Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/41/20230753

Edge impulse-based pretrained neural network for diagnosing COVID-19

Wenbin Gao

The Department of Medicine, University of Leeds, Leeds, LS2 9JT, United Kingdom

ml21w2g@leeds.ac.uk

Abstract. COVID-19 has wreaked havoc on a global scale, primarily owing to its extraordinary contagiousness, thereby straining local healthcare systems to their limits. While the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test is known for its specificity, it suffers from time-consuming procedures and a notable false negative rate. Consequently, there is an immediate imperative for a swift and precise diagnostic approach. This paper introduces a novel concept, employing artificial intelligence, to address these challenges effectively. Specifically, this study employed a transfer learning model provided by the Edge Impulse platform and a dataset containing chest X-ray images of children. This study pre-processed these images and trained and tested them several times using different image sizes and network architectures. The experimental results show that the model achieves very high accuracy (>99%) with 160*160 image size and version 1.0 or 0.35 of the network architecture. These results clearly support the hypothesis that migration learning can play an important role in the fast and accurate diagnosis of COVID-19 with appropriate image size and network architecture. This research can be used as a way to rapidly train locally adapted AI models to achieve rapid assisted diagnosis of this type of acute infectious disease.

Keywords: COVID-19, Artificial Intelligence, Transfer Learning, Edge Impulse.

1. Introduction

An unknown pneumonia was first reported in December 2019,in the following months similar conditions were reported in most countries around the world. In 2020 this pneumonia was identified as being caused by a type of coronavirus called Severe Acute Respiratory Syndrome Coronavirus Type 2 (SARS-CoV-2), which the World Health Organisation (WHO) named COVID-19 and declared it to constitute a global pandemic [1] Until 13 August 2023, there were more than 769 million confirmed cases and more than 6.9 million deaths globally, according to data compiled by the WHO[2]. Due to the highly contagious nature of COVID-19, especially in the early stages of the disease. Although a highly sensitive test called Reverse Transcription Polymerase Chain Reaction (RT-PCR) is used, hospitals and medical professionals are facing a substantial surge in their workload, owing to the extended 4-18 hours testing duration [3] Furthermore, the potential for initial false-negative diagnoses among COVID-19 patients intensifies the gravity of disease transmission [1,3]. Thus, an imperative exists for the swift development of technology that can provide indispensable support to clinicians in achieving precise diagnoses for COVID-19 cases.

DOI: 10.54254/2755-2721/41/20230753

Since patients with COVID-19 generally have lesions characterised in their lung medical imaging pictures and other examination reports, Artificial Intelligence (AI), as one of the hottest technological topics at present, has a great potential in this aspect of medical diagnosis [1,3]. Nowadays, the use of AI for COVID-19 is mainly researched from two aspects: Machine Learning (ML) and Deep Learning (DL). Machine learning is mainly used for early disease detection and technical support for diagnosis. Some studies have pointed out that the blood characteristics of COVID-19 patients can be classified into three categories by analysing a large amount of clinical data, and algorithms are used to continue the classification. It was possible to distinguish between COVID-19 and influenza cases with a sensitivity and specificity over 90 % [4]. Another study noted that an ML-based algorithm, which can make an initial diagnosis based on a patient's clinical symptoms during a fever clinic visit, was more than 90 % accurate after specifying a number of COVID-19 highly relevant features [5]. The deep learning is used more in the field of imaging as a form of ML, and most of the research has been done on training medical image databases with convolutional neural networks [3]. These imaging databases include X-rays, Computed Tomography (CT), Magnetic Resonance (MRI), and ultrasound. An experiment used lung CT images of 49 patients from a hospital in China, including confirmed and suspected cases, and the AI model obtained through the DL approach had an accuracy of 90% and a specificity of 83% [6]. Although the accuracy is not high, this could prove a great potential in the diagnosis of AI COVID-19.

This paper proposes new ideas for training artificial intelligence models. Due to lack of data and resource constraints, migration learning can use models obtained by training on large-scale datasets as initial models [7]. Reducing the need for large amounts of labelled data and computational resources is achieved by fine-tuning the parameters to suit the new task [7]. Therefore, the Edge Impulse platform is used as a vehicle for transfer learning, where data is fed through edge devices to train AI models that match local realities. Thereby, faster AI models with up-to-date accuracy can be obtained.

