A Computationally-Inexpensive Strategy in CT Image Data Augmentation for Robust Deep Learning Classification in the Early Stages of an Outbreak

Hou, Yikun; Navarro-Cia, Miguel

[Link to publication on Research at Birmingham portal](#)

Endorsed by **IPEM**
Institute of Physics and
Engineering in Medicine

PAPER • OPEN ACCESS

A computationally-inexpensive strategy in CT image data augmentation for robust deep learning classification in the early stages of an outbreak

View the article online for updates and enhancements.

Biomedical Physics & Engineering Express

PAPER

A computationally-inexpensive strategy in CT image data augmentation for robust deep learning classification in the early stages of an outbreak

Yikun Hou[1] and Miguel Navarro-Cía[1,2,*]

1 Department of Electronic, Electrical and Systems Engineering, University of Birmingham, Birmingham B15 2TT, United Kingdom
2 School of Physics and Astronomy, University of Birmingham, Birmingham B15 2TT, United Kingdom
* Author to whom any correspondence should be addressed.

E-mail: yxh060@alumni.bham.ac.uk and m.navarro-cia@bham.ac.uk

Abstract

Coronavirus disease 2019 (COVID-19) has spread globally for over three years, and chest computed tomography (CT) has been used to diagnose COVID-19 and identify lung damage in COVID-19 patients. Given its widespread, CT will remain a common diagnostic tool in future pandemics, but its effectiveness at the beginning of any pandemic will depend strongly on the ability to classify CT scans quickly and correctly when only limited resources are available, as it will happen inevitably again in future pandemics. Here, we resort into the transfer learning procedure and limited hyperparameters to use as few computing resources as possible for COVID-19 CT images classification. Advanced Normalisation Tools (ANTs) are used to synthesise images as augmented/independent data and trained on EfficientNet to investigate the effect of synthetic images. On the COVID-CT dataset, classification accuracy increases from 91.15% to 95.50% and Area Under the Receiver Operating Characteristic (AUC) from 96.40% to 98.54%. We also customise a small dataset to simulate data collected in the early stages of the outbreak and report an improvement in accuracy from 85.95% to 94.32% and AUC from 93.21% to 98.61%. This study provides a feasible Low-Threshold, Easy-To-Deploy and Ready-To-Use solution with a relatively low computational cost for medical image classification at an early stage of an outbreak in which scarce data are available and traditional data augmentation may fail. Hence, it would be most suitable for low-resource settings.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and its variants cause the coronavirus disease 2019 (COVID-19) pandemic. There have been 662 million confirmed cases of COVID-19 and approximately 6.7 million deaths as of January 2023 [1]. Fever, cough, myalgia and fatigue are common symptoms of patients infected by COVID-19 [2].

Patients with COVID-19 are usually confirmed by reverse transcription polymerase chain reaction (RT-PCR) testing. However, RT-PCR tests for specific virus need to be carefully designed, hence early prototype cannot effectively detect COVID-19 at the early stage of the outbreak due to its low sensitivity [3, 4]. Besides, suspected patients often cannot be tested in time

because of the shortage/unavailability (especially in undeveloped countries) of RT-PCR test kits during the same period. Therefore, radiological imaging methods like x-rays and chest computer tomography (CT) become complementary examinations to help clinicians diagnose COVID-19 correctly although it cannot detect patients without any lung damage at the earliest stages of infection [5–7]. In addition, imaging methods especially CT can provide semi-quantitative analysis of pulmonary damage severity [8] and monitor the long-term lung damage of patients who have recovered from COVID-19 [9]. CT scans provide more detailed tissue and organ information than x-rays, and CT is a useful tool to efficiently distinguish 'probably positive' and 'probably negative' patients

[10]. Also, x-rays cannot detect any abnormalities of early infection of COVID-19 [11].

Since CT image analysis is time-consuming, researchers proposed an artificial intelligence (AI) model and proved it has potential to identify COVID-19 patients rapidly [12]. B Wang *et al* built an AI system to carry out the task of COVID-19 CT image classification, which can save about 30%-40% detection time [13]. S Wang *et al* modified the inception transfer-learning model and obtained an accuracy of 79.3% in a dataset that included 740 COVID-19 and 325 non-COVID-19 CT images [14]. Wu *et al* proposed a multi-view deep learning fusion model based on ResNet50, and achieved an accuracy of 76% [15]. Chen *et al* applied UNet++ on a CT dataset that contained 35355 images, and achieved an accuracy of 98.85% [16]. Ardakani *et al* tested ten different convolutional neural network (CNN) models and got the best performance with an accuracy of 99.51% and Area Under the Receiver Operating Characteristic (AUC) of 99.4% [17]. In addition, AI methods can not only discriminate COVID and Non-COVID images but also simultaneously classify other type of lung diseases such as lung cancer, viral pneumonia, bacteria pneumonia and so on [18].

Unfortunately, most COVID-19 CT datasets cannot be shared with the public because they involve patients' privacy, which is a common problem in medical image analysis. Einstein *et al* summarised COVID-19 medical images datasets and such datasets with sufficient high-quality data were not open source [19]. Meanwhile, research results based on these datasets are difficult to reproduce. Although several datasets are publicly available, they do not have sufficient data for the training of deep learning models. To solve these two problems, He *et al* proposed a self-supervised transfer learning approach and obtained an accuracy of 86% on a customised public COVID-19 CT dataset they built [20].

Transfer learning and data augmentation are helpful for image classification when only limited data are available [21, 22]. Zhao *et al* pre-trained the ResNet-v2 model on ImageNet-21k, then applied transfer learning and achieved an accuracy of 99.2% while detecting the COVID-19 cases [23]. Loey *et al* explored a combination of traditional data augmentation methods and Conditional Generative Adversarial Nets (CGAN); the performances of COVID-19 CT classification on five deep learning models (AlexNet, VGGNet16, VGGNet19, GoogleNet, and ResNet50) were improved [24]. However, these two approaches are not always beneficial. Transfer learning may only slightly improve image classification performance because of the differences in data and tasks between source and target domain [25, 26]. Furthermore, pre-trained weights are usually obtained from general-purpose datasets like ImageNet without COVID-19 CT scans. Data augmentation strategy significantly affects discriminative performance, but little work

mentioned how to build a suitable strategy for medical image classification [27].

This work mainly aims to provide a potential Low-Threshold, Easy-To-Deploy and Ready-To-Use tool that can quick response to similar outbreaks in the future. At the early stage of such epidemics, rapid diagnosis with timely isolation is an effective method of preventing the spread of outbreaks. Therefore, we focus on using existing methodologies and interlocking them effectively to build a rapid reaction tool rather than developing a completely novel model. The desired outcome can improve COVID-19 CT classification performance based on a deep transfer learning model in a realistic scenario that reflects the early stage of the outbreak of the COVID-19 and any epidemic: (*i*) Scarcity of labelled COVID-19 CT images for training; (*ii*) Data may come from multiple sources; (*iii*) Only limited computing resources may be accessed.

We improve the accuracy from 91.15% to 95.50% in a typical early open-source COVID-19 CT dataset by using synthetic CT images synthesised by Advanced Normalisation Tools (ANTs) as augmented data in EfficientNet-B2. A customised dataset is built to verify the benefit of synthetic images. Notably, most layers are frozen in the process of transfer learning, and we adjust hyperparameters empirically, so that the classification task is done with relatively low computational cost.

Results imply that ANTs could be a potential alternative to Generative Adversarial Networks (GANs) to synthesise images in medical image classification tasks. We hope that this study could provide a new possibility for rapid computer-aided diagnosis in the field of medical imaging in the early stage of future epidemics.

The rest of the paper is organised as follows. Section 2 introduces the methodology, including datasets, synthetic images, and the design of experiments. Results and discussion are described in sections 3 and 4, respectively. Finally, section 5 presents the conclusion.

2. Materials and methodology

In this section, we introduce the datasets, synthesis of images and configuration of the deep learning model. Figure 1 illustrates a flowchart reflecting all datasets we used and the experimental design.

