

1 Development and External Validation of a 2 Mixed-Effects Deep Learning Model to 3 Diagnose COVID-19 from CT Imaging

4 Joshua Bridge, Yanda Meng, Wenyue Zhu, Thomas Fitzmaurice, Caroline McCann, Cliff Addison,
5 Manhui Wang, Cristin Merritt, Stu Franks, Maria Mackey, Steve Messenger, Renrong Sun, Yitian
6 Zhao, Yalin Zheng

7 Institute of Life Course and Medical Sciences, University of Liverpool, Liverpool, United Kingdom, L7
8 8TX Joshua Bridge PhD. Student, Yanda Meng PhD. Student, Wenyue Zhu PhD. Student, Thomas
9 Fitzmaurice Clinical Research Fellow in Respiratory Medicine, Yalin Zheng Reader in Ophthalmic
10 Imaging. Department of Respiratory Medicine, Liverpool Heart and Chest Hospital NHS Foundation
11 Trust, Liverpool, United Kingdom, L14 3PE Thomas Fitzmaurice Clinical Research Fellow in
12 Respiratory Medicine. Department of Radiology, Liverpool Heart and Chest Hospital NHS Foundation
13 Trust, United Kingdom, L14 3PE Caroline McCann Consultant Cardiothoracic Radiologist. Advanced
14 Research Computing, University of Liverpool, Liverpool, United Kingdom, L69 3BX Cliff Addison,
15 Manhui Wang. Alces Flight Limited, Bicester, United Kingdom, OX26 4PP Cristin Merritt, Stu Franks.
16 Amazon Web Services, 60 Holborn Viaduct, London EC1A 2FD Maria Mackey, Steve Messenger.
17 Department of Radiology, Hubei Provincial Hospital of Integrated Chinese and Western Medicine,
18 Hubei University of Chinese Medicine, Wuhan 430000, China Renrong Sun. Cixi Institute of
19 Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese
20 Academy of Sciences, Ningbo 315201, China Yitian Zhao.

21
22 Correspondence to:

23 Dr Yalin Zheng,
24 Institute of Life Course and Medical Sciences,
25 University of Liverpool,
26 William Henry Duncan Building,
27 6 West Derby Street,
28 Liverpool,
29 L7 8TX,
30 Yalin.Zheng@liverpool.ac.uk
31

1 Abstract

2 Objectives

3 To develop and externally geographically validate a mixed-effects deep learning model to diagnose
4 COVID-19 from computed tomography (CT) imaging following best practice guidelines and assess the
5 strengths and weaknesses of deep learning COVID-19 diagnosis.

6 Design

7 Model development and external validation with retrospectively collected data from two countries.

8 Setting

9 Hospitals in Moscow, Russia, collected between March 1, 2020, and April 25, 2020. The China
10 Consortium of Chest CT Image Investigation (CC-CCTI) collected between January 25, 2020, and
11 March 27, 2020.

12 Participants

13 1,110 and 796 patients with either COVID-19 or healthy CT volumes from Moscow, Russia, and
14 China, respectively.

15 Main outcome measures

16 We developed a deep learning model with a novel mixed-effects layer to model the relationship
17 between slices in CT imaging. The model was trained on a dataset from hospitals in Moscow, Russia,
18 and externally geographically validated on a dataset from a consortium of Chinese hospitals. Model
19 performance was evaluated in discriminative performance using the area under the receiver
20 operating characteristic (AUROC), sensitivity, specificity, positive predictive value (PPV), and negative
21 predictive value (NPV). In addition, calibration performance was assessed using calibration curves,
22 and clinical benefit was assessed using decision curve analysis. Finally, the model's decisions were
23 assessed visually using saliency maps.

24 Results

25 External validation on the large Chinese dataset showed excellent performance with an AUROC of
26 0.956 (95% CI: 0.943, 0.970), with a sensitivity and specificity, PPV, and NPV of 0.879 (0.852, 0.906),
27 0.942 (0.913, 0.972), 0.988 (0.975, 1.00), and 0.732 (0.650, 0.814).

28 Conclusions

29 Deep learning can reduce stress on healthcare systems by automatically screening CT imaging for
30 COVID-19. However, deep learning models must be robustly assessed using various performance
31 measures and externally validated in each setting. In addition, best practice guidelines for
32 developing and reporting predictive models are vital for the safe adoption of such models.

33

1 Statements

2 The authors do not own any of the patient data, and ethics approval was not needed. The lead
3 author affirms that this manuscript is an honest, accurate, and transparent account of the study
4 being reported, that no important aspects of the study have been omitted, and that any
5 discrepancies from the study as planned (and, if relevant, registered) have been explained. Patients
6 and the public were not involved in the study.

7 Funding

8 This study was funded by EPSRC studentship (No. 2110275), EPSRC Impact Acceleration Account
9 (IAA) funding, and Amazon Web Services.

10 Summary

11 What is already known on this topic

- 12 • Deep learning can diagnose diseases from imaging data automatically
- 13 • Many studies using deep learning are of poor quality and fail to follow current best practice
14 guidelines for the development and reporting of predictive models
- 15 • Current methods do not adequately model the relationship between slices in CT volumetric
16 data

17 What this study adds

- 18 • A novel method to analyse volumetric imaging data composed of slices such as CT images
19 using deep learning
- 20 • Model developed following current best-practice guidelines for the development and
21 reporting of prediction models

22

1 Introduction

2 Coronavirus disease 2019 (COVID-19) is an infectious respiratory disease caused by severe acute
3 respiratory syndrome coronavirus 2 (SARS-CoV-2). Virus clinical presentation ranges from mild cold-
4 like symptoms to severe viral pneumonia, which can be fatal¹. While some countries have achieved
5 relative control through lockdowns, future outbreaks and new strains are expected to continue, with
6 many experts believing the virus is here to stay². Detection and isolation is the most effective way to
7 prevent further spread of the virus. Even with effective vaccines becoming widely available, with the
8 threat of continued waves and new potentially vaccine-resistant variants, it is vital to further
9 develop diagnostic tools for COVID-19. These tools will likely also apply to future outbreaks of other
10 similar diseases as well as common diseases such as pneumonia.

11 The diagnosis of COVID-19 is usually determined by Reverse Transcription Polymerase Chain
12 Reaction (RT-PCR), but this is far from being a gold standard. A negative test does not necessarily
13 indicate a negative diagnosis, with one recent review finding that RT-PCR has a real-world sensitivity
14 of around 70% and a specificity of 95%³. Furthermore, an individual patient data systematic review⁴
15 found that RT-PCR often fails to detect COVID-19, and early sampling is key to reducing false
16 negatives. Therefore, these tests are often more helpful to rule in COVID-19 rather than ruling out.
17 If a patient presents with symptoms of COVID-19, but an RT-PCR test is negative, then further tests
18 are often required¹. Consecutive negative tests with at least a one-day gap are recommended;
19 however, this still does not guarantee that the patient is negative for COVID-19⁵. Computed
20 tomography (CT) can play a significant role in diagnosing COVID-19⁶. Given the excessive number of
21 COVID-19 cases worldwide and the strain on resources expected, automated diagnosis might reduce
22 the burden on reporting radiologists.

23 CT images are made up of many slices, creating a three dimensional (3D)-like structure. Previous
24 approaches, such as those used by Li et al.⁷ and Bai et al.⁸, treat the image as separate slices and use
25 a pooling layer to concatenate the slices. An alternative approach assumes the slices form a 3D
26 structure and use a 3D CNN, such as that proposed in CoviNet⁹. A fundamental limitation of these
27 methods is the need for the same number of slices as their inputs, but the number of slices often
28 varies between different CT volumes. Instead, we propose using a novel mixed-effects layer to
29 consider the relationship between slices in each scan. Mixed-effects models are commonly used in
30 traditional statistics^{10,11}, but we believe this is the first time that mixed-effects models have been
31 utilised in such a way. It has been observed that some lobes of the lung are more often affected by
32 COVID-19 than others^{12,13} with lower lobe distribution being a prominent feature of COVID-19¹⁴, the
33 fixed-effects take this into account by considering where each slice is located within the scan.

