Interpretable Covid-19 Detection from Chest Radiographs using Vision Transformer

Abstract. The massive insurge of COVID-19 infected patients over the last 2 years has made it difficult to consistently diagnose the disease without manual support. An example of this is Chest X-rays, which is a useful testing method but still requires manual checking. Thus, results are likely to be erroneous as well. In light of these drawbacks and the evolution of technology in data science, we propose a Vision Transformer-based deep learning pipeline for COVID-19 detection from chest X-ray-based imaging. The model achieves an accuracy of 60.21% along with an AUC score of 80.25% in the multi-class classification task. Furthermore, by implementing a Grad-CAM based visualization we make it easier for radiologists to readily interpret and observe the progression of the disease in the affected lungs, assisting healthcare.

Keywords: vision transformer. COVID-19. deep learning. healthcare. localization. grad-CAM

1 Introduction

Since June 2021, the world has seen an astonishing peak of 173 million COVID-19 cases. The shocking numbers only increase day by day without any signs of ever declining [1]. If not detected early, COVID-19 proceeds to develop a flu-like sickness which evolves into acute respiratory distress syndrome (ARDS), which is rather deadly [2, 3]. Despite the recent attempts to detect the disease, with the lack of appropriate resources, as well as limited amounts of accessible data, it has been difficult to strive for an improved diagnosis [4]. So far, the only acceptable method of diagnosis for COVID-19 is RT-PCR, but it is costly, risky to medical staff, and there are few diagnostic test kits available. On the other hand, we have other methods such as medical imaging techniques such as X-ray and CT-based screening. These are relatively safe, faster, and rather easily accessible. X-ray imaging is preferred over CT scans since it is cost effective, time-saving and can be found even in the most remote locations [5]. Due to the complicated structural patterns of lungs which change in degree and appearance over time, the accuracy of chest imaging for a COVID-19 infection detection is heavily dependent on radiological proficiency. However, the reliability of a sophisticated chest examination interpretation is greatly threatened by the lack of skilled radiologists. The implementation of Deep Learning and Data Science in medical fields, such as imaging, shows outstanding performance Thoracic Imaging [6]. In similar fashion, recently there have been several techniques involving CT and X-ray images which use Deep Learning and Data Science for the diagnosis of COVID-19.

This paper presents a novel deep learning pipeline for automatic analysis of COVID-19 using chest radiograph images. The main contributions of this study are as follows:

- We develop a Vit model which outperforms popular CNN methods
- We utilize a very challenging dataset which is similar to the real world data.
- We show the affected regions for better diagnosis.

2 Literature Review

The goal behind COVID-19 image processing is to ascertain the existence of any potentially infectious features such as unilateral or bilateral ground-glass opacities, which are distributed peripherally, and found mostly in round and oval shapes [7, 8].A comprehensive review for machine learning techniques used for COVID-19 detection and classification based on chest radiographs and CT images was provided in [9].

In certain studies, a rather conventional method is followed, which is combining such features with a classifier to deduce the presence of infection. For instance, Mahdy et al. [10] applied a multi-level thresholding for segmenting the X-ray images, which were later classified using a Support Vector Machine (SVM) classifier. Barstugan [11] initially presented a model with SVM-based classification without any feature selection and later with features selected using five feature selection methods. Comparing several different techniques, it was found that the best score was obtained from a grey level size zone matrix feature selector, supported with SVM classification. So far, this literature has cited several deep learning methods for COVID-19 detection via X-ray and CT images. In case of X-rays, Marques et al. presented an EffecientNet pipeline to classify chest X-ray images into the classes COVID-19, normal, or pneumonia after completing 10-fold cross validation [12]. In the study presented by Zabirul Islam et al. we see them merging a convolutional neural network (CNN) and a long short-term memory network to identify features indicating COVID-19 infection in X-ray images [13]. The authors in [14], brought forth a deep network, which is multiscale attention-guided with soft distance regularization to observe and find the symptoms COVID-19 from X-ray images. Their network formulated a prediction vector and attention from multiscale feature maps. In order to amplify the robustness of the model and to enrich the training data, attention-guided augmentations were added accompanying a soft distance regularization. Abbas et al. applied transfer learning from object recognition (i.e. ImageNet dataset) to X-ray images, which can be broken down into three stages. The primary stage is decomposition, wherein class decomposition is applied to AlexNet, to extract deep local features. Secondary stage is the transfer phase, which involves fine-tuning the network weights for X-ray images. The final stage is the composing phase which assembles the subclasses of each class [15]. The dependence of these methods on chest radiographs in the diagnosis decreases the sensitivity of the results for early detection because the sensitivity increases with the progression of the disease [16].