2.1. Datasets introduction
2.1.1. COVID-CT dataset
Yang *et al* built the COVID-CT dataset, an open-sourced dataset that includes 349 COVID-19 and 463 non-COVID-19 CT images, and obtained a classification accuracy of 89% in a model based on multi-task learning and self-supervised learning [28]. The dataset was broadly adopted (see table 4) and the utility of it has been confirmed by senior radiologist, although the quality of paper-extracted CTs is worse than the

Figure 1. Experimental design flowchart.

Figure 2. CT scans problems in the COVID-CT dataset. (a) Non-normalisation contrast, (b) Embedded text, (c) White border. a), b) and c) have different resolutions.
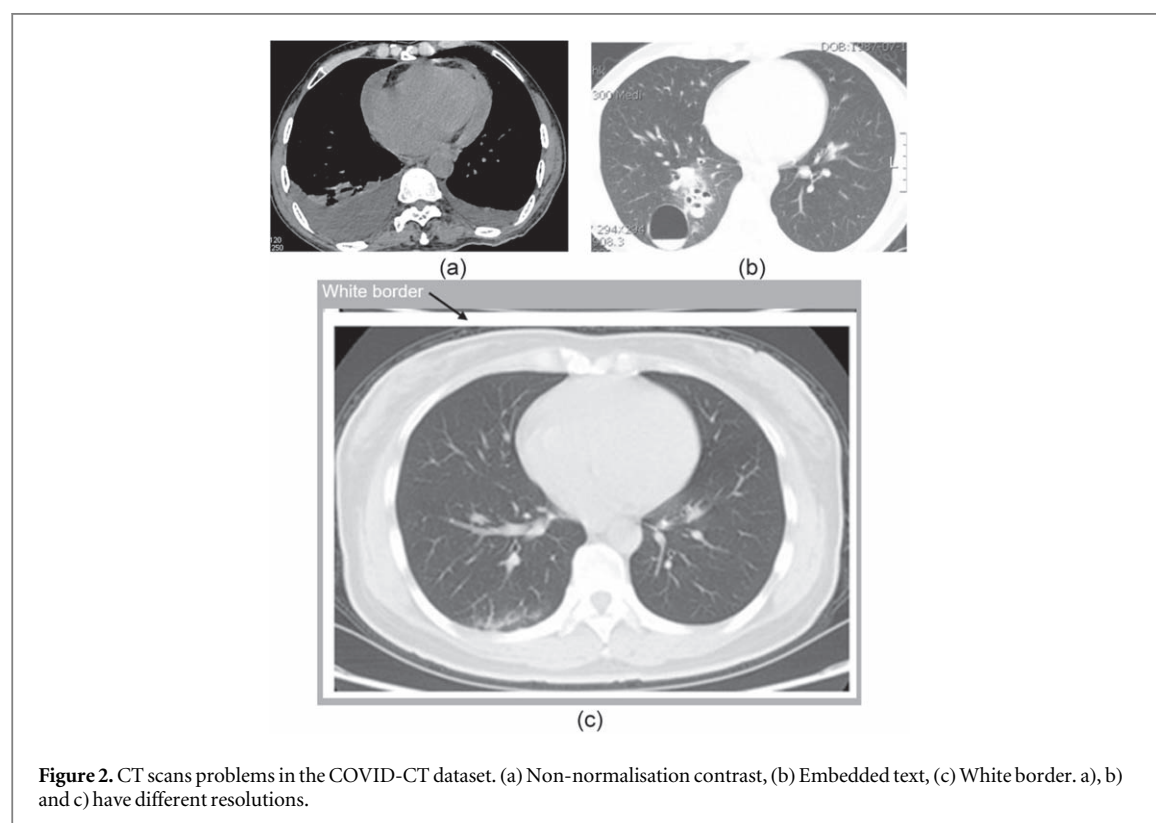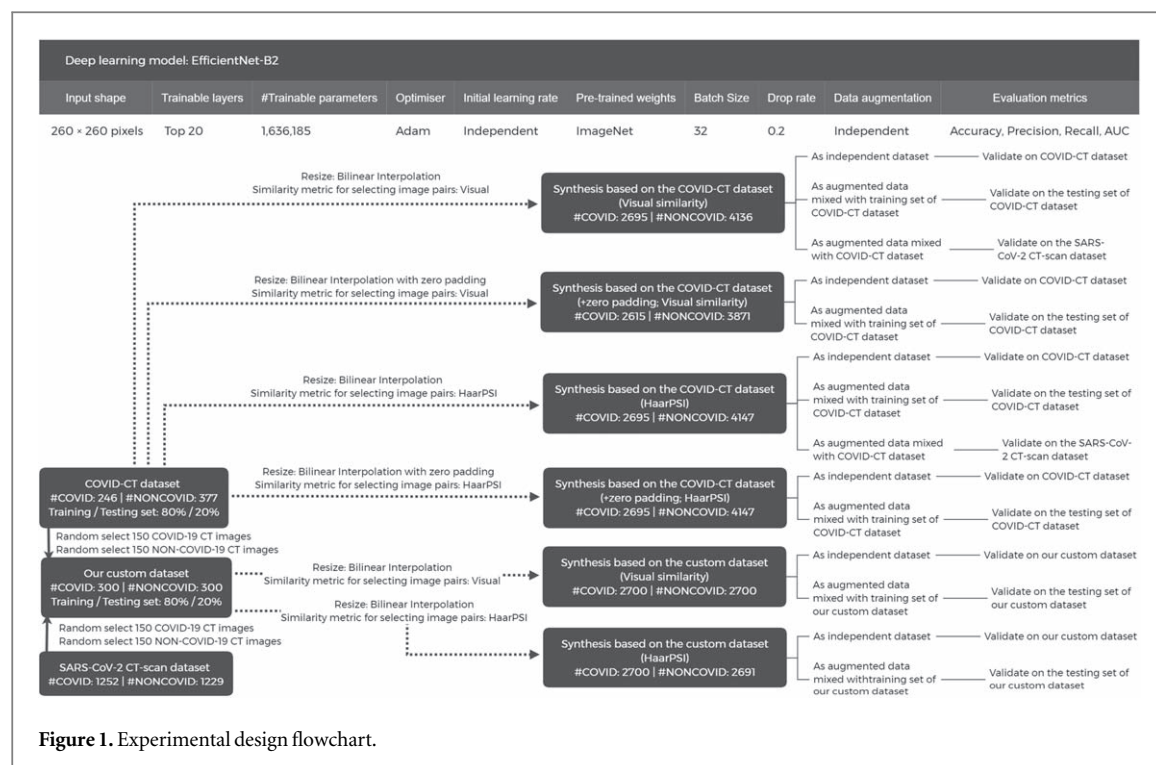
original CTs [28]. CT images were extracted from numerous papers from multiple sources such as medRxiv2, bioRxiv3, MedPix, LUNA, Radiopaedia and PubMed Central. In this case, some data belonging to one single source of this dataset were not continuous (e.g., most of the images were missing from a series of CTs of a patient). Besides, data from different sources were generated by different CT scanners worldwide. Compared with the use of data from a single source, multisource data increase the difficulty of the classification, especially when data are insufficient. However, the dataset represents typical easy-to-obtain and publicly available data at the early stage of

the COVID-19 epidemic, which lowers the threshold for researchers to explore related topics.

Figure 2 illustrates four problems found in Yang's COVID-CT dataset [28]: (*i*) non-normalisation contrast; (*ii*) embedded text; (*iii*) white border; (*iv*) resolution inconsistency. Since synthetic images are generated based on images from this dataset and such problems adversely affect the quality of the synthetic images, only 246 COVID-19 and 377 non-COVID-19 CTs are retained in the dataset after selection.

Specifically, contrast intensity was re-mapped in the range of [0, 1]. The embedded text was an irrelevant feature for this classification task and interfered with model performance. Therefore, images with two or more lines of embedded text were discarded, but the rest were kept as noisy data to prevent possible overfitting in the following classification task. We removed the white border by cropping to avoid generating a large number of synthetic images with irregular white borders. The solution for various resolutions is described in the 'resolution normalisation' subsection.

2.1.2. Custom dataset

The custom dataset was derived from the COVID-CT dataset [28] and the SARS-CoV-2 CT-scan dataset [29]. To this end, we randomly selected 300 CT images (150 COVID-19 and 150 non-COVID-19 images) from each of these two datasets, and then built the custom dataset (600 images in total).