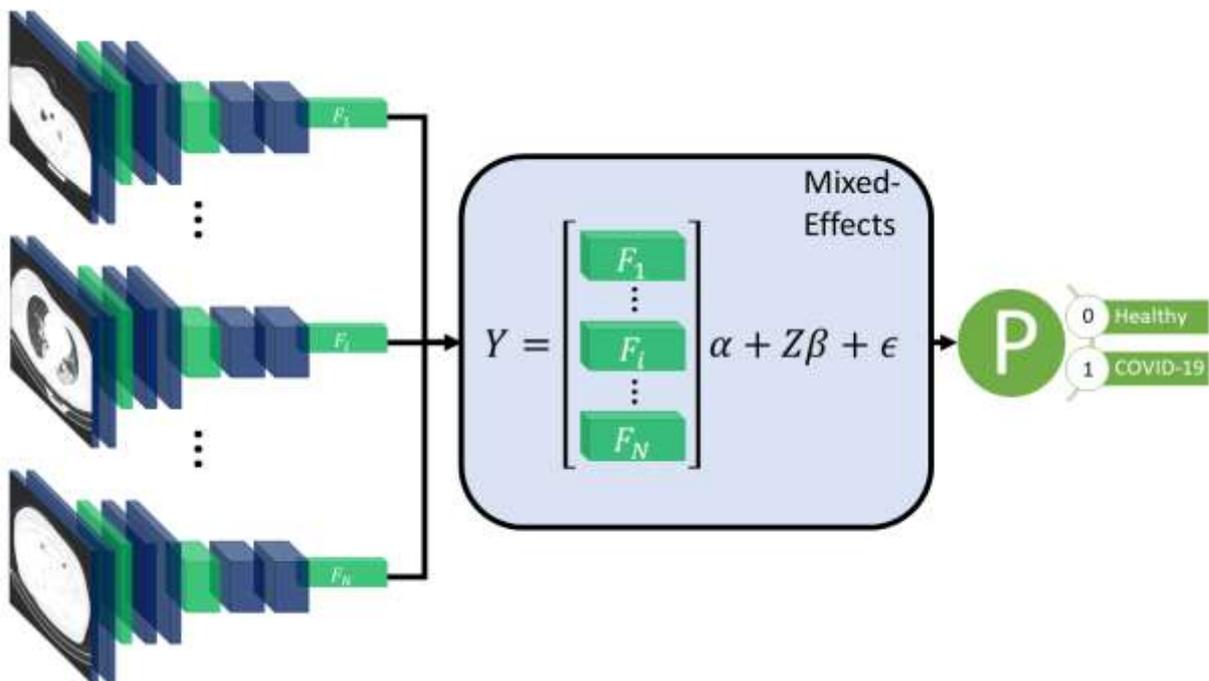
34 Deep learning has shown great potential in the automatic classification of disease, often achieving
35 expert-level performance. Such models could screen and monitor COVID-19 by automatically
36 analysing routinely collected CT images. As observed by Wynants et al.¹⁵ and Roberts et al.¹⁶, many
37 models are already developed to diagnose COVID-19, which often obtain excellent discriminative
38 performance; however, very few of these models, if any, are suitable for clinical use, mainly due to a
39 lack of robust analysis and reporting. These models often suffer from common pitfalls, making them
40 unsuitable for broader adoption. Roberts et al.¹⁶ identified three common areas in which models
41 often fail these are: (1) a lack of adequately documented methods for reproducibility, (2) failure to
42 follow established guidelines and best practices for the development of deep learning models, and
43 (3) an absence of external validation displaying the model's applicability to a broader range of data
44 outside of the study sample. Failure to address these pitfalls leads to profoundly flawed and biased
45 models, making them unsuitable for deployment.

1 In this work, we aim to address the problems associated with previous models by following
2 guidelines for the reporting^{17,18} and development¹⁹ of prediction models to ensure that we have
3 rigorous documentation allowing the methods developed here to be replicated. In addition, we will
4 make code and the trained model publicly available at [github.com/JTBridge/ME-COVID19] to
5 promote reproducible research and facilitate adoption. Finally, we use a second dataset from a
6 country other than the development dataset to externally validate the model and report a range of
7 performance measures evaluating the model's discrimination, calibration, and clinical usefulness.

8 Hence, our main aim is to develop a mixed-effects deep learning model to accurately classify images
9 as healthy or COVID-19, following best practice guidelines. Our secondary aim is to show how deep
10 learning predictive algorithms can satisfy current best practice guidelines to create reproducible and
11 less biased models.

12 Methods

13 Our proposed method consists of a feature extractor and a two-stage generalised linear mixed-
14 effects model (GLMM)²⁰, with all parameters estimated within the deep learning framework using
15 backpropagation. First, a series of CT slices forming a CT volume is input to the model. In our work,
16 we use 20 slices. Next, a convolutional neural network (CNN) extract relevant features from the
17 model and creates a feature vector for each CT slice. Then, a mixed-effects layer concatenates the
18 feature vectors into a single vector. Finally, a fully connected layer followed by a sigmoid activation
19 gives a probability of COVID-19 for the whole volume. The mixed effects and fully connected layer
20 with sigmoid activation are analogous to a linear GLMM in traditional statistics. The overall
21 framework is shown in Figure 1.



22
23 **Figure 1: Diagram of the overall framework.** Twenty slices are chosen from a CT volume. Each slice is
24 fed into a CNN with shared weights, which outputs a feature vector of length 2048 for each image.
25 The feature vectors form a 20-by-2048 fixed effects matrix, X, for the GMM model with a random-
26 effects matrix, Z, consisting of an identity matrix. A mixed-effects model is used to model the
27 relationship between slices. Finally, a fully connected layer and sigmoid activation return a
28 probability of the diagnosis.

1 Feature extractor

2 For the feature extractor, we use a CNN. In this work, we chose InceptionV3²¹ as it is relatively
3 efficient and commonly used. InceptionV3 outputs a feature vector of length 2048. To reduce the
4 time needed to reach convergence, we pretrained the CNN on ImageNet²². A CNN is used for each
5 slice, with shared weights between CNNs; this reduces the amount of computational power
6 required. Following the CNN, we used a global average pooling layer to reduce each image to a
7 feature vector for each slice. We then added a dropout of 0.6 to improve generalizability to unseen
8 images. We form the feature vectors into a matrix of shape 20×2048 . Although we used
9 InceptionV3²¹ here, other networks would also work and may provide better performance on other
10 similar tasks. We then need to concatenate these feature vectors into a single feature vector for the
11 whole volume; normally, pooling is used, in our work we propose using a mixed-effects models.