3 Proposed Methodology

We describe our proposed methodology in this section. We discuss about the dataset collection and analysis, data processing, ViT model creation and Grad-CAM visualization.

3.1 Dataset

We use SIIM-FISABIO-RSNA COVID-19 Detection Dataset[17] for this study. The train dataset comprises 6,334 chest scans in DICOM format, which were de-identified to protect patient privacy. All images were labeled by a panel of experienced radiologists for the presence of opacities as well as overall appearance.

3.2 Vision Transformer

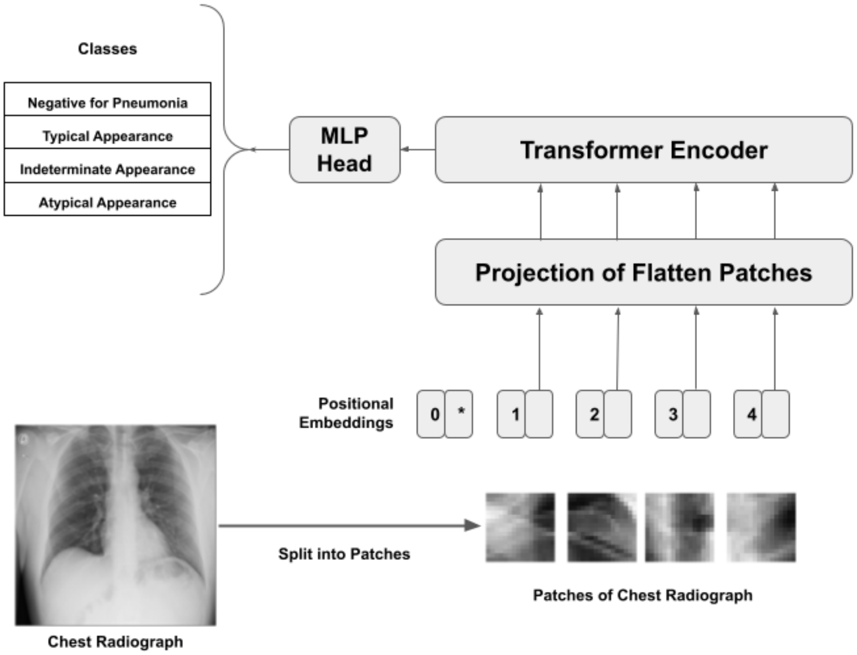


Fig. 1: The proposed ViT model for COVID-19 detection.