The COVID-CT dataset [28] was introduced in the previous subsection. The SARS-CoV-2 CT scan dataset contained 1252 COVID CTs and 1230 non-COVID CT scans collected from hospitals in São Paulo, Brazil [29]. Angelov *et al* built it and achieved an accuracy of 97.38% in an eXplainable Deep Learning approach (xDNN) [29]. As the previous subsection mentioned, problems (i) and (iv) were observed in this dataset. We only obtained 1252 COVID-19 and 1229 non-COVID-19 CT images when we accessed the dataset [30]. We mainly used partial data from this dataset to build a custom dataset, and the entire dataset was treated as a test set in our cross-dataset (i.e., training and test set are from different datasets instead of splitting one dataset into training, validation and test sets) experiment.

2.2. Synthetic CT images

2.2.1. Selection of synthesis methods

Generative Adversarial Networks (GANs) [31] are commonly used to expand datasets by synthesising diverse and realistic images, particularly in the biomedical domain [32–34]. Methods based on GANs have been applied to generate high-quality COVID-19 CT images [35, 36]. However, GANs usually require enormous data with high computational costs, especially when high-quality and high-resolution synthetic images are needed [37, 38]. The time required to

customise and fine tune the model is ill-advised for rapid response in a fast-spreading pandemic. Besides, Yi *et al* pointed out that most works on synthesising medical images through GANs adopt metrics like Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM) that could not correspond to the visual quality of images [39].

To reduce the dependence on the high-performance hardware, we utilised the Advanced Normalisation Tools (ANTs), initially designed for deformable image registration with small or large deformations, to synthesise CT images [40]. Moreover, default configurations of functions provided by ANTs are good enough so further careful fine-tuning is not necessary.

ANTs provided a technique called 'morphing' based on Geodesic Image Interpolation (GII). Avants *et al* used GII to simulate the missing volumetric brain images from two in a series of images, and proved that it offers 25%–30% better intensity accuracy than linear interpolation [41]. It is a feasible and potentially efficient method to synthesise images, especially when dealing with images from multiple data sources or defective image sets with partial missing data.

Suppose there are two 'controlled' images, one is a 'fixed' image and the other is a 'moving' image. Applying 'morphing' will force the 'moving' image to be partially deformed to the 'fixed' image. The 'morphing' function allows us to synthesise one or more images at a specific position between two images. Figure 3 illustrates an example of the synthesis. Figure 4 shows synthetic images obtained when the same set of 'controlled' images is applied with different parameters in ANTs.

Given that some of the images in the COVID-CT dataset only presented a single slice per patient, and both patients as well as CT scanners information was erased, we could not use ANTs as we discussed above. Besides, to the best of our knowledge, currently no guidance for synthesising images through ANTs under such conditions exists. This void is addressed here as described next.

Previous works demonstrated that the larger the feature gap between the 'controlled' images, the higher the probability of generating a greatly distorted image. Figure 5 shows examples of heavily distorted synthetic images. Intuitively, visual similarity based on the subjective visual perception of researchers can be used to select image pairs. We also introduce Haar wavelet-based perceptual similarity index (HaarPSI) [42] as a measurable metric to do the same job for a comparison. HaarPSI is a computationally inexpensive image similarity and quality evaluation metric widely used in the medical image domain [43–45].

2.2.2. Resolution normalisation

Images should be resized to a uniform resolution before inputting into a convolutional neural network.

Figure 3. The synthesis process through ANTs. (a) Original "fixed" image, (b) Original "moving" image, (c) Scaled "fixed" image, (d) Scaled "moving" image, (e) Synthesised image through ANTs, (f) an example of synthetic image without zero padding.

Figure 4. Synthetic images with different parameters in ANTs (the input image pair is the same as figure 3).
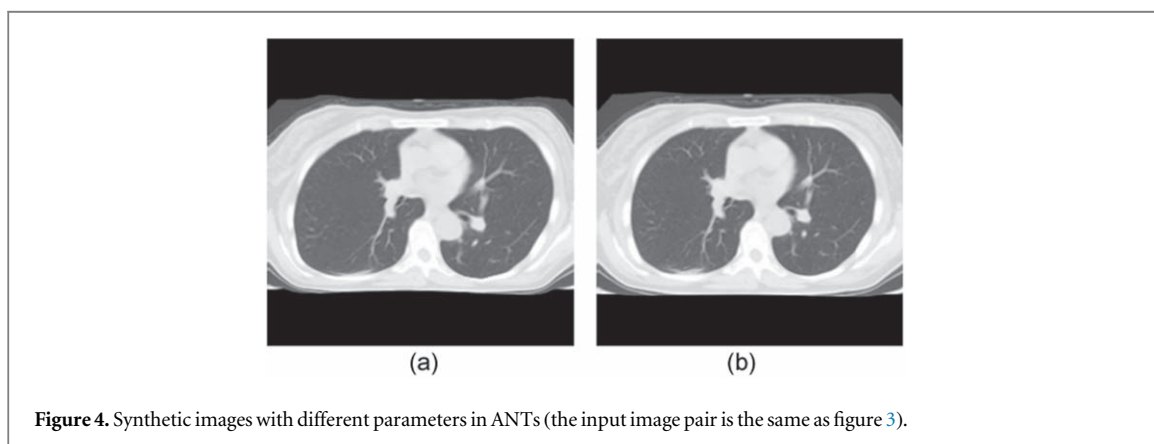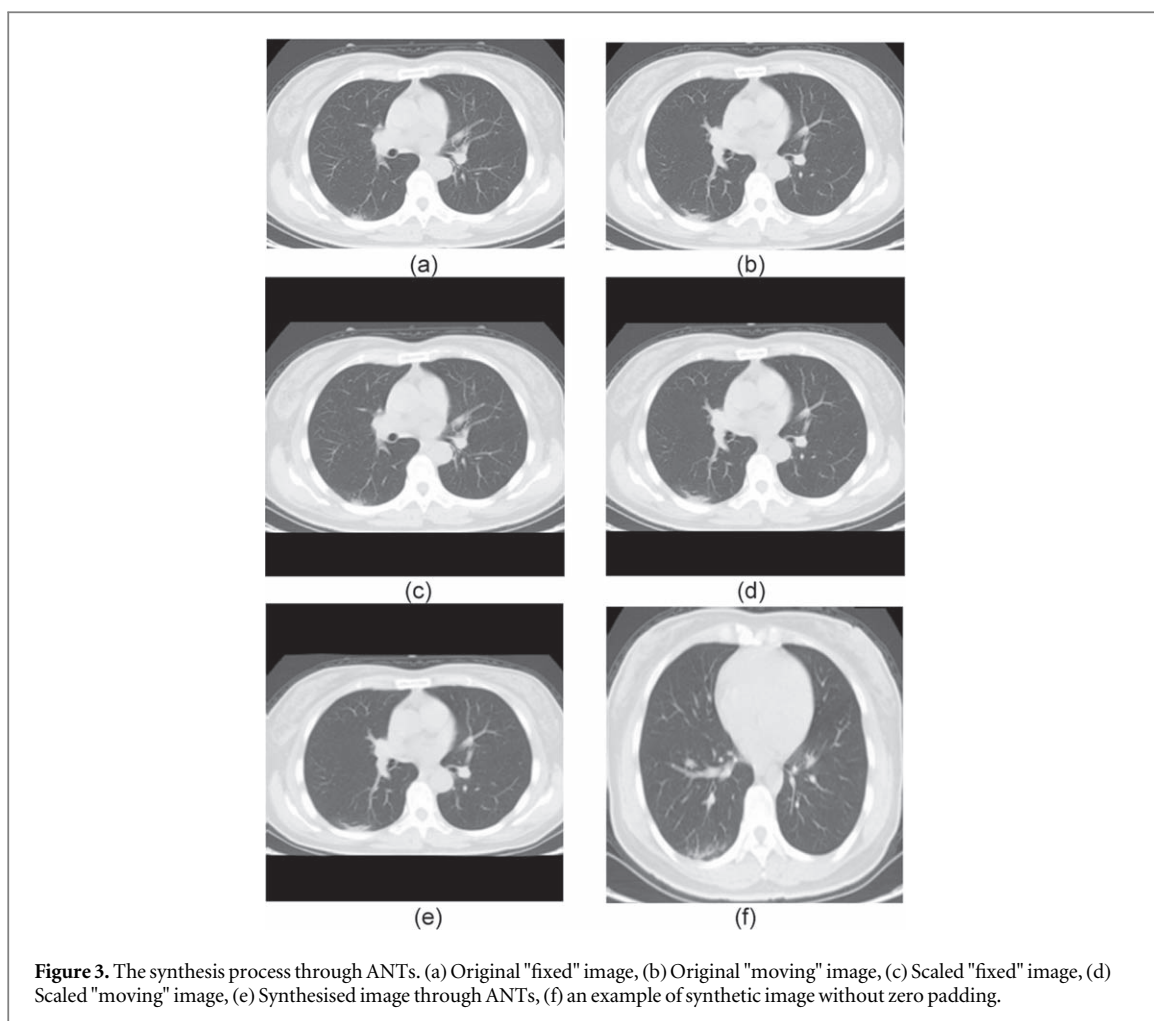
In addition, the probability of synthesising images with large degree distortion can be reduced by using the same size images.