12 Mixed-effects network

13 We propose utilising a novel mixed-effects layer to model the relationship between slices. Mixed-
14 effects models are a statistical model consisting of a fixed-effects part and a random-effects part.
15 The fixed-effects part models the relationship within the CT slice; the random effects can model the
16 spatial correlation between CT slices within the same image¹¹. For volumetric data, the number of
17 slices may differ significantly due to various factors such as imaging protocol and the size of the
18 patient. Some volumes may have fewer images than the model is designed to use, which leads to
19 missing data. Mixed-effects models can deal with missing data provided the data are missing at
20 random. The mixed-effects model is given by

$$21 \quad Y_i = X_i\alpha + Z_i\beta + e_i$$

22 where Y_i , X_i , Z_i , e_i are vectors of outcomes, fixed effects design matrix of shape 20×2048 ,
23 random effects design matrix of shape 20×20 , and vector of error unknown random errors of the
24 i th patient of shape 20, respectively, and α , β are fixed and random effects parameters, both of
25 length 20. We assume that the random effects β are normally distributed with 0 and variance G

$$26 \quad \beta \sim N(0, G)$$

27 We also assume independence between the random effects and the error term.

28 The fixed effects design matrix, X , is made up of the feature vectors output from the feature
29 extraction network. For the random effects design matrix, Z , we use an identity matrix with the
30 same size as the number of slices; in our experiments, this is 20. The design matrix is then given by

$$31 \quad Z_{20 \times 20} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix}.$$

32 This matrix easily generalises to any number of slices. If the distance between slices is not uniform,
33 the values can be altered accordingly. We assumed no particular correlation matrix. We included the
34 fixed and random intercept in the model. All parameters for the mixed-effects layer were initialised
35 using the Gaussian distribution with mean 0 and standard deviation 0.05.

36 A type of mixed-effects modelling has previously been combined with deep learning for gaze
37 estimation²³. However, their mixed-effects method is very different from our proposed method;
38 they used the same design matrix for fixed and random effects. In addition, they also estimated
39 random-effects parameters with an expectation-maximisation algorithm, which was separate from
40 the fixed effects estimation, which used deep learning. In our work, we utilise a spatial design matrix

1 to model the spatial relationship between slices and estimate parameters within the deep learning
2 framework using backpropagation without the need for multiple stages.

3 Loss function

4 As the parameters in the model are all estimated using backpropagation, we must ensure that the
5 assumption of normally distributed random effects parameters with mean zero is valid. We achieve
6 this by introducing a loss function for the random effects parameters, which enforces a mean,
7 skewness, and excess kurtosis of 0. We measure skewness using the adjusted Fisher–Pearson
8 standardised moment coefficient

$$9 \quad Skew(\beta) = \frac{\sqrt{n(n-1)}}{n-2} \frac{E[(\beta - \bar{\beta})^3]}{(E[(\beta - \bar{\beta})^2])^{3/2}}$$

10 and the excess kurtosis using

$$11 \quad Kurt(\beta) - 3 = \frac{1}{n^2} \sum_{i=1}^n \left(\frac{E[(\beta - \bar{\beta})^4]}{(E[(\beta - \bar{\beta})^2])^2} - 3 \right),$$

12 where n is the length of β , $\bar{\beta}$ is the mean of β and $E[\cdot]$ is the expectation function. The Gaussian
13 distribution has a kurtosis of 3; therefore, the excess kurtosis is given by the kurtosis minus 3. This
14 formula for this fixed-effects parameters loss function which we aim to minimise is then given by

$$15 \quad L_{fixed} = |E(\beta) + Skew(\beta) + Kurt(\beta) - 3|.$$

16 For the classification, we use the Brier Score²⁴ as the loss function, which is given by

$$17 \quad L_{Brier} = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2$$

18 where N is the total number of samples, p_i is the predicted probability of sample i and o_i is the
19 observed outcome of sample i . The Brier score is the same as the mean squared error of the
20 predicted probability.

21 We chose to use the Brier Score over the more commonly used binary cross-entropy because it can
22 be decomposed into two components: refinement and calibration. Calibration is often overlooked in
23 deep learning models but is vital to assess the safety of any prediction model. The refinement
24 component combines the model's resolution and uncertainty and measures the model's
25 discrimination. The calibration component can be used as a measure of the model calibration.
26 Therefore the Brier Score can be used to optimize both the discrimination and calibration of the
27 model. The overall loss function is given by

$$28 \quad L = L_{Brier} + L_{fixed}.$$

29 A scaling factor could be introduced to weight one part of the loss function as more important than
30 the other; however, we give both parts of the loss function equal weighting in our work.

31 We also transformed the labels as suggested by Platt²⁵ to reduce overfitting. The negative and
32 positive labels become

$$33 \quad o_- = \frac{1}{N_- + 2}$$

1 and

$$2 \quad o_+ = \frac{N_+ + 1}{N_+ + 2}$$

3 respectively, where N_- and N_+ are the total number of negative and positive cases in the training
4 set. This is similar to label smoothing as commonly used in deep learning, but the new targets are
5 chosen by applying Baye's Rule to the out-of-sample data to prevent overfitting.

6 Classification layer

7 The output of the mixed-effects layer is a single vector, which is the same length as the number of
8 slices used. For example, in our work, we had a vector of length 20. Furthermore, we used a fully
9 connected layer with sigmoid activation to obtain a probability of the scan showing COVID-19; the
10 sigmoid activation is analogous to the logistic link function in traditional statistics. Finally, we added
11 an L1 regularisation term of 0.1 and an L2 regularisation term of 0.01 to the kernel to reduce
12 overfitting.

13 Model performance

14 Many deep learning models focus on assessing discriminative performance only, using measures
15 such as the area under the receiver operating characteristic curve (AUROC), sensitivity, and
16 specificity. To better understand the model performance and impact, we report performance
17 measures in three broad areas: discrimination, calibration, and clinical usefulness²⁶. Discrimination
18 assesses how well a model can discriminate between healthy and COVID-19 positive patients.
19 Models with excellent discriminative performance can still produce unreliable results, with vastly
20 overestimated probabilities regardless of the true diagnosis²⁷. Model calibration is often overlooked
21 and rarely reported in deep learning, if at all; however, poorly calibrated models can be misleading
22 and lead to dangerous clinical decisions²⁷. Calibration can be assessed using four levels, with each
23 level indicating better calibration than the last²⁸. The fourth and most stringent level (strong
24 calibration) requires the correct model to be known, which in turn requires predictors to be non-
25 continuous, and an infinite amount of data to be used and is therefore considered utopic. We
26 consider the third level (moderate calibration) using calibration curves. Moderate calibration will
27 ensure that the model is at least not clinically harmful. Finally, measures of clinical usefulness assess
28 the clinical consequences of the decision and acknowledge that a false positive may be more or less
29 severe than a false negative.

30 Firstly, the discriminative performance is assessed using AUROC using the pROC package in R²⁹, with
31 confidence intervals constructed using DeLong's³⁰ method. For sensitivity, specificity, positive
32 predictive value (PPV), and negative predictive value (NPV), we use the ReportROC³¹ package in R²⁹;
33 with 95% confidence intervals constructed using the simple asymptotic formula. Secondly, we assess
34 model calibration using calibration curves created using the CalibrationCurves package²⁸, which is
35 based on the rms³² package. Finally, we assess the clinical usefulness of the model using decision
36 curve analysis³³. Net benefits are given at various thresholds, and models which reach zero net
37 benefit at higher thresholds are considered more clinically useful. Two brief sensitivity analyses are
38 performed, one assessing the model's ability to deal with missing data and the other assessing its
39 ability to deal with noise. To improve the model's interpretability and reduce the black-box nature,
40 we produce saliency maps³⁴ that show which areas of the image are helpful to the model in the
41 prediction. We also check the assumption of normally distributed random-effects parameters.

1 Comparison models

2 To assess the added benefit of using our mixed-effects method, we compare against networks that
3 use alternative methods. Both COVNet⁷ and a method proposed by Bai et al.⁸ propose deep learning
4 models that consider the slices separately before concatenating the features using max pooling.
5 COVNet uses a ResNet50³⁵ CNN to extract features and pooling layers to concatenate the features
6 before a fully connected classification layer. The model proposed by Bai et al. uses EfficientNetB4³⁶
7 to extract features followed by a series of full-connected layers with batch normalisation and
8 dropout; average pooling is then used to concatenate the feature vectors before classification. While
9 max pooling is simple and computationally efficient, it cannot deal with pose variance and does not
10 model the relationship between slices.