2. Methods

2.1. Preparation of the dataset

In this research, a dataset named Chest X-ray (Pneumonia) sourced from Kaggle [reference] was used. The dataset is X-ray images taken from children patients aged 1-5 years old from Guangzhou Women and Children's Medical Centre during their clinical care. The entire dataset consists of 5863 images and is divided into three folders, the training set, the training set and the validation set, and each folder also includes subfolders for COVID-19 and normal. All the images are in JPGE format instead of the original image format DICOM and according to the description of the dataset, the images have been identified and diagnosed by two expert doctors and the validation set has been examined by a third doctor before it was finally approved to be used for training the AI. Therefore, the dataset has high credibility.

2.2. Introduction of Edge Impulse

Today's smartphones and some embedded systems are becoming more powerful, making it possible to perform complex computing (e.g., machine learning) on these edge devices. Edge Impulse is a platform designed to assist developers in deploying machine learning on edge devices. It enables real-time data classification, anomaly detection, and other tasks using machine learning models for tasks that might otherwise require hard coding. Edge Impulse typically provides data collection, data labelling and management, model training, model testing and validation, model deployment, real-time monitoring, and more. It can be used in a wide range of applications in agriculture, manufacturing and healthcare. Edge Impulse user interface as shown in Figure 1.

DOI: 10.54254/2755-2721/41/20230753

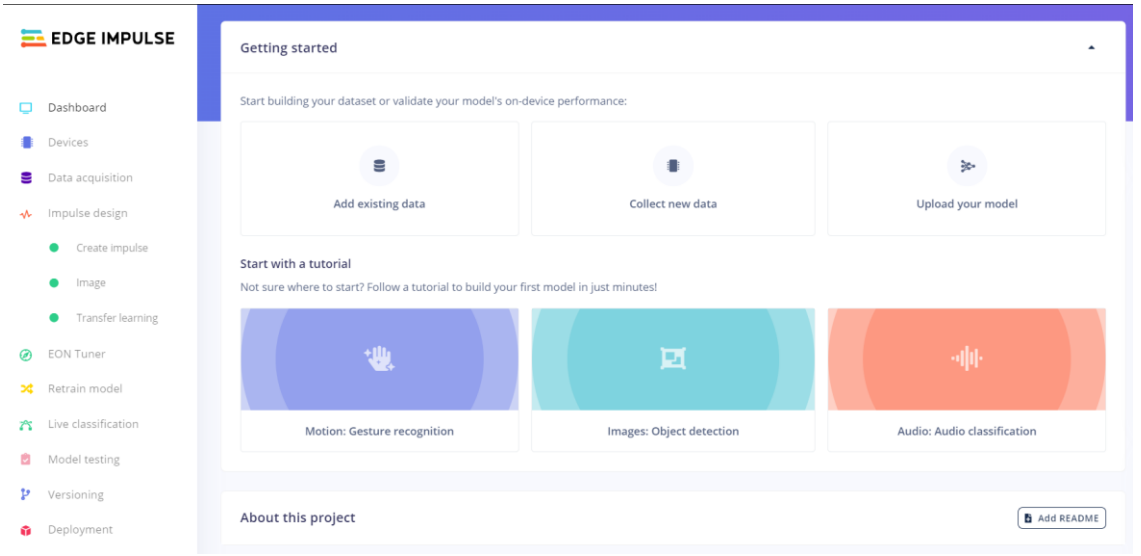

Figure 1. Edge Impulse User Interface (Photo/Picture credit: Original).

2.3. Procedure of realization

Convolutional Neural Network (CNN) is a type of neural network that plays an important role in the field of image recognition and classification. It can not only recognise objects and faces, but also provide visual support for robots and self-driving cars. A typical CNN typically comprises five fundamental components: the convolutional layer, activation function, pooling layer, fully connected layer, and classification layer. The convolutional layer serves as the cornerstone of the CNN, primarily responsible for extracting valuable features through convolutional operations applied to the input data. Following the convolution operation, the activation function comes into play, introducing non-linear characteristics into the system. The pooling layer, on the other hand, censors the image dimensions and retains the maximum feature values of the image. The fully connected layer is usually located at the end and is used to connect each neuron to each neuron in the next layer. The output layer will classify the input image according to the training requirements [8].

Transfer learning is a machine learning method that takes knowledge and features learned from one task and transfers them to another task that is related to it but not exactly the same, which has been widely used in many studies [9, 10]. This reduces a lot of training time and computational resources. This is because when training on data, it takes a lot of computational resources and time to obtain its features. If a well-trained model on a certain task has highly generalised features, then these features can be easily used for multiple tasks that are different but related, thus reducing unnecessary waste of computational resources [7].