Usually, with image resolution between 256×256 pixels and 448×448 pixels, AUC achieved the maximum value in binary classification tasks of the chest radiograph undertaken by the convolutional neural network (CNN) [46]. To reduce the computational cost, we resized all CT images to 260×260 pixels which is the input shape of EfficientNet-B2 architecture (table 1). Although it seems that scaling the image to 256×256 pixels can minimise the use of

computational resources, we must re-code the deep learning model that can be directly transferred, which does not fit our original intention: a Ready-To-Use method.

To balance the computational cost and quality of synthetic images, we used bilinear interpolation (i.e., linear interpolation in two dimensions sequentially) to scale images instead of nearest-neighbour or bicubic interpolation [47]. Two main scaling methods were considered: conventional bilinear interpolation with or without zero padding.
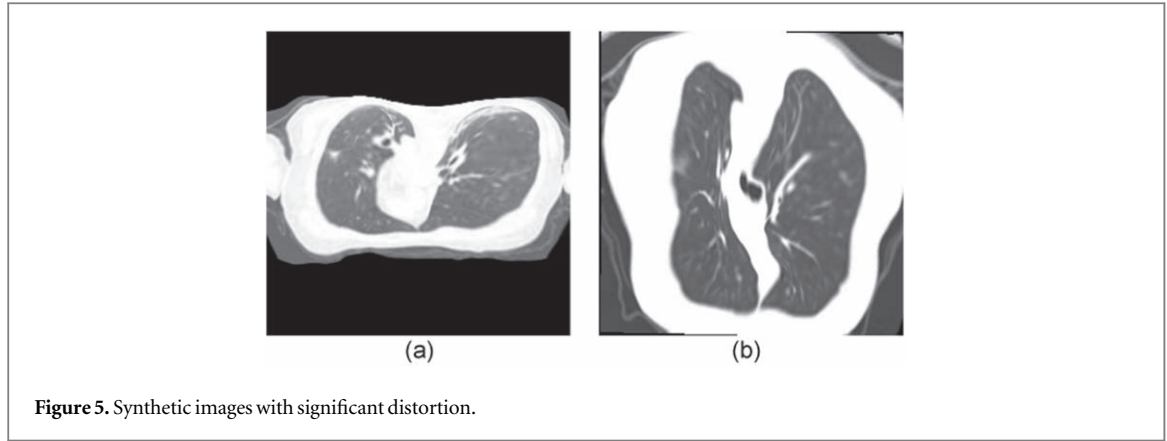
Figure 5. Synthetic images with significant distortion.

Table 1. Input shapes of efficientnet base models.

EfficientNet model	Input shape
EfficientNet-B0	224 pixels × 224 pixels
EfficientNet-B1	240 pixels × 240 pixels
EfficientNet-B2	260 pixels × 260 pixels
EfficientNet-B3	300 pixels × 300 pixels
EfficientNet-B4	380 pixels × 380 pixels
EfficientNet-B5	456 pixels × 456 pixels
EfficientNet-B6	528 pixels × 528 pixels
EfficientNet-B7	600 pixels × 600 pixels

Zero padding (i.e., adding zero-value pixels to the borders of images) is proposed to enlarge small images to a fixed size without loss and improve image classification tasks' accuracy as well as time performance in CNNs. However, Hashemi pointed out that it did not affect the accuracy but significantly reduced the convergence time because zero input values did not activate convolutional units [48]. Hence, we attempted to combine the bilinear interpolation and zero padding. Furthermore, the aspect ratio was kept, and loss only came from interpolation.

We scaled each image based on the scale factor of width. For example, an image is $W_O \times H_O$ pixels and the target size is $W_T \times H_T$ pixels. The scale factor F_s is as follows:

$$F_S = \frac{W_T}{W_O} \tag{1}$$

Then the new height H_N is calculated as shown in the equation below:

$$H_N = H_O \times F_S = H_O \times \frac{W_T}{W_O} \tag{2}$$

If the new height H_N is smaller than the target height H_T, zero-value pixels are used to fill the blank between them, as shown in figure 3(c) and figure 3(d). On the contrary, if the new height H_N is bigger than the target height H_T, the image will not participate in the synthesis. Cropping is not accepted because it causes feature loss. Fortunately, the pre-processed COVID-CT dataset does not contain such images. In

brief, the combined method only filled the top and bottom borders of the image instead of around the image compared with the original zero padding.

Another way to maintain the aspect ratio is to add cropping to the interpolation process. However, we believe that this impairs classification performance because cropping results in a loss of information.

2.3. Implementation details
2.3.1. Deep learning architecture
Deep learning has various applications in radiology, especially classification, segmentation and detection [49]. Many deep learning models can undertake the classification of COVID-19 CT scans, such as AlexNet, ResNet-50, Inception-v3, and Xception [50–52]. However, they have a large number of trainable parameters. For example, AlexNet has about 61 million parameters, which needs enormous computing resources and training time.

Tan and Le [53] developed the EfficientNet family that outperforms all previous models we mentioned in accuracy and efficiency when applied to the ImageNet dataset. Compared with traditional methods that scale one dimension (width, depth or resolution) of the network, the EfficientNet scales all these dimensions uniformly by a compound coefficient. Therefore, EfficientNet allows people to arbitrarily choose width/depth/resolution according to the compound scaling formula:

$$\begin{aligned} &\textit{Depth}: d = \alpha^\phi \\ &\textit{Width}: w = \beta^\phi \\ &\textit{Resolution}: r = \gamma^\phi \\ &\textit{Constrain 1}: \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ &\textit{Constrain 2}: \alpha \geqslant 1, \ \beta \geqslant 1, \ \gamma \geqslant 1 \end{aligned} \tag{3}$$

Where $\varnothing$ is a user-specified coefficient that reflects computing resources, and $\alpha = 1.2$, $\beta = 1.1$ as well as $\gamma = 1.15$ are calculated by a grid search based on the EfficientNet-B0. However, the actual implementation is restricted by many factors (e.g., the channel size should be a multiple of 8 required by the building block). Hence, Keras only provides 8 classic EfficientNet models (B0-B7) with specific width/depth/resolution. Table 1 shows the input shape of these models.

Finally, we chose the EfficientNet-B2 (Input shape is 260×260 pixels) due to the computational cost and the effect of CT image size concern (see the subsection 'resolution normalisation').

2.3.2. Training configurations

The EfficientNet-B2 model based on transfer learning with pre-trained weights from ImageNet was deployed in the experiment. Since we used a much smaller dataset than ImageNet, we applied extremely small learning rates to obtain incremental changes in performance. Besides, a large learning rate may cause the model to fail to converge in our experiments.

To further reduce the computational cost, we strictly limited some of the hyperparameters of the model. Only the top 20 layers could be trained, except for the built-in BatchNormalisation layers because they had non-trainable weights. Therefore, only 1,636,185 out of 7,775,610 were trainable parameters in Keras. Meanwhile, we empirically adjusted hyperparameters instead of grid or random search that cost enormous resources.