11 An alternative method to pooling is treating the scans as 3D, such as in CoviNet³⁷. CoviNet takes the
12 whole scan and uses a 16 layer 3D CNN followed by pooling and fully connected layers. We
13 implemented these models as described in their respective papers. Data augmentation was the
14 same for all experiments for fair comparisons.

15 In all comparison experiments, we kept hyperparameters, such as learning rate, learning rate decay,
16 and data augmentation, the same to ensure the comparisons were fair. For COVNet⁷ and the model
17 proposed by Bai et al.⁸, we pretrained the CNNs on ImageNet as they also did; however, no
18 pretrained models were available for CoviNet. For the loss function, we also used the Brier score²⁴.

19 Computing

20 Models were developed using an Amazon Web Services p3.8xlarge node with four Tesla V100 16GiB
21 GPUs and 244GiB available memory. Model inference was performed on a local Linux machine
22 running Ubuntu 18.04, with a Titan X 12GiB GPU and 32GiB available memory. Model development
23 and inference were performed using Tensorflow 2.4^{38 39}, and R 4.0.5²⁹ was used to produce
24 evaluation metrics^{31 40} and graphs^{32 41}. We used mixed precision to reduce the computational cost,
25 which uses 16-bit floating-point precision in all layers, except for the mixed-effects and classification
26 layers, where 32-bit floating-point precision is used.

27 We used the Adam optimiser⁴² with an initial learning rate of 1e-4; if the internal validation loss did
28 not improve for three epochs, we reduced the learning rate to 20%. In addition, we assumed
29 convergence and stopped training if the loss did not improve for ten epochs to reduce the time
30 spent training and the energy used.

31 Data

32 There is currently no established method for estimating the sample size estimate in deep learning.
33 We propose treating the final fully connected classification layer as the model and treating previous
34 layers as feature extraction. We can then use the number of parameters in the final layer to estimate
35 the required sample size. Using the 'pmsampsize' package⁴³ in R, we estimate the required minimum
36 sample size in the development set. We use a conservative expected C-statistic of 0.8, with 21
37 parameters and an estimated disease prevalence of 80% based on datasets used in other studies.
38 This gives a minimum required sample size of 923 patients in the training set. For model validation,
39 around 200 patients with the disease and 200 patients without the disease are estimated to be
40 needed to assess calibration²⁸.

41 All data used here is retrospectively collected and contains hospital patients with CT scans
42 performed during the COVID-19 pandemic. The diagnosis was determined by examining radiological
43 features of the CT scan for signs of COVID-19, such as ground-glass opacities. For model
44 development, we use the MosMed dataset⁴⁴, which consists of a total of 1,110 CT scans displaying

1 either healthy or COVID-19 infected lungs. The scans were performed in Moscow hospitals between
2 March 1, 2020, and April 25, 2020. We split the dataset into two sets for training and internal
3 validation on the patient level. The training set is used to train the model, and the internal validation
4 set is used to select the best model based on the loss at each epoch; this helps prevent overfitting on
5 the training set. In addition, we obtained images from a publicly available dataset published by
6 Zhang et al.⁴⁵ consisting of CT images from a consortium of Chinese hospitals.

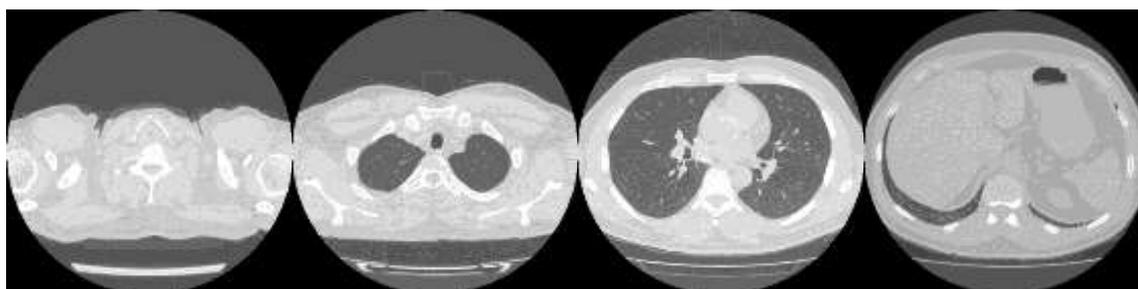
7 Overall, this allows us to perform external geographical validation in another country and to better
8 evaluate the developed model. In addition, we will be able to assess how well a deep learning model
9 generalises to other populations. A summary of all the datasets used is shown in Table 1. We have
10 923 patients in the training set and at least 200 patients in each class for the external validation set.

11 **Table 1:** Summary of the datasets used.

Dataset	Location	Use	Healthy/COVID19
MosMed Training	Moscow, Russia	Training	169/856
MosMed Validation	Moscow, Russia	Internal Validation	85/285
Zhang et al. ⁴⁵	China	External Validation	243/553

12

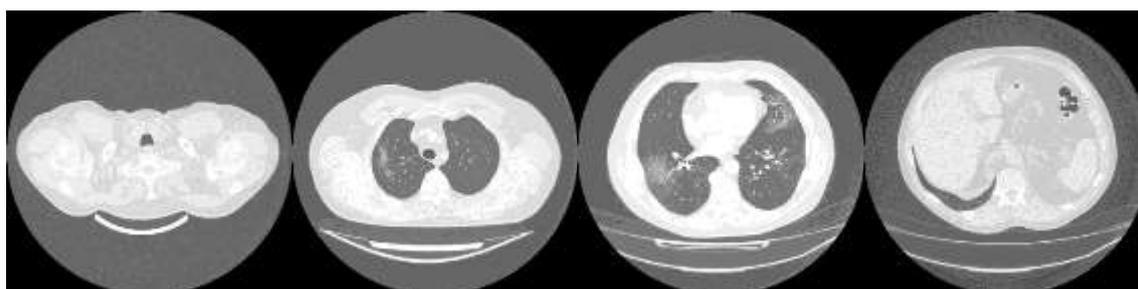
13



14

15

(a)



16

17

(b)

18 **Figure 2:** Example images showing (a) healthy and (b) COVID-19 lungs taken from the Mosmed
19 dataset.

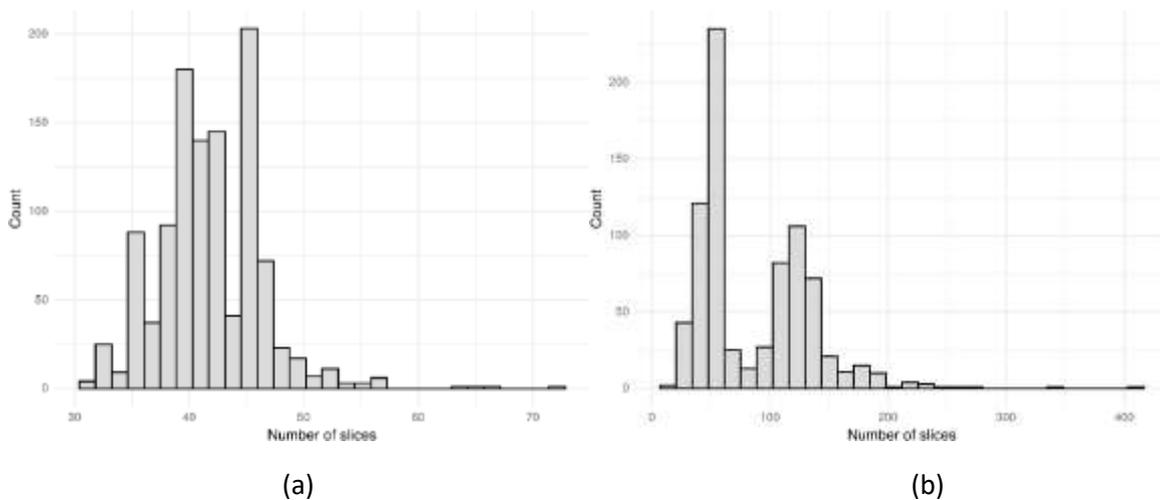
20 Data pre-processing and augmentation

21 The MosMed dataset was converted from Dicom image format into PNG, normalised to have a mean
22 of 120 and a variance of 95. Images were ordered from the top of the lungs to the bottom. During
23 training, we applied random online data augmentation to the images. This alters the image slightly
24 and gives the effect of increasing the training dataset size, although this is not as good as expanding
25 the training dataset with more samples. First, we adjusted the brightness and contrast between 70%