Following the success that Transformers achieved in solving the natural language processing problems [18], Dosovitskiy et al. in [19] proceeded to present the Vision Transformer (ViT) model. ViT tries to mirror the original transformer architecture [20] to its best. The study shows that after sufficient training, the ViT can outperform the state-of-the-art CNN with about one-fourth of the computing resources. Our proposed approach utilizes this Vision Transformer model and fine-tunes it on our dataset with a custom MLP block. The preliminary part of the network boasts a Patch Encoder layer that helps in reshaping the input image into multiple flattened patches. Since only the sequential data is compatible with the Transformer encoders, some positional embeddings are also placed along the patches to form a sequence. The Transformer encoder used is the same as [20] and contains multi-headed self attention layers and multiple Multi-layer Perceptron (MLP) blocks. Information is globally integrated throughout the full picture using ViT's self-attention layer; ViT also learns to encode the relative positioning of the patches in order to recreate the visual formation from the training data. Every pixel in the image is used as input and self attention requires every single pixel to pay attention to every other pixel, thereby causing self-attention to have a quadratic cost. This is rather expensive, and it does not scale to a reasonable input size. Thus, the image is separated into patches. Layer Norm is applied before every block to aid in decreasing the training time and refine the generalization performance, since the Layer Norm does not establish any additional dependencies between the training images.The overall architecture has been illustrated in Fig. 1

Table 1: Results obtained from ViT method

Accuracy	Top-2 Accuracy	Loss	AUC Score
0.6021	0.8120	1.10	0.8025

3.3 Grad-CAM

For better visual representation and model interpretability, the Grad CAM Map based illustration was introduced by Selvaraju et al. [21]. is shown in Fig. 6. The said feature is crucial in highlighting certain areas in lungs, which is critical for disease predictions as well as disease development. We obtain the images by passing the output of the embedding layer present in our model at the beginning just after the input layer. Fig. 3 displays the affected areas in different scenarios. The provided figure proves that our suggested methodology recognizes and distinguishes relevant impacted areas from COVID-19 and other pneumonia images. As COVID-19 is more intensive in terms of damage to the lungs, as compared to other types of pneumonia, our model prioritizes this fact by highlighting yellow and red areas in the COVID patient's x-ray image.

4 Experimental Results and Discussion

In this section we discuss about the fine tuning of the ViT model and obtained results of the proposed model.

4.1 Experimentation

We fine tune the model to get the optimum results. We resize the images to 256 x 256 pixels with keeping it RGB (three channel) image. There are 144 patches

