposed approach uses pretrained models, fine-tuned for detecting the presence of COVID-19 disease on chest X-rays. This approach achieves an accuracy score of 97.61%, precision score of 95.34%, recall score of 93.84% and, f1-score of 94.58%. This result signifies the performance of transformer-based models on chest X-ray.**

Keywords—COVID-19; Chest X-rays; Vision Transformer (ViT); Transfer learning;

I. INTRODUCTION

The term Global Pandemic applies to an epidemic of an infectious disease that is spread all across the globe. The world has not seen a global pandemic after the Spanish flu up until the year 2019. In late 2019, a new strain of corona virus disease was found in Wuhan district of Hubei Province in China. Since this strain of coronavirus disease was not previously identified in human beings, World Health Organization named this strain as "Novel" coronavirus disease. The Novel Coronavirus Disease of 2019 was officially given an abbreviated name of COVID-19 by the World Health Organization, where the acronym CO stands for **Co**rona, VI stands for **Vi**rus, D stands for **D**isease followed by the year 2019 represented as **19**.

This disease abruptly spread all across the globe affecting millions of people and created global tension across all countries. Covid-19 is caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [17]. This virus reside within the same clan of Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS).The commonly identified symptoms for this disease include coughing, shortness of breath, fever, pneumonia, and severe respiratory distress. The symptomatology of this disease is so diverse and are usually identified within day one to day fourteen after the person is being exposed to this contagious virus.

After the initial identification of the presence of covid 19 disease in China, the cases started expanding exponentially, affecting millions of people. This sudden burst of cases, forced all countries to take severe measures to keep the outbreak in control. Figure 1 shows the chart of new cases from January 2020 to July 2021 published in World Health Organization website.

With an exponential increase in covid 19 cases,there is a high demand for rapid testing. Baseline gold standard approach used to test this coronavirus disease 2019 is reverse transcription polymerase chain reaction (RT-PCR). The results for this test takes somewhere
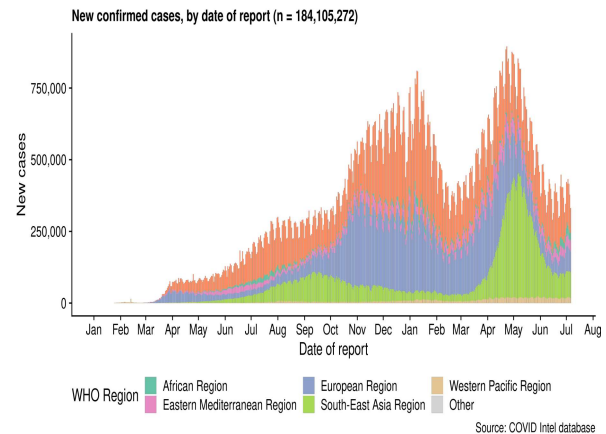

Fig. 1. COVID-19 cases worldwide published by WHO [1]

between 48 to 72 hours of wait time for people to know if they have been affected by this contagious virus or not. Other testing methods include Antigen testing for coronavirus disease. The results of this test usually takes around 15 minutes to 30 minutes of wait time for people to know if they have been exposed to the contagious virus or not. Typically, patients experiencing symptoms of this coronavirus disease 2019 are asked to take Chest CT scans or X-ray scans. Then a radiologist expert examines the scans to confirm if there is a presence of this disease in the chest area. This process is time consuming and is hardly inefficient considering the rise in cases every minute. Rapid and rigorous testing is required in-order to identify as much as cases possible and prevent them from spreading this virus to others. This calls for the need of artificial intelligence to help identify covid-19 patients with minimum time consumption and maximum efficiency.

Artificial Intelligence and Machine learning/Deep learning are currently the state of the art in predicting results in almost all fields. The AI based systems are showing massive results in healthcare domain and this would be helpful in mitigating the time delay for covid-19 virus identification. The vision transformer is the latest state-of-the-art deep learning architecture as it achieves similar performance in benchmark datasets compared to CNN-based models and sometimes even outperforms them in some tasks.

II. RELATED WORK

Deep Learning has become a great area of interest for researchers in recent days. Many have successfully applied deep learning / machine learning techniques [6] for medical classification and obtained promising results. The transfer learning-based approaches have shown remarkable performance boosts in almost all Computer Vision and Natural Language Processing tasks. This approach saves a lot of computational time and can be trained with a fraction of the data.