1 and 130%. We then rotated the image plus or minus 45 degrees and cropped the image up to 20%
2 on each side. Finally, we flipped the image horizontally and vertically with a probability of 50% each.
3 All random values were chosen using the uniform distribution except for the flips, which were
4 chosen using a random bit.

5 The dataset taken from Zhang et al.⁴⁵ required a large amount of sorting to be made suitable for use.
6 Some of the scans were pre-segmented and only showed the lung areas, while others showed the
7 whole CT scan. We removed any pre-segmented images. Identifying information on some images
8 had to be cropped to reduce bias in the algorithm. In addition, many of the scans were duplicates
9 but were not labelled as such, and many scans were incomplete, only showing a few lung slices or
10 not showing any lung tissue at all. We only used complete scans with one scan per patient. Finally,
11 some scans needed to be ordered top to bottom. Using the bilinear sampling algorithm, all images
12 were resized to 256 by 256 pixels, and image values were divided by 255 to normalise between 0
13 and 1.

14 The MosMed dataset has a median of 41 slices, a minimum of 31 slices and a maximum of 72 slices.
15 The Zhang et al. dataset has much greater variability in scan size with a median of 61 slices, a
16 minimum of 19 slices, and a maximum of 415 slices. We present histograms showing the number of
17 slices per scan in Figure 3. We require a fixed number of slices as input, and we chose 20 as the slice
18 size. For all scans, we included the first and last images. If scans had more than 20 slices, we sampled
19 uniformly to select 20. Only one scan in the Zhang et al. dataset had less than 20 slices; a blank slice
20 replaced this slice; the mixed-effects model can account for missing data.



23 **Figure 3:** Histogram showing the number of slices per scan for (a) the MosMed⁴⁴ dataset and (b) the
24 Zhang et al.⁴⁵ dataset. The MosMed dataset has much fewer slices on average with a much smaller
25 spread.

26 Results

27 On the internal validation dataset, the proposed model attained an AUROC of 0.936 (95%CI: 0.910,
28 0.961). Using an optimal cut-off point of 0.740, the sensitivity, specificity, NPV, and PPV were 0.807
29 (0.761, 0.853), 0.953 (0.908, 0.853), 0.983 (0.966, 1.0), and 0.596 (0.513, 0.678), respectively. The
30 model proposed by Bai et al.⁸ attained an AUROC of 0.737 (0.674, 0.80). Using an optimal cut-off
31 point of 0.431, the sensitivity, specificity, NPV, and PPV were 0.747 (0.697, 0.798), 0.659 (0.558,
32 0.760), 0.880 (0.839, 0.921), and 0.438 (0.352, 0.523), respectively. Covinet⁹ attained an AUROC of
33 0.768 (0.713, 0.824). Using an optimal cut-off point of 0.501, the sensitivity, specificity, NPV, and

1 PPV were 0.691 (0.638, 0.745), 0.741 (0.648, 0.834), 0.90 (0.860, 0.939), and 0.417 (0.339, 0.496),
2 respectively. COVNet⁷ attained an AUROC of 0.935 (0.912, 0.959). Using an optimal cut-off point of
3 0.428, the sensitivity, specificity, NPV, and PPV were 0.825 (0.780, 0.869), 0.988 (0.965, 1.0), 0.996
4 (0.987, 1.0), and 0.627 (0.545, 0.709), respectively. Full results are shown in Table 2.

5 Calibration curves in Figure 5 show reasonable calibration for the mixed-effects model, although
6 recalibration may improve performance. The other models do not have good calibration and likely
7 provide harmful predictions. The decision curve in Figure 6 shows that the model is of great clinical
8 benefit compared to the treat all and treat none approach.

9 It is vital to remember that the model was selected using this internal testing set to avoid overfitting
10 on the training set; therefore, these results are biased, and the external validation results are more
11 representative of the true model performance.

12 **Table 2:** Area under the receiver operating characteristic curve (AUROC), sensitivity, specificity,
13 positive predictive value (PPV) and negative predictive value (NPV) on the internal validation
14 dataset. Point estimates and 95% confidence intervals were calculated using De Long's method for
15 AUROC and 2000 sample bootstrapping for sensitivity, specificity, PPV, and NPV. Full results are
16 displayed in Table 2.

Model	AUROC	Sensitivity	Specificity	PPV	NPV
Bai et al. ⁸	0.731 (0.674, 0.80)	0.747 (0.697, 0.798)	0.659 (0.558, 0.760)	0.880 (0.839, 0.921)	0.438 (0.352, 0.523)
CoviNet ⁹	0.768 (0.713, 0.824)	0.691 (0.638, 0.745)	0.741 (0.648, 0.834)	0.90 (0.860, 0.939)	0.417 (0.339, 0.496)
CovNet ⁷	0.935 (0.912, 0.959)	0.825 (0.780, 0.869)	0.988 (0.965, 1.0)	0.996 (0.987, 1.0)	0.627 (0.545, 0.709)
Mixed-Effects (Ours)	0.936 (0.910, 0.961)	0.807 (0.761, 0.853)	0.953 (0.908, 0.998)	0.983 (0.966, 1.0)	0.596 (0.513, 0.678)

17

18 On the external geographical validation dataset, the proposed model attained an AUROC of 0.930
19 (0.914, 0.947). Using an optimal cut-off point of 0.878, the sensitivity, specificity, NPV, and PPV
20 were 0.758 (0.722, 0.793), 0.963 (0.939, 0.987), 0.979 (0.965, 0.993), and 0.636 (0.587, 0.685),
21 respectively. The model proposed by Bai et al.⁸ attained an AUROC of 0.805 (0.774, 0.836). Using an
22 optimal cut-off point of 0.434, the sensitivity, specificity, NPV, and PPV were 0.716 (0.679, 0.754),
23 0.778 (0.726, 0.830), 0.880 (0.850, 0.910), and 0.546 (0.494, 0.599), respectively. CoviNet⁹ attained
24 an AUROC of 0.734 (0.698, 0.770). Using an optimal cut-off point of 0.505, the sensitivity, specificity,
25 NPV, and PPV were 0.675 (0.635, 0.714), 0.687 (0.629, 0.746), 0.831 (0.796, 0.865), and 0.481 (0.429,
26 0.534), respectively. COVNet⁷ attained an AUROC of 0.808 (0.775, 0.841). Using an optimal cut-off
27 point of 0.995, the sensitivity, specificity, NPV, and PPV were 0.826 (0.795, 0.858), 0.679 (0.620,
28 0.738), 0.854 (0.824, 0.884), and 0.632 (0.574, 0.691), respectively. Full results are shown in Table 3.