2.4. Implementation details

Before conducting this experiment, the authors experimented with Edge Impulse using a validation set to familiarise themselves with the various settings, but found a problem. The limitations of the platform prevented it from processing too much data at the same time, so the authors randomly culled the original dataset, and ultimately retained 2,442 images for subsequent experiments. When the data was uploaded on the platform, the COVID-19 and normal images were already labelled and grouped into the training set, while the test set in the dataset was uploaded and labelled for testing. This was followed by the impulse design phase, this item was designed to pre-process the data, extract image features using signal processing and then classify the new data using the learning module. Firstly the image size was set, the image size was set to 96*96, then the image size mode was adjusted, there are three modes of sizing which are fit to short axis, fit to long axis and squash, these three are common methods used by machine learning to prepare images, since most of the image features of the COVID-19 that need to be recognised

DOI: 10.54254/2755-2721/41/20230753

are in the centre of the image, so the fit to short axis mode was chosen , which is suitable for cases where the aspect ratio needs to be preserved and where the key information is in a more centralised location. The second step is to add the processing module, this study chose to preprocess and normalise the image data, which can be more convenient for the AI to read the data information. The third step is to select the learning module, choose to use migration learning to process the image data, this is because migration learning is suitable for image recognition, and in the case of fewer images can also get very good data. The image preprocessing interface is shown in Figure 2. The fourth step the parameters of the image are adjusted, the main thing is to adjust the colour of the image, there are two options RGB and greyscale. The greyscale mode is chosen because the chest X-ray image is a greyscale image, which may increase the computational complexity and storage space if it is trained as an RGB image. At last, the moment arrived to fine-tune the intricacies of the transfer learning process. Initially, for the neural network configuration, this study established the number of training cycles at 15 and adjusted the learning rate to 0.001 shown in Figure 3. These parameters are better suited for acclimatization operations when using the validation set. Following that, with regard to the neural network architecture, Edge Impulse provides an array of architectures tailored to image sizes of 96*96 and 160*160, each diverging in terms of RAM and ROM utilization and supporting various image color schemes. The 96*96 version 0.25 of the neural network architecture was first selected and then clicked on Start Training to get the accuracy and loss of the training along with the confusion matrix of the model. The model was tested using the test set in this module in Model Testing to get the accuracy of the final model. Transfer learning parameter adjustment as shown in image 3.

The above experiment was done 6 times as the model with the highest fitness needed to be selected. Two experiments were done using 96*96 size, the neural network frameworks selected were version 0.25 and 0.1, the RAM and ROM size of version 0.25 is less than that of version 0.1. 160*160 size images were trained four times and four versions of neural network frameworks were taken which were 1.0,0.75,0.5,0.35. From version 1.0 to 0.35 the RAM and ROM size is reduced sequentially.

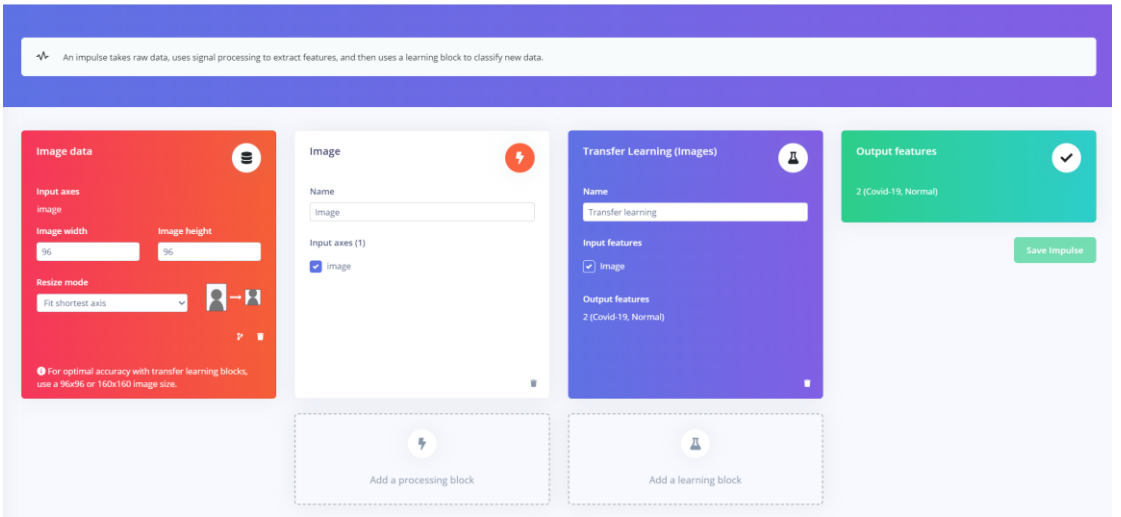

Figure 2. Image Preprocessing Interface (Photo/Picture credit: Original).