We used Adam optimiser to update weights and separately set the learning rate of the top layer and other unfrozen layers. The dropout rate [54] of the top layer was set to 0.2 to prevent overfitting. Datasets in baseline tests were split into a proportion of 80% and 20% for training and testing, respectively. Batch size and maximum epochs were set to 32 and 100 separately.

2.3.3. Data augmentation

Data augmentation expands training datasets and enhances the data quality to solve the problems when meagre data can be accessed, especially medical data [55, 56]. It has been shown to improve the performance of deep learning models and help to correct overfitting [57]. Generally, it can be divided into two methods in image classification tasks: transformations of images and introducing new synthetic data. Although this work focuses on the effect of synthetic images, we still introduce the traditional data augmentation to compare performances.

Data augmentation methods are not omnipotent, and their specific drawbacks make them be unequally popular [58]. A commonly used combination was applied to our experiments: (*i*) rotation by a random amount in the range $[-10\% \times 2\pi, 10\% \times 2\pi]$; (*ii*) random translation vertically or horizontally in the range $[-10\%, 10\%]$; (*iii*) flip each image vertically or horizontally; (*iv*) randomly adjust the contrast of images.

However, combining augmentation brings a complex impact and no guaranteed benefits. A study reported that data augmentation harmed deep learning models in detecting COVID-19 x-ray images [59]. Therefore, we did not expect the typical augmentation combination to be advantageous, particularly when the capabilities of the deep learning model were limited.

2.3.4. Evaluation criteria

Four metrics were applied to evaluate the classification performance: Accuracy, Precision, Recall and Area Under the Receiver Operating Characteristic (AUC) score. For these metrics, the higher, the better.

2.3.5. Design of experiments

Experiments were carried on a laptop with Intel(R) Core(TM) i7-10875H CPU @ 2.30 GHz, 32GB RAM, NVIDIA GeForce RTX 2060 6 G and Windows 10. Image normalisation was completed by Matlab, and image synthesis was done by ANTs on Linux. Keras/Tensorflow undertook the classification task in Python.

The first experiment explored the effect of different image resizing methods with or without traditional data augmentation. Then synthetic images were introduced and compared with the best model of the first experiment. Specifically, all synthetic images were firstly treated as an independent dataset, then trained on it and tested on the source dataset (i.e., the dataset provides 'controlled' image pairs). Next, synthetic images were treated as augmented data to mix with the training set and validated on the testing set.

The aim of using synthetic images as an independent dataset is to evaluate the utility of synthetic data. For example, if we achieve 100% accuracy when we train on synthetic data and test on original data, it may indicate such synthetic data perfectly replicates the characteristics of real data but may exacerbate overfitting when testing on unseen data (i.e., test set). Instead, we can say that the synthetic data includes many wrong features if the experiment shows low accuracy.

Silva *et al* [60] proposed a cross-dataset test to evaluate the generalisation power of deep learning models and reported the best accuracy of 56.16% when training on the SARS-CoV-2 CT-scan dataset and testing on the COVID-CT dataset. The opposite scenario produced worse results because the training set was much smaller than the testing set [60]. Hence, we merged the COVID-CT dataset and synthetic images as a training set and tested it on the SARS-CoV-2 CT-scan dataset to explore whether synthetic images enhance generalisation.

Finally, the custom dataset was built (see subsection 'custom dataset') and tested to verify latent conclusions derived from previous experiments.

3. Results

The baseline model applied bilinear interpolation as resizing method without data augmentation and achieved the best accuracy of 91.15% on the COVID-CT dataset. The use of cropping or interpolation with zero padding harmed the performance of the model (accuracy of 88.32% and 86.06% separately), and typical data augmentation methods had more severe

Table 2. Comparison of different image pre-resizing methods with or without traditional data augmentation for the COVID-CT dataset.

Resizing method	Traditional data augmentation	Learning rate of top layer	Learning rate of other layers	Accuracy	AUC	Precision	Recall
Bilinear interpolation	No	1×10^{-2}	1×10^{-4}	91.15%	96.40%	89.22%	89.89%
	Yes	5×10^{-4}	2×10^{-5}	74.81%	85.45%	89.38%	45.51%
Bilinear interpolation with cropping	No	1×10^{-2}	1×10^{-4}	88.32%	96.26%	88.60%	83.07%
	Yes	5×10^{-4}	2×10^{-5}	77.11%	86.10%	76.07%	66.30%
Bilinear interpolation with zero padding	No	3×10^{-4}	3×10^{-5}	86.06%	94.55%	85.64%	80.68%
	Yes	5×10^{-4}	2×10^{-5}	84.60%	91.18%	82.06%	81.23%

adverse effects in this case. Table 2 reports the performance difference among the three resizing ways with or without traditional data augmentation.

The number of synthetic images is presented in figure 1. Synthetic images were treated as independent datasets and augmented data separately. Performances were less well than the source dataset baseline model when synthetic images were used as independent datasets. However, performance was improved when images were synthesised by visual similarity and became augmented data (table 3). Figure 6 illustrates the accuracy curve and loss curve, which achieved an average accuracy of 95.50% when synthetic images were considered as augmented data. Table 4 shows a comparison of our best results with other studies using the COVID-CT dataset [28]. From the perspective of AUC metric, the improvement may not significant compare with others results (from 94.2% [61] to 98.54%), but the model we used is much smaller (EfficientNet-B2 has 9.2 million parameters in total but ResNet-50 has 26 million).

Unfortunately, synthetic images seemed no benefit on generalisation capability because the model did not converge in the cross-dataset test. Training on the COVID-CT and testing in the SARS-CoV-2 CT-scan dataset presented a poor accuracy of 49.31%. Adding images based on visual similarity or HaarPSI and both pre-resized by interpolation obtained the accuracy of 48.48% and 50.12%, respectively.

The baseline performance of the custom dataset without synthetic images obtained 85.95%, 93.21%, 87.27% and 84.60% accuracy, AUC, precision and recall, respectively. When traditional data augmentation was applied, they dropped to 78.29%, 84.80%, 81.07% and 75.16%, respectively. Table 5 shows the performance when synthetic images were considered and gives a similar performance trend to previous experiments. The best scenario increased the accuracy and AUC to 94.32% and 98.61% separately.

4. Discussion

The main goal of our experiments is to find a convenient and efficient solution for classification tasks based on deep learning when limited data and

computing resources are available. In such cases, traditional data augmentation methods based on basic image operations may fail. In the experiments, we selected a typical dataset, the COVID-CT dataset, created on the early stage of the epidemic and can be accessed by the public. To simulate a low computational power environment, we froze most of the trainable layers of EfficientNet-B2 and synthesised images through ANTs instead of GANs. Meanwhile, grid search, random search or other expensive hyperparameter tuning methods were forbidden.

In this work, we first proposed an image scaling method based on interpolation and zero padding and compared it with two other ways: bilinear interpolation or interpolation with cropping. As expected, although cropping maintains the aspect ratio of images, the loss of features impairs the model's performance. Unfortunately, the proposed resizing method also adversely affects the deep learning model in this case (table 2). It seems to be attributed to the same reason that zero values cannot activate the convolutional unit as [48] reported. Furthermore, the proposed method scaled all images to a given resolution, but the images were not filled with the same number of black pixels. Intuitively, the area of the black pixels generated by zero padding was not the same between scaled images, which directly led black pixels to blend into the surroundings and produce irregular black borders during synthesis, as shown in figure 5 (a).

Then we used interpolation and interpolation with padding to further synthesise images through ANTs. Visual similarity and HaarPSI were applied to select image pairs. When resizing methods were analysed independently, there was little difference in the impact of pre-resizing the image by interpolation or interpolation with zero padding. Synthesis based on visual similarity showed better performance improvements than HaarPSI when the effects of resizing approaches were ignored (table 3). In the best case, the accuracy and AUC improved from 91.15% to 95.50% and 96.40% to 98.54% separately after the synthetic images based on visual similarity and pre-resized by bilinear interpolation were added to the training set as augmented data.