29 Similar to the internal validation, Figure 7 shows reasonable calibration for the mixed-effects model,
30 although some recalibration may improve performance. Again, the comparison models could give
31 harmful predictions as they are poorly calibrated. The decision curve in Figure 8 shows that the
32 model is of great clinical benefit compared to the treat all and treat none approach.

33 Although our proposed method and the Covnet model showed similar performance on the internal
34 validation set, the Covnet model could not generalise to the external geographical validation set, and
35 calibration showed that the Covnet model would provide harmful risk estimates. This highlights the

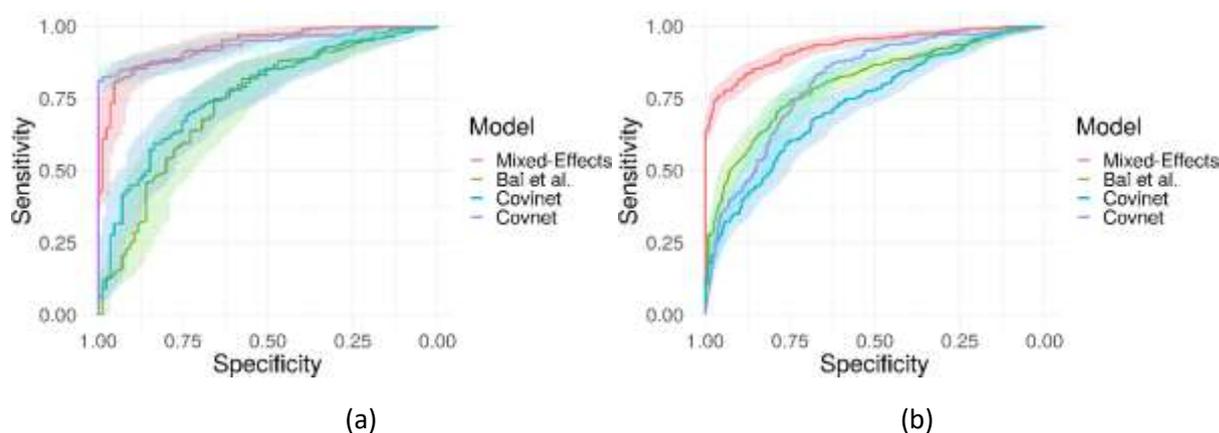
1 need for robust external validation in each intended setting. Nevertheless, the results show that the
2 proposed method better generalises to external geographical datasets and provides less harmful
3 predictions when compared to the four previously proposed methods.

4

5 **Table 3:** Area under the receiver operating characteristic curve (AUROC), sensitivity, specificity,
6 positive predictive value (PPV) and negative predictive value (NPV) on the external validation
7 dataset. Point estimates and 95% confidence intervals were calculated using De Long's method for
8 AUROC and 2000 sample bootstrapping for sensitivity, specificity, PPV, and NPV.

Model	AUROC	Sensitivity	Specificity	PPV	NPV
Bai et al. ⁸	0.805 (0.774, 0.836)	0.716 (0.679, 0.754)	0.778 (0.726, 0.830)	0.880 (0.850, 0.910)	0.546 (0.494, 0.599)
CoviNet ⁹	0.734 (0.698, 0.770)	0.675 (0.635, 0.714)	0.687 (0.629, 0.746)	0.831 (0.796, 0.865)	0.481 (0.429, 0.534)
CovNet ⁷	0.845 (0.818, 0.873)	0.801 (0.768, 0.834)	0.741 (0.686, 0.796)	0.875 (0.847, 0.904)	0.621 (0.565, 0.677)
Mixed-Effects (Ours)	0.930 (0.914, 0.947)	0.758 (0.722, 0.793)	0.963 (0.939, 0.987)	0.979 (0.965, 0.993)	0.636 (0.587, 0.685)

9



10

11

12 **Figure 4:** Receiver operating characteristic curves for (a) the MosMed internal validation set and (b)
13 the Zhang et al.⁴⁵ external validation set.

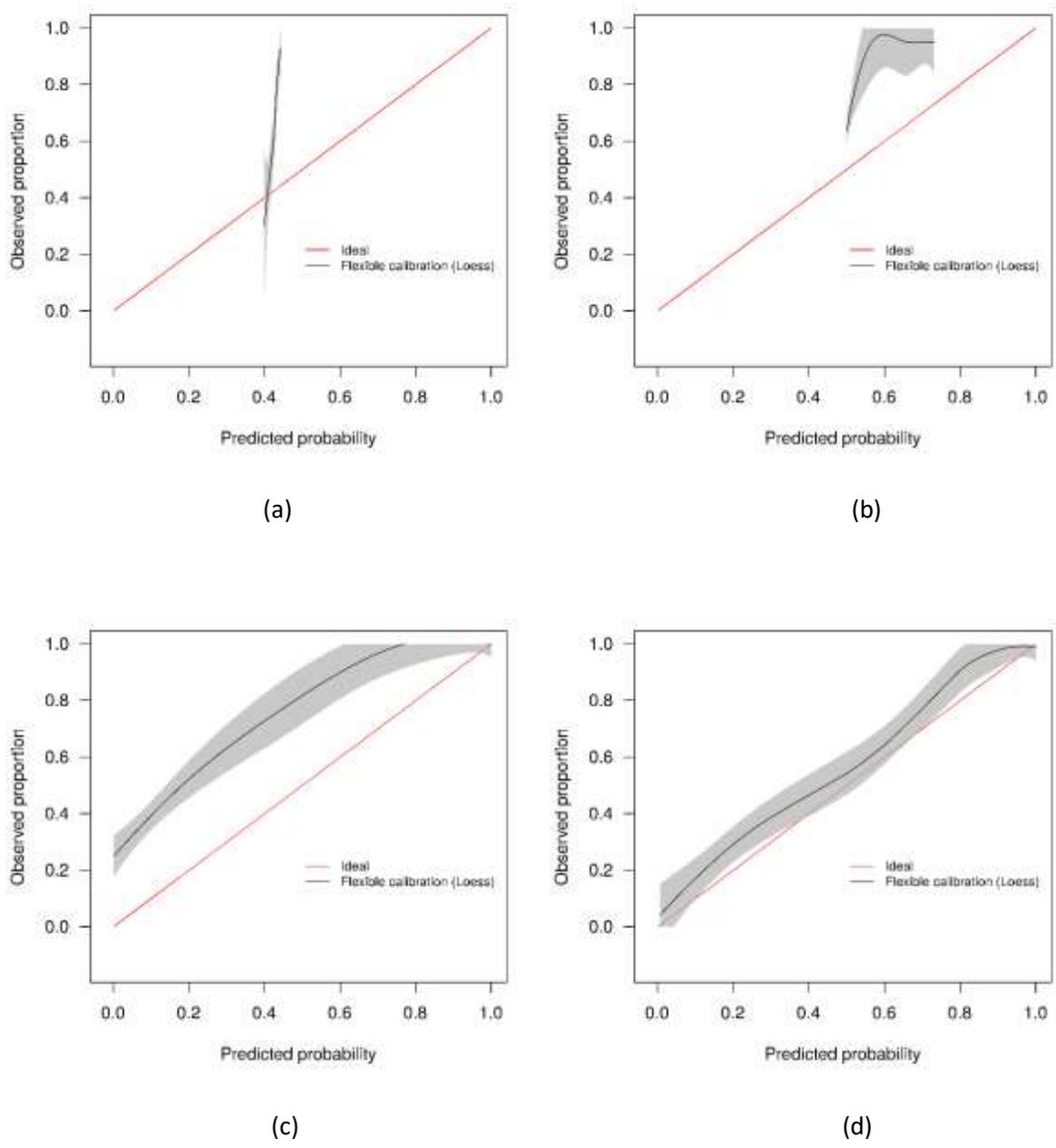
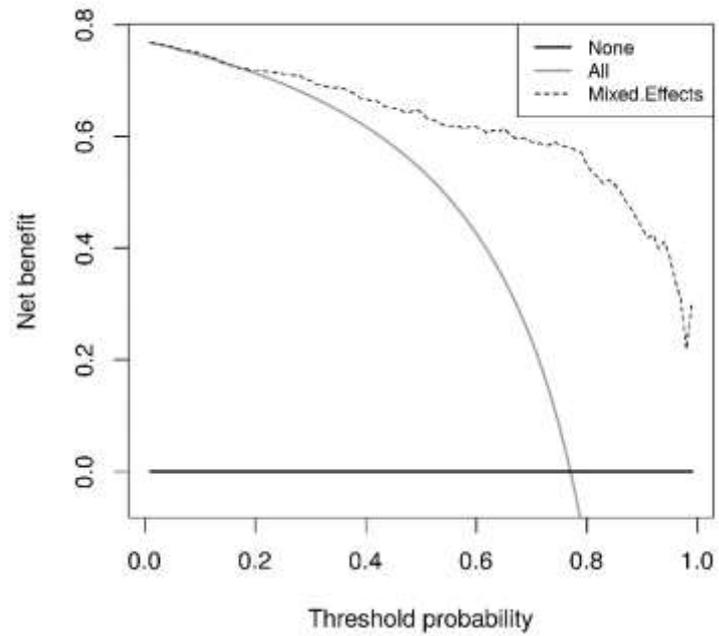
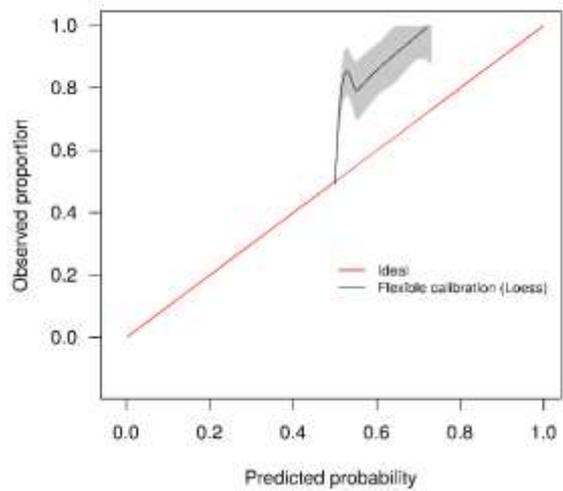
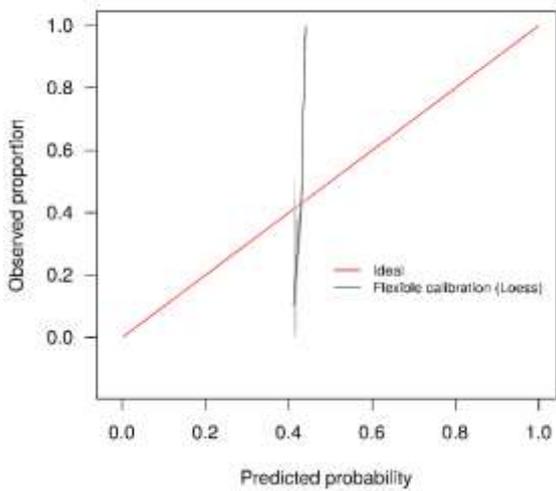