DOI: 10.54254/2755-2721/41/20230753

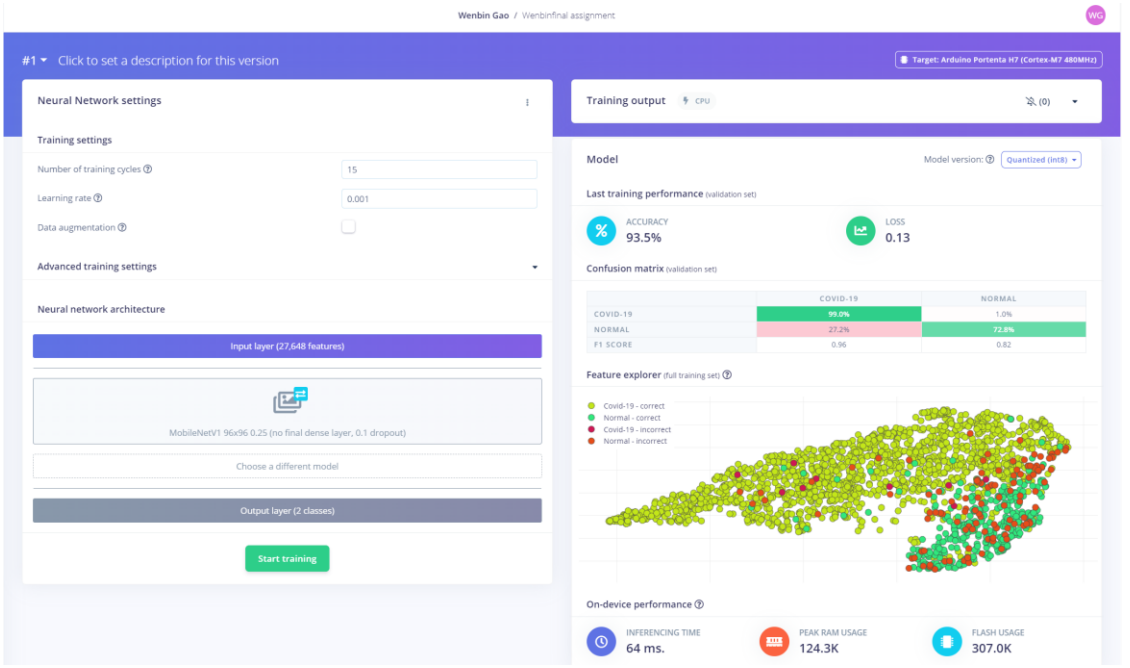

Figure 3. Transfer Learning Parameter Adjustment Interface (Photo/Picture credit: Original).

3. Results and Discussion

The experimental results are shown in Table 1. Firstly, observe the TPR, which is called the true positive rate, also called sensitivity, and it represents the percentage of the model that classifies COVID-19 correctly. It can be seen that all the models have a high value of TPR, which is greater than 95%. The value is further improved when the image size is changed from 96*96 to 160*160. However, it was found that the size 160*160 had the same TPR for version 1.0 as version 0.35 and version 0.75 as version 0.5.FNR stands for False Negative Case Rate, which is the proportion of COVID-19 diagnosed as normal. It can be seen that the FNR is low for all models, which indicates that all models are accurate for COVID-19 recognition with high probability. It is also found that image size 96*96 using version 0.25 and image size 160*160 using versions 0.75 and 0.5 have the same FNR. image size 160*160 using version 1.0 and version 0.35 have the same FNR. tNR stands for True Negative Ratio which is also called as Specificity and indicates the proportion of images that are correctly recognised as normal. It can be seen that using the model with low image size its accuracy in TNR is less. The TNR of the model version 1.0 is the highest when the image size is 160*160 and it is more than 90%.FPR stands for False Positive Rate and the result is opposite to the TNR and the FNR of the model version 1.0 is the lowest when the image size is 160*160.Accuracy stands for the probability of the model to correctly classify the COVID-19 and the normal image, where the image size is 160*160 is over 95% using both version 1.0 and version 0.35.F1 score represents the reconciled average of precision and recall, and it can be clearly seen that the scores have increased when the image size is 160*160, which proves that the model performs best when the image size is 160*160, and the results of version 1.0 and version 0.35 are much closer.