A. *Transfer Learning*

Transfer Learning has become a common practice in almost all Natural Language Processing and Computer Vision tasks, given the enormous computational power required to train them. This approach is a two-step process. At first, you train the model on large volumes of benchmark datasets. Then you fine-tune them by freezing the initial layers and replacing the last layer based on the task at hand, thus using the previously learnt features for representation. This approach requires a fraction of the data required to train a model from scratch and can be trained in a shorter duration, with less computational resources.

B. *Transformers*

The Transformer architecture[7] was initially proposed in 2017 for neural machine translation. This opened up a wide range of new possibilities in the domain of Natural Language Processing. Since then, several variants and advancements were proposed in transformer-based architectures and all of them are generally pretrained on a large corpus of text data. This architecture proved that combining self-attention with linear layers outperformed the traditional sequence-to-sequence LSTM-RNN based approaches in Neural Machine Translation and other Natural Language Processing tasks.

C. *Vision Transformers*

Since the introduction of transformer architecture[7], many researchers have tried to apply them to Computer Vision tasks. Many even added CNN before self-attention to extract features from images, but most of these approaches could not be applied to real-world Computer Vision problems. The successful utilization of transformer architecture for images was proposed by Dosovitskiy et al. [18]. They managed to get higher accuracy with less computational time for training. This approach also has a much less image-specific bias when compared to CNN-based models. In this approach, they split the images into small patches and stacked them linearly to pass them as input to the transformer model.

D. *Transfer learning based approaches*

Talha Anwar and others. [12] used the pretrained EfficientNet family of architectures to detect COVID-19 in CT scans and managed to get an accuracy score of 89.7%, an F1 score of 89.6%, and an AUC score of 89.5%.

The authors, Rahaman and others. [16] have developed a CAD system for detecting COVID-19 X-rays from others by fine-tuning the pretrained VGG19 model. They managed to obtain a precision score of 97.5% and a recall score of 82.0%. This study experimented with various pretrained CNN-based models and found VGG19 outperforming other latest models.

Sethy et al.[21] suggested a classification model that combined both pretrained ResNet50 and SVM. Here, the author used the Convolutional Neural Network to extract feature from images. This extraced features is then passed into SVM. They showed that this approach is superior to using just a CNN-based network. They achieved an FPR of 95.38%, F1 score of 95.52%, MCC score of 91.41%, and Kappa score of 90.76% for detecting COVID-19 in X-ray images. The authors of [11] also used convolutional neural networks as a feature extraction technique.

III. METHODOLOGY

A. *Dataset*

To test how good our proposed method is, we used chest X-rays of COVID19 patients, healthy patients, patients with opaque lungs, and patients with viral pneumonia. We collected images from two publicly available data sets.

COVID19 X-ray database dataset [14] [19]. This was created by a team of researchers from various universities, in cooperation with doctors from Malaysia. They examined various X-rays and classified them into COVID-19, Viral Pneumonia, Lung Opacity and Healthy chest X-rays.

The COVID19 Pneumonia Normal Chest X-ray PA Dataset from Kaggle was used as the test set.

COVID-19 image from the COVID19 X-ray database[14] [19] was originally around 3500. We upsampled it to 7,000 by using various image augmentation techniques. We then took equal samples of images from Viral Pneumonia, Lung Opacity and Healthy chest X-rays for the Non-COVID class.

I Dataset Split

Dataset	COVID	NON-COVID	TOTAL
Train	6880	6980	13,860
Validation	350	369	719
Test	2313	2313	4626
Total	9443	9662	19,105

B. *Dataset Preprocessing*

The COVID-19 images are up-sampled using image augmentation techniques with the help of albumentations [13] library. The chest X-ray is flipped vertically or horizontally, then randomly rotated to the limit of 270 degrees with constant edges, and the brightness and contrast are arbitrarily adjusted to the limit of 0.4.