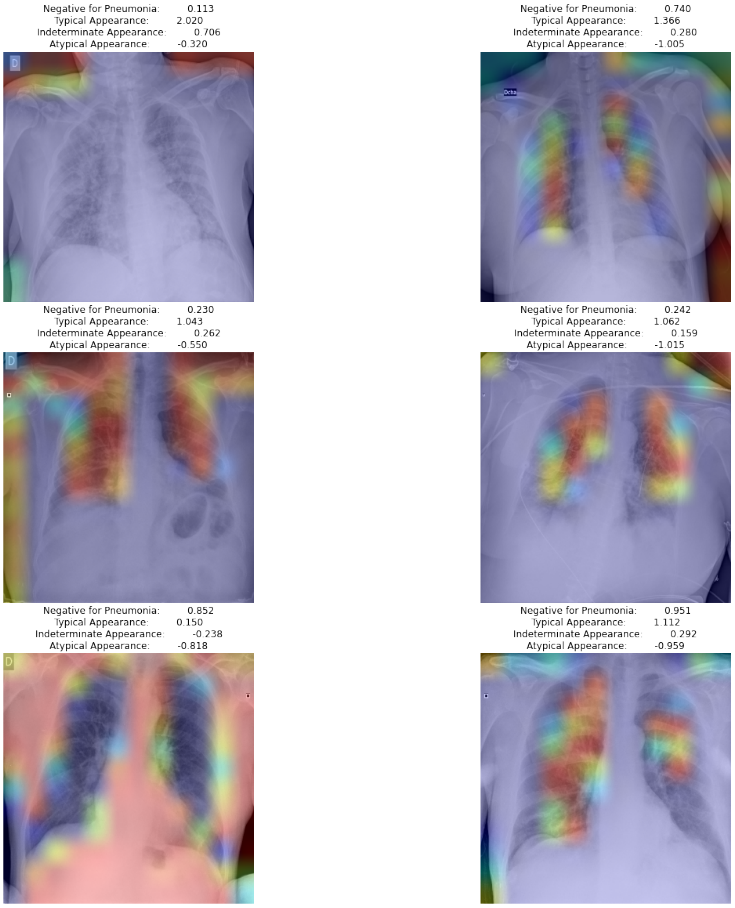

Negative for Pneumonia: 0.113
Typical Appearance: 2.020
Indeterminate Appearance: 0.706
Atypical Appearance: -0.320
D
Negative for Pneumonia: 0.740
Typical Appearance: 1.366
Indeterminate Appearance: 0.280
Atypical Appearance: -1.005
Dcha
Negative for Pneumonia: 0.230
Typical Appearance: 1.043
Indeterminate Appearance: 0.262
Atypical Appearance: -0.550
D
Negative for Pneumonia: 0.242
Typical Appearance: 1.062
Indeterminate Appearance: 0.159
Atypical Appearance: -1.015
Negative for Pneumonia: 0.852
Typical Appearance: 0.150
Indeterminate Appearance: -0.238
Atypical Appearance: -0.818
D
Negative for Pneumonia: 0.951
Typical Appearance: 1.112
Indeterminate Appearance: 0.292
Atypical Appearance: -0.959

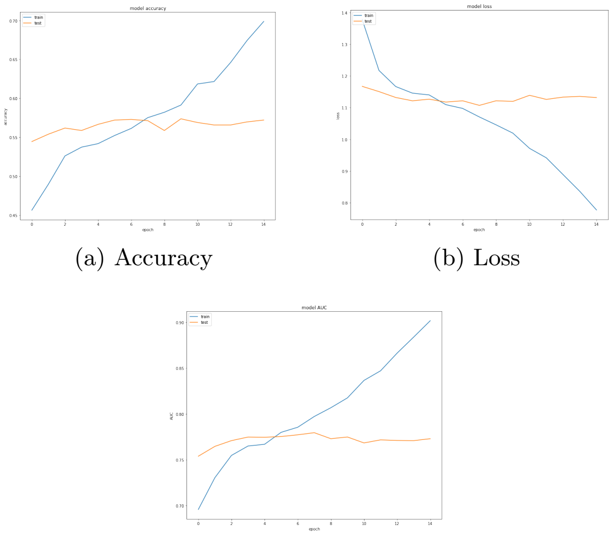
Fig. 2: Highlighted Regions

per image with 20 x 20 pixel size. There are 400 elements per patch. We use 16 training examples per iteration to achieve the optimal result. We train the model with 15 epochs. Our ViT model contains 5 transformer layers which enables the model to learn effectively. We use Adam optimizer with 0.001 learning rate.

4.2 Result Analysis

For our study, we evaluate the model performance using Accuracy, Top-2 Accuracy, Loss, and The Area Under the Curve of the Receiver Operating Characteristic (ROC), also known as AUC of ROC. These metrics give us a quantitative measure to evaluate the proposed ViT model performance with other popular methods.

Our achieved results are shown in Table 1. And in Fig. 3 we have shown the train vs validation curves which represents the learning curve of the model.



(a) Accuracy

(b) Loss

(c) AUC

Fig. 3: Train vs Validation of The Proposed Model

5 Conclusion and Future Work

Our study shows that the model we proposed is promising both in terms of robustness and easy interpretation. A Vision Transformer based architecture is used in this model as it achieves an astonishing accuracy of 98% and an AUC score as high as 99%. To further assure the reliability of our model, an interpretable inference pipeline with Grad-CAM based visualizations per image is added. We believe that our proposed model, which utilizes chest x-ray images can also be used as a cheaper, make-shift, bed-side diagnostic tool for detecting COVID-19. Such a tool is especially useful in areas with a scarcity for rapid testing, and even serve as a secondary diagnostic method following the regular RT-PCR test. Screening a second time may very well help us to verify that any true negative or false positive cases do not occur. Our future work will focus on proposing another variant of the Vision Transformer for further improving the performance, given the availability of larger data sets.