Figure 5: Calibration curves for (a) the Bai et al.⁸ model (b) the Covinet model⁹, (c) the Covnet model⁷, (d) the proposed mixed-effects model on the Mosmed internal validation dataset.



1

2 **Figure 6:** Decision curves for the proposed mixed-effects model on the Mosmed internal validation
3 dataset.



4

5

(a)

(b)

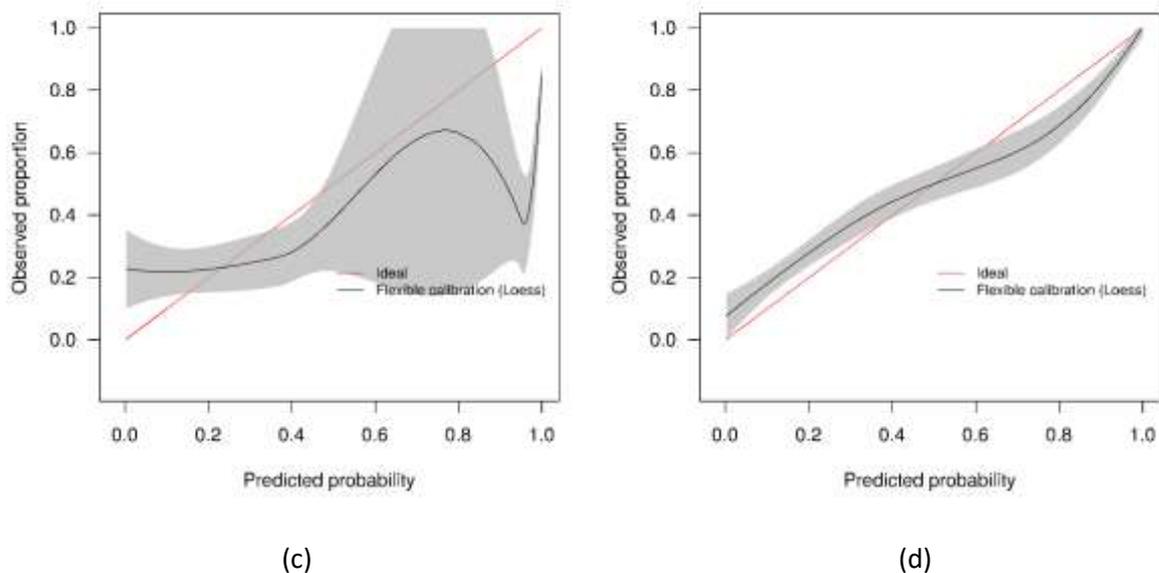


Figure 7: Calibration curves for (a) the Bai et al.⁸ model (b) the Covnet model⁹, (c) the Covnet model⁷, (d) the proposed mixed-effects model on the Zhang et al. external validation dataset.