chest X-rays are of different sizes from 447 × 530 to 4200 × 3290 pixels. Therefore, we made the target size of images as 224 x 244 pixels. As the models were pretrained on RGB images, we created fake RGB images by stacking the channel over itself. Contrast Limited Adaptive Histogram Equalization (CLAHE) [13] is used as an image enhancement method. CLAHE is a revision of Adaptive
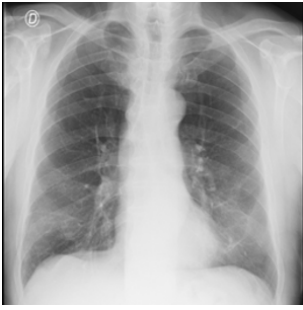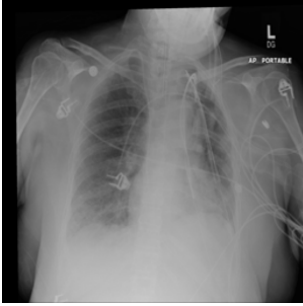
Fig. 2.   (a) COVID
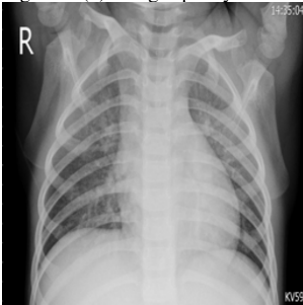

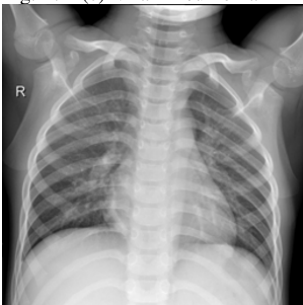Fig. 3.   (b) Lung Opacity

Fig. 4.   (c) Viral Pneumonia

Fig. 5.   (d) Normal

Histogram Equalization (AHE) that avoids excessive contrast enhancement in the image.

IV.   EXPERIMENTS AND ANALYSIS

A. Experimental Setup

In this research, we have used DenseNet[8], InceptionV3 [4], WideResNet101[5] and, Vision Transformer(ViT-B/32)[18] models in the transfer learning process, where the

model is pre-trained on the ImageNet [2] dataset. The bottom layer of the architecture is frozen, as it only extracts general features. We replaced the top layer of the model so that it returns dataset-specific output from the linear layer to match our dataset.

In ViT-B/32[18] pretrained model, the images are split into 32x32 patches and passed into an embedding layer to get the patch embeddings. As the transformers are permutationally invariant, 1-D position embeddings extracted from the images are added to the patch embeddings. This resultant embedding vector is then fed into the encoder part of ViT[18]. The encoder of ViT[18] consists of alternating Layer Normalization, Multiheaded Self-Attention and, MLP[20] blocks. Skip connections are applied after every block.
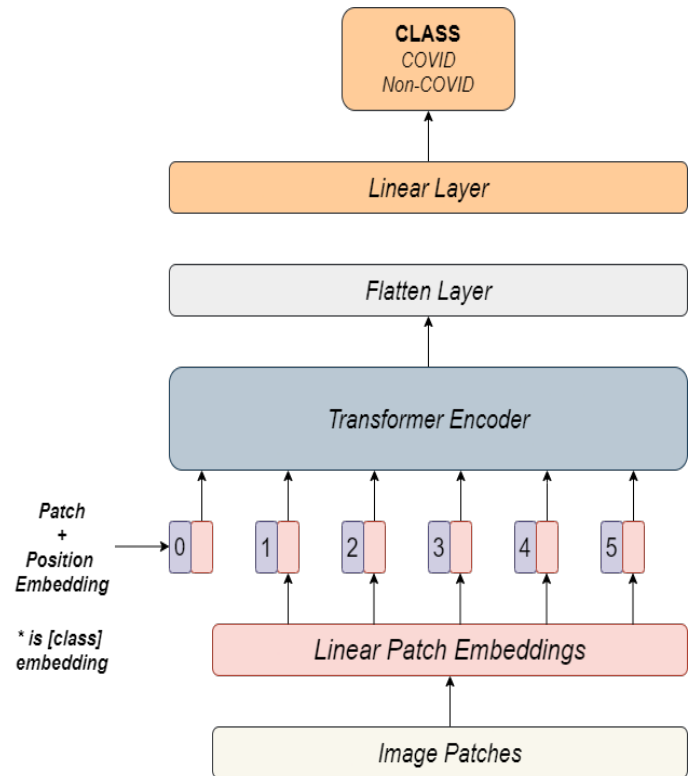