Table 3. Comparison of the Impact of Synthetic Images on Classification Performance on the COVID-CT Dataset.

| Resizing method	Selection criteria for image pairs	Usage of synthetic images	Learning rate of top layer	Learning rate of other layers	Accuracy	AUC	Precision	Recall
Bilinear interpolation	Visual similarity	Independent	7×10^{-9}	3×10^{-5}	89.67%	96.60%	92.21%	80.71%
		Augmented	1×10^{-9}	1×10^{-5}	95.50%	98.54%	94.85%	93.81%
	HaarPSI	Independent	2×10^{-4}	2×10^{-6}	75.49%	82.87%	66.24%	77.39%
		Augmented	2×10^{-9}	3×10^{-5}	89.91%	96.78%	81.89%	96.00%
Bilinear interpolation with zero padding	Visual similarity	Independent	7×10^{-9}	3×10^{-5}	86.24%	97.25%	96.39%	67.81%
		Augmented	1×10^{-9}	1×10^{-5}	94.07%	98.65%	90.79%	94.86%
	HaarPSI	Independent	2×10^{-4}	2×10^{-6}	77.87%	84.03%	71.86%	72.28%
		Augmented	2×10^{-8}	1×10^{-5}	89.04%	94.70%	84.52%	88.67%

Y Hou and M Navarro-Cía
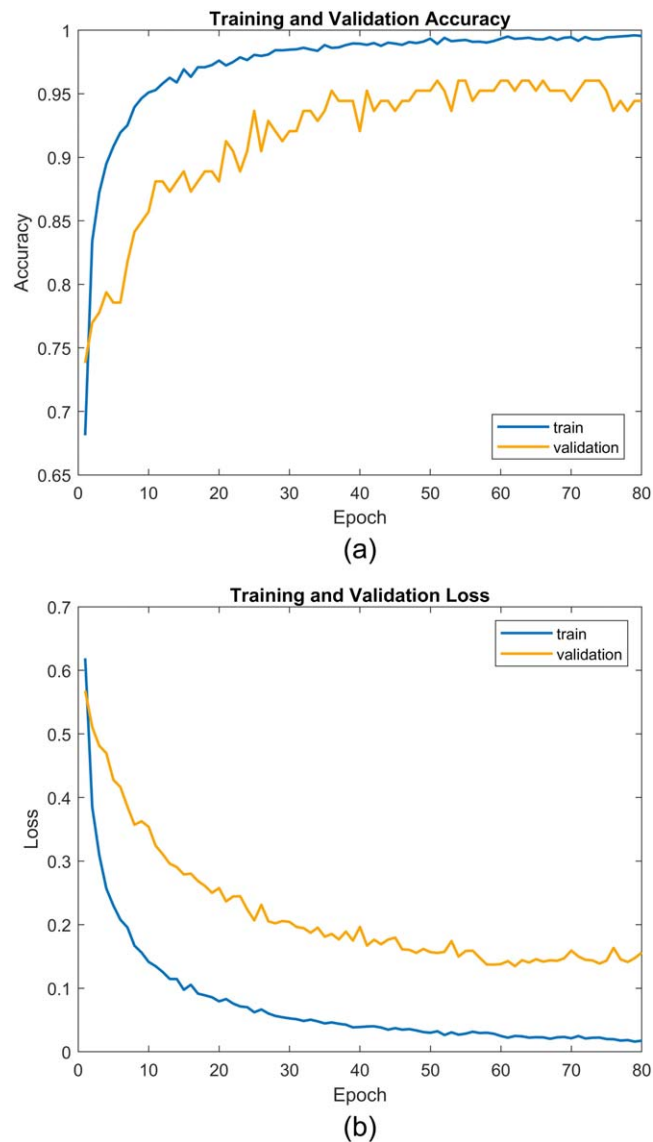
Figure 6. The training and validation accuracy (a) and loss curve (b) when synthetic images as augmented data on the COVID-CT dataset.

Table 4. Performance comparison between our proposed method and some existing works on the COVID-CT dataset.
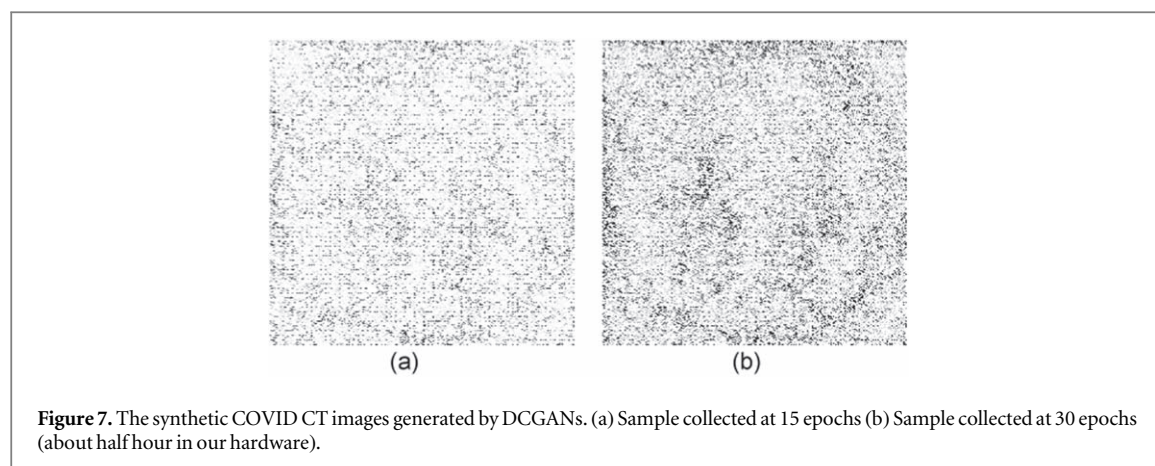
Author(s)	Model	Note	Accuracy	AUC	Precision	Recall
M Loey *et al* [24]	ResNet50	Traditional augmentation with CGAN	81.41%	—	—	80.85%
P Silva *et al* [60]	EfficientCovidNet	Traditional augmentation	87.68%	—	93.98%	79.59%
A Shamsi *et al* [61]	ResNet50	Linear SVM	87.9%	94.2%	—	86.5%
X Yang *et al* [28]	DenseNet-169	Contrastive self-supervised learning; Traditional augmentation	89.10%	98.10%	—	—
A Saeedi *et al* [62]	DenseNet-121	Nu-SVM classifier	90.61%	95.05%	89.76%	90.80%
Ours	EfficientNet-B2	ANTs	95.50%	98.54%	94.85%	93.81%
S Chattopadhyay *et al* [63]	Deep features of ResNet18	Clustering-based Golden Ratio Optimizer (CGRO)	99.31%	—	99%	100%

Notably, synthesising images through ANTS is more efficient than GANs or Deep Convolutional Generative Adversarial Networks (DCGANs). We have tried to build a traditional GAN to synthesise images, but the training process hardly converges after tuning. Then we built a DCGAN and applied many approaches (e.g., batch normalization, gaussian weight initialization, images re-scale to the range [−1, 1], etc) to stabilise it. The best DCGAN model in our experiment had a three convolution layers

Table 5. Comparison of the impact of synthetic images on the classification performance of the custom datasets.

Resizing method	Selection criteria for image pairs	Usage of synthetic images	Learning rate of top layer	Learning rate of other layers	Accuracy	AUC	Precision	Recall
Bilinear interpolation	Visual similarity	Independent	5×10^{-6}	3×10^{-5}	82.32%	95.87%	95.77%	67.73%
		Augmented	2×10^{-3}	2×10^{-5}	94.32%	98.61%	93.66%	95.24%
	HaarPSI	Independent	5×10^{-6}	1×10^{-5}	70.02%	83.29%	84.20%	49.43%
		Augmented	5×10^{-3}	5×10^{-5}	90.08%	96.57%	89.46%	90.95%

Figure 7. The synthetic COVID CT images generated by DCGANs. (a) Sample collected at 15 epochs (b) Sample collected at 30 epochs (about half hour in our hardware).

discriminator with 650,241 parameters and a three transposed convolution layers generator with 18,771,073 parameters. The model was trained from scratch and generated only an extremely obscure outline of CT scans (see figure 7) within half hour. However, ANTs could generate about 2600 usable synthetic images at the same time in a virtual machine with half computational resources compared with the DCGAN model. Unlike GANs and their descendants, fine-tuning is not necessary for ANTs to generate usable synthetic images, which significantly reduces the time and resources required. In addition, the training on GANs usually not stable and the trade-off between performance and computational cost should be reflected.