DOI: 10.54254/2755-2721/41/20230753

Table 1. Experimental data.

	TPR	FNR	TNR	FPR	accuracy	F1 score for Covid-19	F1 score for Normal
96 0.25	97.90 %	0.50 %	37.60 %	50.40 %	75.32 %	0.86	0.54
96 0.1	97.90 %	0.80 %	41.90 %	50.90 %	76.92	0.86	0.59
160 1.0	99.70 %	0.30 %	91.10 %	8.90%	98%	0.99	0.95
160 0.75	99.50 %	0.50 %	72.30 %	27.70 %	93.90 %	0.96	0.83
160 0.5	99.50 %	0.50 %	73.30 %	26.70 %	94.10 %	0.96	0.84
160 0.35	99.70 %	0.30 %	82.20 %	17.80 %	96.10 %	0.98	0.90

The provided data unmistakably demonstrates that the image size can exert a significant impact on the model's accuracy. This effect likely arises from the fact that a larger image size provides the model with more information, thereby enhancing its overall accuracy [7]. Regarding the framework version employed, there is limited performance improvement observed between 0.35 and 1.0. This suggests that the model is nearing its upper performance limit with an image size of 160, and further optimization may be required for finer-tuned results. Interestingly, both version 0.35 and 1.0 exhibit similar outcomes, indicating the model's robustness and its ability to maintain consistent performance across varying resource constraints. It also shows a good match between the model and the dataset used, and a side effect of the correctness and importance of transfer learning. It is noted that the model discriminates normal patients poorly compared to COVID-19 patients, presumably due to data imbalance, where there are more images of COVID-19 than of normal people, which makes the model tend to recognise COVID-19[8]. It is also due to the fact that the lung images of normal people are featureless, and that some external factors on the images may be recognised as COVID-19 due to some data bias, factors to be identified as the basis for the judgement of COVID-19, which leads to a higher probability of misdiagnosis of the correct patient [7,8]. Finally, it is possible that the model was trained to better recognise COVID-19 while ignoring the recognition of normal cases.

4. Conclusion

In this work, the authors propose a new idea for training diagnostic COVID-19 AI, using edge devices to upload local COVID-19 image databases, and training locally appropriate COVID-19 diagnostic models by using the transfer learning models provided on the Edge Impulse platform. Experimental results show that the AI trained by this training method has high COVID-19 diagnostic accuracy. This experiment still has limitations, the quality of images within the dataset was not carefully examined and some data bias was ignored, while the number of samples was not large enough, which made the data unbalanced, with the trained model still having flaws. The tuning of the relevant parameters was also not the best. In the future, the authors will pay more attention to the quality of the dataset before the experiment to avoid the problem of model accuracy caused by data bias. The parameters of migration learning should also be adjusted several times to ensure the highest accuracy of the model.

References

[1] Suri J S et al 2021 A narrative review on characterization of acute respiratory distress syndrome in COVID-19-infected lungs using artificial intelligence Computers in Biology and Medicine 130 104210

DOI: 10.54254/2755-2721/41/20230753

[2] World Health Organization 2023 https://covid19.who.int/
[3] Safiabadi T Seyed H et al 2021 Tools and techniques for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)/COVID-19 detection Clinical microbiology reviews 34.3 10-1128.
[4] Kukar M et al 2021 COVID-19 diagnosis by routine blood tests using machine learning Scientific reports 11.1 10738
[5] Domínguez O Juan L et al. 2021 Machine learning applied to clinical laboratory data in Spain for COVID-19 outcome prediction: model development and validation Journal of medical Internet research 23.4 e26211
[6] Li L et al 2020 Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy Radiology 296.2 E65-E71
[7] Singh T et al 2022 Ftl-CoV19: A Transfer Learning Approach to Detect COVID-19 Computational Intelligence and Neuroscience 2022
[8] Maior C B S et al. 2021 Convolutional neural network model based on radiological images to support COVID-19 diagnosis: Evaluating database biases Plos one 16.3 e0247839.
[9] Qiu Y et al 2019 Semantic segmentation of intracranial hemorrhages in head CT scans In 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS) pp 112-115) IEEE
[10] Lehečka J Psutka J V and Psutka J 2023 Transfer Learning of Transformer-based Speech Recognition Models from Czech to Slovak arXiv preprint arXiv:2306.04399