Fig. 6.   Model Architecture

The logits from the linear layer pass through a sigmoid activation function, which returns the predicted probabilities. The target size of the image was set as 224 x 224 x 3 for all models, with an initial learning rate of 1e-4. We used Adam [3] optimizer for CNN based models and RectifiedAdam [15] optimizer for ViT-B/32[18] with the "ReduceLROnPlateau" and "EarlyStopping" schedulers. The learning rate and optimizer were found using hyperparameter optimization framework Optuna [10].

We used Optuna[10] for 50 trials, using various optimizers and learning rates from 1e-3 to 1e-6 and found Adam optimizer works well for all pre-trained CNN models plus 1e-4 as the best initial learning rate. We used the RectifiedAdam[15] optimizer as it showed considerable performance improvement over vanilla Adam[3] optimizer for the ViT[18] pre-trained model. ReduceLROnPlateau scheduler monitors the validation

646

accuracy and reduces the learning rate by a factor of 0.2 if it does not improve for more than three epochs. EarlyStopping scheduler monitors validation accuracy and stops training if the monitored metric does not improve.

InceptionV3[4], DenseNet[8] and, WideResNet101[5] were trained on Google Colaboratory with a batch size of 32 for 25 epochs. ViT-B/32[18] model was trained on NVIDIA RTX 2060 mobile GPU with a batch size of 16 and was early stopped at an epoch of 10.

B. Evaluation Criteria

In this study, we used Accuracy, Precision, Recall, and f1-score as scoring metrics. The accuracy gives a basic idea of the model's performance and is simply a ratio of the correctly predicted observation to the overall observations. The ratio of the correctly predicted positive output and the entire predicted positive output is known as precision score. The recall score is the ratio of the correctly predicted positive output and the total output of the actual positive observation. The weighted average of precision score and recall score is referred to as the F1 score, where the maximum value is 1 and the minimum value is 0.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (4)$$

V. RESULTS

To derive conclusions, we compared the accuracy, precision, recall and F1-score of each model.

II Performance Analysis Table

Model	A.Precision	A.Recall	A.F1	ACC(%)
InceptionV3	0.875	0.870	0.872	0.88
DenseNet	0.910	0.922	0.915	0.92
WideResNet101	0.865	0.855	0.845	0.85
ViT-B/32	**0.953**	**0.938**	**0.946**	**0.976**

The above table illustrates the performance of different deep learning models for detection of COVID-19 from chest X-rays. The test accuracy gives us a general idea about the model's real world performance.

It can be observed from the above table that the highest avg. precision, avg. recall, avg. F1-score and, test accuracy of 95.3%, 93.8%, 94.6%, 97.6% respectively was achieved by ViT-B/32[18] model, followed by DenseNet[8] with value of 91.0%, 92.2%, 91.5% and, 92% respectively. The WideResNet101[5] model attained the lowest avg. precision, avg. recall, avg. F1-score and, test accuracy of 86.5%, 85.5%,

84.5%, 85.0% respectively.

Among all the tested models, InceptionV3[4] and WideResNet101[5] attained the least accuracy of 88% and 85% respectively, while ViT-B/32[18] and DenseNet[8] were the best performers with accuracy of 97.6% and 92%. The ViT[18] model took around 10 minutes to train while other CNN based models took around 35 minutes. This proves that transformer based approaches consume less time when compared to CNN based approaches. Thus ViT[18] baseline model with 32x32 as input patch size is selected as the best model for classifying COVID-19 X-ray images.

VI. CONCLUSION AND FUTURE WORK

The COVID-19 viral infection is a global pandemic and is rapidly mutating. The lungs of the infected people are inflamed due to this viral infection. Hence, examining chest X-rays is one of the possible approaches in the detection of COVID-19. In this study, we have proposed an automated and accurate technique for distinguishing COVID-19 cases from Viral Pneumonia, Lung Opacity, and Healthy chest X-rays. We experimented with four different transfer learning-based architectures and, their performance is evaluated based on four performance metrics. The results of ViT-B/32[18] confirm that transformer-based models are on par with professional radiologists.