When synthetic images were used as an independent dataset and validated on the source dataset, the performances were lower than the baseline model but still acceptable. It indicated that the generated images were diverse. A small number of images with significant distortion were synthesised, and we did not remove these data. We believe keeping these data can prevent overfitting when they are considered as augmented data. Additionally, cleaning this data may require the supervision of a radiologist.

Since the above results (table 2 and table 3) and discussions have proved that the proposed resizing method (bilinear interpolation with zero padding) did not show any benefits, we decided only to adopt bilinear interpolation, the best in the previous experiments, as the image resizing method for the following experiments related to the custom dataset and cross-dataset test.

To simulate the dilemma faced by researchers in the early stage of any outbreak (i.e., the lack of data and the wide range of data sources), we customised a dataset based on two open-source datasets: COVID-CT [28] and SARS-CoV-2 CT-scan [29]. When synthetic data were added, the accuracy significantly improved from 85.59% to 94.32%, which was a promising result and proved that synthetic images by ANTs could enhance the performance of the deep learning model. A research combined four datasets that included almost 2200 images, which is larger than our custom dataset, and obtained an accuracy of 90.91% based on machine learning [64].

The cross-dataset test showed current synthetic images used in this experiment did not contribute to the generalisation capability of the deep learning model. We consider two major reasons here. Firstly, our synthetic data are generated from low-quality data and noise also be 'amplified' during data augmentation. Secondly, our cross-dataset test is based on two datasets instead of splitting one dataset into train, validation and test sets. Well-constructed dataset usually has a specific data distribution, and datasets with similar content theoretically also belong to the same distribution, but the small differences will be significantly magnified by the gap in the amount of data. Hence, we look forward to verifying our method on high-quality and bigger dataset in the future. Furthermore, we believe the latter one can be mitigated by federated learning with multiple datasets because the aggregation averages the model parameters that trained on different datasets.

We also found several limitations. Firstly, the dataset we used is small and lower quality compare with the private datasets that we did not have permission to access. Although we have confidence on our method with high quality data, it should be further verified in the future. Secondly, we did not clean the synthetic data, which means some synthetic images with significant distortions were kept and adversely affected the model. However, data cleaning in the field of medical imaging usually requires the assistance of radiology experts. Thirdly, we only tested two similarity metrics: visual similarity and HaarPSI. In this case, the visual similarity is better, but we do not know how other metrics will behave in such scenarios. Since there is no current guidance for synthesising medical images based on similarity measurement through ANTs, we provide a simple approach that could be scrutinised further. In future works, we will evaluate more similarity metrics and pay close attention to advanced metrics that can better reflect visual similarity. Finally, only one commonly used combination of data augmentation was considered in our work. Although it

performed poorly in this experiment, it may achieve a better result with careful fine-tuning. Future work should explore the efficient application of data augmentation to small datasets with diverse data.

The experiment was based on transfer learning to overcome the problem of data lack. However, some researchers pointed out that transfer learning that adopted pre-trained weights from general datasets like ImageNet offered limited performance gains due to the large discrepancy between the source and target data [20, 25, 65]. They also pointed out that much smaller deep learning architectures could perform comparably to the standard ImageNet models, which would further reduce the computational cost [25].

Using synthesised images by ANTs improved the image classification performance on the restricted EfficientNet-B2. Also, the improvement did not require careful fine-tuning or any additional search strategy of hyperparameters. Our results may hold true in a high computational cost situation like more complex deep learning models with larger datasets. A greater understanding of our findings may make synthetic medical images based on ANTs an alternative to GANs.

5. Conclusion

In this study, we maximised the classification performance and minimised computational cost by combining existing efficient methods, and provided a feasible solution for classifying COVID-19 CT images based on deep learning with limited computing resources and data. Although small dataset usually does not satisfy the typical requirement of standard deep learning, it can perform promising results with the help of transfer learning and data augmentation. Distributed learning frameworks such as federated learning which benefits multiple clients by aggregating model parameters instead of raw data to minimise the privacy concern, but the applications on sensitive medical data still have many ethics related problems. Hence, designing a simple tool which works well on small dataset can be applied by hospitals independently under privacy restriction is necessary. Experiments showed that synthetic images based on ANTs could improve classification performance when traditional data augmentation failed or even backfired. We highlight three features: Low-Threshold, Easy-To-Deploy and Ready-To-Use. Publicly accessible data does not usually contain high-quality images because of their size (difficult to transfer, download, and process), the lack of standards in medical imaging, the large diversity of medical imaging devices, etc However, it avoids data privacy issues and is easy to be obtained. To meet our motivation and serve for practical low-resource settings, we assemble several existing tools and technologies, such as transfer learning and interpolation, and the method we

proposed does not require advanced machine learning skills or fine-tuning experiences on deep learning models. High-performance hardware is not essential for such tools, which indicates our method is easier to be accepted and deployed in local hospitals, clinics and other medical institutions, especially in developing countries. Besides, ANTs use NIFTI (Neuroimaging Informatics Technology Initiative) files [a common format for medical images such as functional magnetic resonance imaging (fMRI)] as input to synthesise images which can be easily deployed in hospitals, and distort patient information during synthesis to ensure anonymity. The above characteristics lower the research threshold, allowing scholars and healthcare workers with few resources to have an opportunity to explore more possibilities. Although ideal laboratory condition with novel methods that lead advanced breakthroughs, our simulations address brutal reality where our method may be easily deployed in practical. Hence, this work offers a new possibility for rapid image classification to assist diagnosis in the early stages of future epidemics.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files). Data will be available from 28 February 2023.

ORCID iDs

Yikun Hou ⓘ https://orcid.org/0000-0002-9456-8224
Miguel Navarro-Cía ⓘ https://orcid.org/0000-0003-0256-6465

References

[1] WHO Coronavirus (COVID-19) Dashboard 2023 World Health Organization (https://covid19.who.int/)
[2] Huang C *et al* 2020 Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China *Lancet* **395** 497–506
[3] Dramé M *et al* 2020 Should RT-PCR be considered a gold standard in the diagnosis of COVID-19? *J. Med. Virol.* **92** 2312–3
[4] Ai T *et al* 2020 Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in china: a report of 1014 cases *Radiology* **296** E32–40
[5] Islam M M, Karray F, Alhajj R and Zeng J 2021 A review on deep learning techniques for the diagnosis of novel coronavirus (covid-19) *IEEE Access* **9** 30551–72
[6] Dong D *et al* 2021 The role of imaging in the detection and management of covid-19: a review *IEEE Rev. Biomed. Eng.* **14** 16–29

[7] Yang W *et al* 2020 The role of imaging in 2019 novel coronavirus pneumonia (COVID-19) *European Radiology* **30** 4874–82

[8] Ding X, Xu J, Zhou J and Long Q 2020 Chest CT findings of COVID-19 pneumonia by duration of symptoms *Eur. J. Radiol.* **127** 109009

[9] Han X *et al* 2021 Six-month follow-up chest CT findings after severe COVID-19 pneumonia *Radiology* **299** E177–86

[10] Gietema H A *et al* 2020 CT in relation to RT-PCR in diagnosing COVID-19 in the netherlands: a prospective study *PLoS One* **15** e0235844

[11] Horry M J *et al* 2020 COVID-19 detection through transfer learning using multimodal imaging data *IEEE Access* **8** 149808–24

[12] Mei X *et al* 2020 Artificial intelligence–enabled rapid diagnosis of patients with COVID-19 *Nat. Med.* **26** 1224–8

[13] Wang B *et al* 2021 AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system *Appl. Soft Comput.* **98** 106897

[14] Wang S *et al* 2021 A deep learning algorithm using CT images to screen for corona virus disease (COVID-19) *European radiology* **31** 6096–104

[15] Wu X *et al* 2020 Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: a multicentre study *Eur. J. Radiol.* **128** 109041

[16] Chen J *et al* 2020 Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography *Sci. Rep.* **10** 19196

[17] Ardakani A A, Kanafi A R, Acharya U R, Khadem N and Mohammadi A 2020 Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks *Comput. Biol. Med.* **121** 103795