References

1. World-Health-Organization. *COVID-19 weekly epidemiological update* (23, May 2021).
2. Lang, T. Plug COVID-19 research gaps in detection, prevention and care. *Nature* **583,** 333–333 (July 2020).
3. Yang, P. & Wang, X. COVID-19: a new challenge for human beings. *Cellular Molecular Immunology* **17** (Mar. 2020).

4. Ting, D., Carin, L., Dzau, V. & Wong, T.-Y. Digital technology and COVID-19. *Nature Medicine* **26** (Mar. 2020).
5. Nayak, S. R., Nayak, D. R., Sinha, U., Arora, V. & Pachori, R. B. Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study. *Biomedical Signal Processing and Control* **64,** 102365. ISSN: 1746-8094. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7674150/ (2022) (Feb. 2021).
6. Zhou, S. K. *et al.* A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. *Proceedings of the IEEE* **109,** 820–838. https://doi.org/10.1109%2Fjproc.2021.3054390 (May 2021).
7. Kanne, J. *et al.* COVID-19 imaging: What we know now and what remains unknown. *Radiology* **299,** 204522 (Feb. 2021).
8. Schmitt, W. & Marchiori, E. Covid-19: Round and oval areas of ground-glass opacity. *Pulmonology* (Apr. 2020).
9. Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* **3** (Mar. 2021).
10. Hassanien, A. E., Mahdy, L., Ezzat, K., Elmousalami, H. & Aboul-Ella, H. *Automatic X-ray COVID-19 Lung Image Classification System based on Multi-Level Thresholding and Support Vector Machine* Apr. 2020.
11. Barstugan, M., Ozkaya, U. & Ozturk, S. *Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods* 2020. https://arxiv.org/abs/2003.09424.
12. Marques, G., Agarwal, D. & de la Torre Díez, I. Automated Medical Diagnosis of COVID-19 through EfficientNet Convolutional Neural Network. *Appl. Soft Comput.* **96.** ISSN: 1568-4946. https://doi.org/10.1016/j.asoc.2020.106691 (Nov. 2020).
13. Islam, M. Z., Islam, M. M. & Asraf, A. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Informatics in Medicine Unlocked* **20,** 100412. ISSN: 2352-9148. https://www.sciencedirect.com/science/article/pii/S2352914820305621 (2020).
14. Li, J. *et al.* Multiscale Attention Guided Network for COVID-19 Diagnosis Using Chest X-ray Images. *IEEE journal of biomedical and health informatics* **PP** (Feb. 2021).
15. Abbas, A., Abdelsamea, M. M. & Gaber, M. M. *Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network* 2020. https://arxiv.org/abs/2003.13815.
16. Stephanie, S. *et al.* Determinants of Chest Radiography Sensitivity for COVID-19: A Multi-Institutional Study in the United States. *Radiology: Cardiothoracic Imaging* **2.** PMID: 33778628, e200337. eprint: https://doi.org/10.1148/ryct.2020200337. https://doi.org/10.1148/ryct.2020200337 (2020).

17. Lakhani, P. *et al. The 2021 SIIM-FISABIO-RSNA Machine Learning COVID-19 Challenge: Annotation and Standard Exam Classification of COVID-19 Chest Radiographs.* Oct. 2021. osf.io/532ek.
18. Wang, B. *et al. Pre-trained Language Models in Biomedical Domain: A Systematic Survey* 2021. https://arxiv.org/abs/2110.05006.
19. Dosovitskiy, A. *et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* 2020. https://arxiv.org/abs/2010.11929.
20. Vaswani, A. *et al. Attention Is All You Need* 2017. https://arxiv.org/abs/1706.03762.
21. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* **128,** 336–359. https://doi.org/10.1007%2Fs11263-019-01228-7 (Oct. 2019).