Though this model achieves leading results in classifying COVID-19 chest X-rays, there is still scope for development. Noise is a key factor in radiography that affects the model's performance. Applying Generative Adversarial Network(GAN) based noise reduction [9] techniques on the dataset can greatly improve the performance of our model. An ensemble learning-based approach can enhance the final performance. Using a large version of ViT[18] with a larger dataset can surely reach higher performance metrics. We are also planning to extend our work to the segmentation of COVID-19 chest X-rays & CT scans to give even more insights for the radiologists.

REFERENCES

[1] WHO COVID. *Explorer. Geneva: World Health Organization. 2020.* 19.

[2] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition.* 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[3] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization.* cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. 2014. URL: http://arxiv.org/abs/1412.6980.

[4] Christian Szegedy et al. *Rethinking the Inception Architecture for Computer Vision.* 2015. arXiv: 1512.00567 [cs.CV].

[5] Sergey Zagoruyko and Nikos Komodakis. "Wide Residual Networks". In: *CoRR* abs/1605.07146 (2016). arXiv: 1605.07146. URL: http://arxiv.org/abs/1605.07146.

[6] Karthik Sivarama Krishnan et al. "Recognition of human arm gestures using Myo armband for the game of hand cricket". In: *2017 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS)*. 2017, pp. 389–394. DOI: 10.1109/IRIS.2017.8250154.

[7] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[8] Gao Huang et al. *Densely Connected Convolutional Networks*. 2018. arXiv: 1608.06993 [cs.CV].

[9] Yuewen Sun et al. "Digital radiography image denoising using a generative adversarial network". In: *Journal of X-Ray Science and Technology* 26 (2018). 4, pp. 523–534. ISSN: 1095-9114. DOI: 10.3233/XST-17356. URL: https://doi.org/10.3233/XST-17356.

[10] Takuya Akiba et al. *Optuna: A Next-generation Hyperparameter Optimization Framework*. 2019. arXiv: 1907.10902 [cs.LG].

[11] Karthik Sivarama Krishnan and Ferat Sahin. "ORB-DeepOdometry - A Feature-Based Deep Learning Approach to Monocular Visual Odometry". In: *2019 14th Annual Conference System of Systems Engineering (SoSE)*. 2019, pp. 296–301. DOI: 10.1109/SYSOSE.2019.8753848.

[12] Talha Anwar and Seemab Zakir. "Deep learning based diagnosis of COVID-19 using chest CT-scan images". In: *2020 IEEE 23rd International Multitopic Conference (INMIC)*. 2020, pp. 1–5. DOI: 10.1109/INMIC50486.2020.9318212.

[13] Alexander Buslaev et al. "Albumentations: Fast and Flexible Image Augmentations". In: *Information* 11.2 (2020). ISSN: 2078-2489. DOI: 10.3390/info11020125. URL: https://www.mdpi.com/2078-2489/11/2/125.

[14] Muhammad E. H. Chowdhury et al. "Can AI Help in Screening Viral and COVID-19 Pneumonia?" In: *IEEE Access* 8 (2020), pp. 132665–132676. DOI: 10.1109/ACCESS.2020.3010287.

[15] Liyuan Liu et al. *On the Variance of the Adaptive Learning Rate and Beyond*. 2020. arXiv: 1908.03265 [cs.LG].

[16] Md Mamunur Rahaman et al. "Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches". In: *Journal of X-Ray Science and Technology* 28 (2020). 5, pp. 821–839. ISSN: 1095-9114. DOI: 10.3233/XST-200715. URL: https://doi.org/10.3233/XST-200715.

[17] Alexandra C Walls et al. "Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein". In: *Cell* 181.2 (2020), pp. 281–292.

[18] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[19] Tawsifur Rahman et al. "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images". In: *Computers in Biology and Medicine* 132 (2021), p. 104319. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.104319. URL: https://www.sciencedirect.com/science/article/pii/S001048252100113X.

[20] Ilya Tolstikhin et al. "MLP-Mixer: An all-MLP Architecture for Vision". In: *arXiv preprint arXiv:2105.01601* (2021).

[21] Prabira Kumar Sethy and Santi Kumari Behera. "Detection of coronavirus disease (covid-19) based on deep features". In: ().