[18] Shi F *et al* 2020 Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19 *IEEE Rev. Biomed. Eng.* **14** 4–15

[19] El-Rashidy N *et al* 2021 Comprehensive survey of using machine learning in the COVID-19 pandemic *Diagnostics* **11** 1155

[20] He X *et al* 2020 *medRxiv* 10.1101/2020.04.13.20063941 2021.12.01Sample-efficient deep learning for COVID-19 diagnosis based on CT scans

[21] Weiss K, Khoshgoftaar T M and Wang D 2016 A survey of transfer learning *Journal of Big Data* **3** 9

[22] Taylor L and Nitschke G 2018 Improving deep learning with generic data augmentation *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* 1542–7

[23] Zhao W, Jiang W and Qiu X 2021 Deep learning for COVID-19 detection based on CT images *Sci. Rep.* **11** 14353

[24] Loey M, Manogaran G and Khalifa N E M 2020 A deep transfer learning model with classical data augmentation and CGAN to detect COVID-19 from chest CT radiography digital images *Neural Computing and Applications* 1–13

[25] Raghu M, Zhang C, Kleinberg J and Bengio S 2019 Transfusion: understanding transfer learning for medical imaging *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*

[26] Alzubaidi L *et al* 2020 Towards a better understanding of transfer learning for medical imaging: a case study *Applied Sciences* **10** 4523

[27] Hussain Z, Gimenez F, Yi D and Rubin D 2017 Differential data augmentation techniques for medical imaging classification tasks *Amia Annual Symposium Proceedings* (American Medical Informatics Association) 979–84

[28] Yang X, He X, Zhao J, Zhang Y, Zhang S and Xie P 2020 COVID-CT-dataset: a CT scan dataset about COVID-19 2003.13865arXiv preprint, dataset available: (https://github.com/UCSD-AI4H/COVID-CT)

[29] Angelov P and Soares E A 2020 *medRxiv* 10.1101/2020.04.24.20078584 2021.12.01SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification (https://kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset)

[30] SARS-COV-2 Ct-Scan Dataset 2021 (https://kaggle.com/plameneduardo/sarscov2-ctscan-dataset)

[31] Goodfellow I *et al* 2014 Generative adversarial nets *Advances in Neural Information Processing Systems 27 (NIPS 2014)* 27

[32] Han C *et al* 2019 Combining noise-to-image and image-to-image GANs: brain MR image augmentation for tumor detection *IEEE Access* **7** 156966–77

[33] Calimeri F, Marzullo A, Stamile C and Terracina G 2017 Biomedical data augmentation using generative adversarial neural networks *International conference on artificial neural networks* (Springer) pp 626–34

[34] Madani A, Moradi M, Karargyris A and Syeda-Mahmood T 2018 Chest x-ray generation and data augmentation for cardiovascular abnormality classification *Medical Imaging 2018: Image Processing* 10574 (International Society for Optics and Photonics)

[35] Jiang Y, Chen H, Loew M and Ko H 2021 COVID-19 CT image synthesis with a conditional generative adversarial network *IEEE Journal of Biomedical and Health Informatics* **25** 441–52

[36] Liu S *et al* 2020 3d tomographic pattern synthesis for enhancing the quantification of covid-19 2005.01903arXiv preprint

[37] Wiatrak M, Albrecht S V and Nystrom A 2019 Stabilising generative adversarial networks: A survey 1910.00927arXiv preprint

[38] Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J and Aila T 2020 Training generative adversarial networks with limited data *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*

[39] Yi X, Walia E and Babyn P 2019 Generative adversarial network in medical imaging: A review *Med. Image Anal.* **58** 101552

[40] Avants B B, Tustison N and Song G 2009 Advanced normalisation tools (ANTS) *Insight j* **2** 1–35

[41] Avants B B, Epstein C L and Gee J C 2005 Geodesic image interpolation: Parameterizing and interpolating spatiotemporal images *International Workshop on Variational, Geometric, and Level Set Methods in Computer Vision* (Berlin: Springer) 247–58

[42] Reisenhofer R, Bosse S, Kutyniok G and Wiegand T 2018 A Haar wavelet-based perceptual similarity index for image quality assessment *Signal Process. Image Commun.* **61** 33–43

[43] Sahu P, Huang H, Zhao W and Qin H 2019 Using virtual digital breast tomosynthesis for de-noising of low-dose projection images *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (IEEE) pp 1647–51

[44] Kofler A, Dewey M, Schaeffter T, Wald C and Kolbitsch C 2020 Spatio-Temporal Deep Learning-Based Undersampling Artefact Reduction for 2D Radial Cine MRI With Limited Training Data *IEEE Trans. Med. Imaging* **39** 703–17

[45] Kofler A *et al* 2020 Neural networks-based regularisation for large-scale medical image reconstruction *Phys. Med. Biol.* **65** 135003

[46] Sabottke C F and Spieler B M 2020 The effect of image resolution on deep learning in radiography *Radiology: Artificial Intelligence* **2** e190015

[47] Han D 2013 Comparison of commonly used image interpolation methods *Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)* (Atlantis Press) pp 1556–9

[48] Hashemi M 2019 Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation *Journal of Big Data* **6** 98

[49] Yamashita R, Nishio M, Do R K G and Togashi K 2018 Convolutional neural networks: an overview and application in radiology *Insights into imaging* **9** 611–29

[50] Wang T, Zhao Y, Zhu L, Liu G, Ma Z and Zheng J 2020 Lung CT image aided detection COVID-19 based on Alexnet network *2020 5th International Conference on Communication, Image and Signal Processing (CCISP)* 199–203

[51] Walvekar S and Shinde D 2020 Detection of COVID-19 from CT images using resnet50 *2nd International Conference on Communication & Information Processing (ICCIP)*

[52] Ko H *et al* 2020 COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation *Journal of medical Internet research* **22** e19569

[53] Tan M and Le Q 2019 Efficientnet: rethinking model scaling for convolutional neural networks *International Conference on Machine Learning* (PMLR) 6105–14

[54] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *The journal of machine learning research* **15** 1929–58

[55] Shorten C and Khoshgoftaar T M 2019 A survey on image data augmentation for deep learning *Journal of Big Data* **6** 1–4860

[56] Perez L and Wang J 2017 The effectiveness of data augmentation in image classification using deep learning 1712.04621arXiv preprint

[57] Chlap P, Min H, Vandenberg N, Dowling J, Holloway L and Haworth A 2021 A review of medical image data augmentation techniques for deep learning applications *Journal of Medical Imaging and Radiation Oncology* **65** 545–63

[58] Aslan M F, Unlersen M F, Sabanci K and Durdu A 2021 CNN-based transfer learning–BiLSTM network: A novel approach for COVID-19 infection detection *Appl. Soft Comput.* **98** 106912

[59] Elgendi M *et al* 2021 The effectiveness of image augmentation in deep learning networks for detecting COVID-19: A geometric transformation perspective *Frontiers in Medicine* **8** 629134

[60] Silva P *et al* 2020 COVID-19 detection in CT images with deep learning: a voting-based scheme and cross-datasets analysis *Inform Med. Unlocked* **20** 100427

[61] Shamsi A *et al* 2021 An uncertainty-aware transfer learning-based framework for COVID-19 diagnosis *IEEE Trans Neural Netw. Learn. Syst.* **32** 1408–17

[62] Saeedi A, Saeedi M and Maghsoudi A 2020 A novel and reliable deep learning web-based tool to detect covid-19 infection from chest ct-scan 2006.14419 arXiv preprint

[63] Chattopadhyay S, Dey A, Singh P K, Geem Z W and Sarkar R 2021 COVID-19 detection by optimising deep residual features with improved clustering-based golden ratio optimiser *Diagnostics* **11** 315

[64] Sharma S 2020 Drawing insights from COVID-19-infected patients using CT scan images and machine learning techniques: a study on 200 patients *Environmental Science and Pollution Research* **27** 37155–63

[65] Chen D, Ji S, Liu F, Li Z and Zhou X 2020 A review of automated diagnosis of covid-19 based on scanning images *6th International Conference on Robotics and Artificial Intelligence* pp 97–104