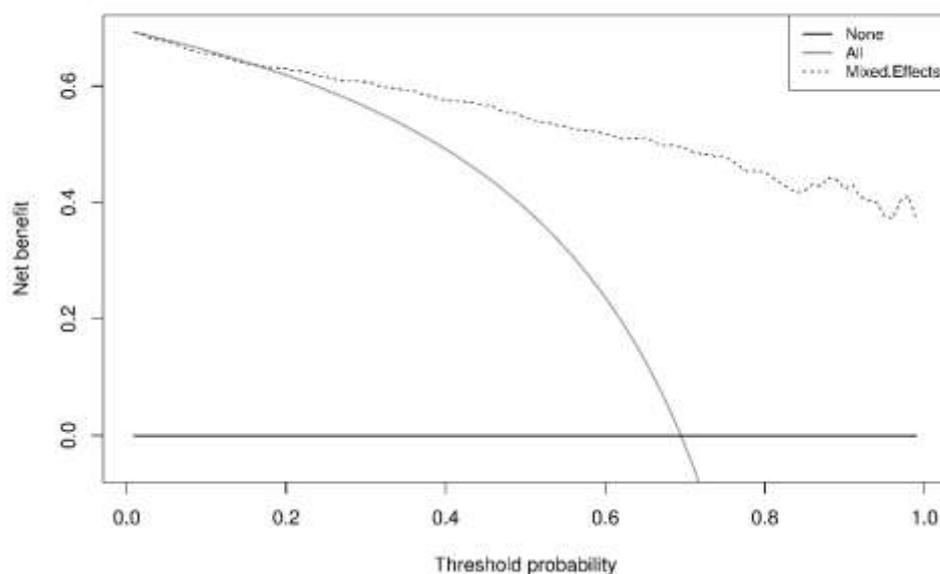
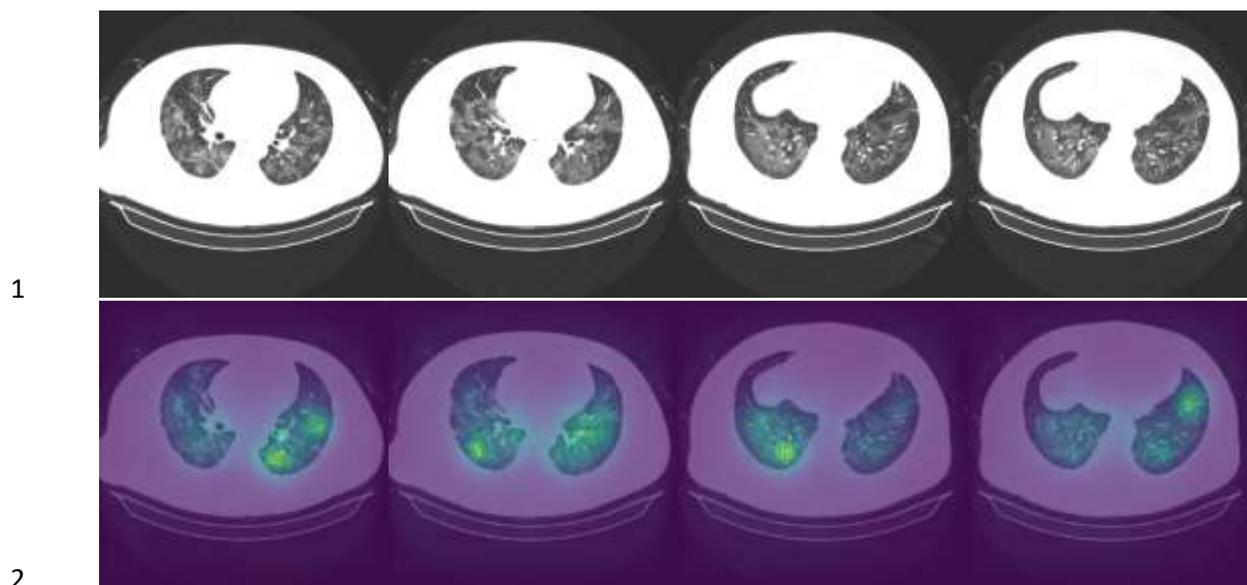


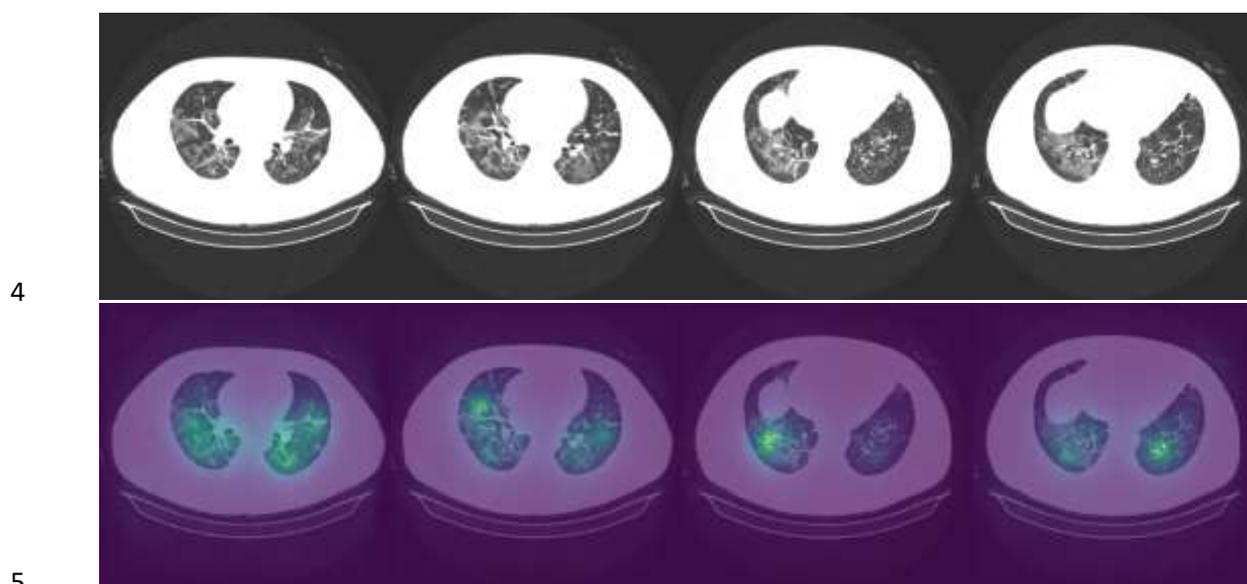
Figure 8: Decision curves for the proposed mixed-effects model on the Zhang et al. external validation dataset.

Saliency maps

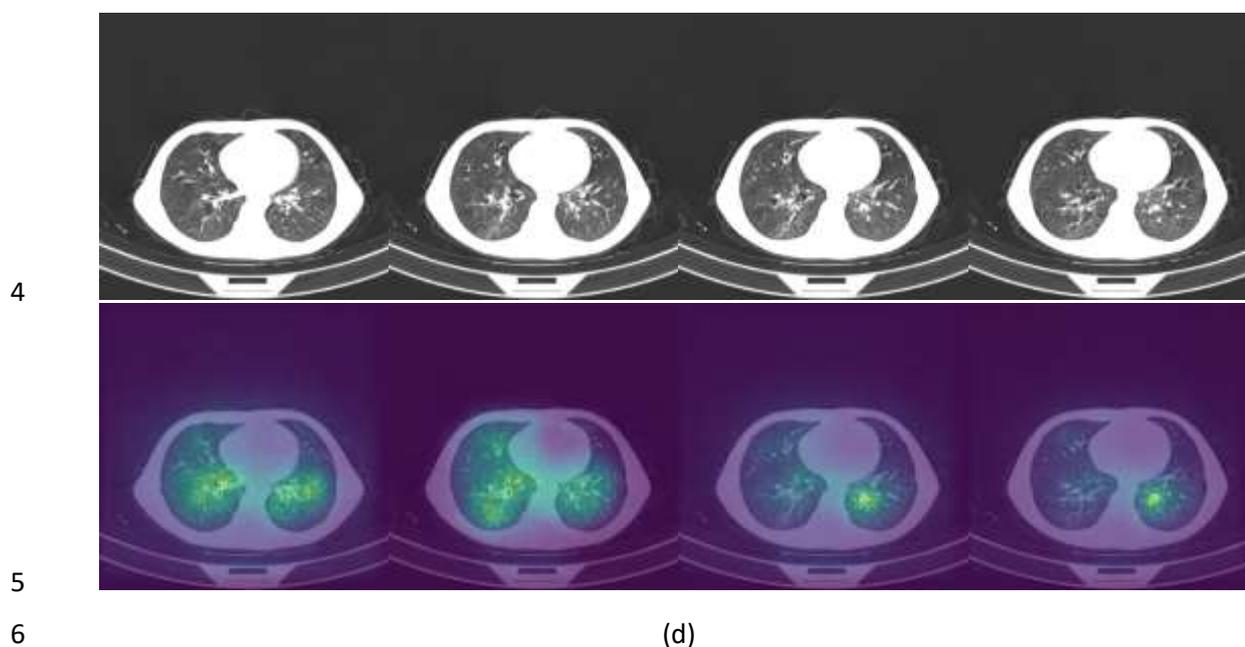
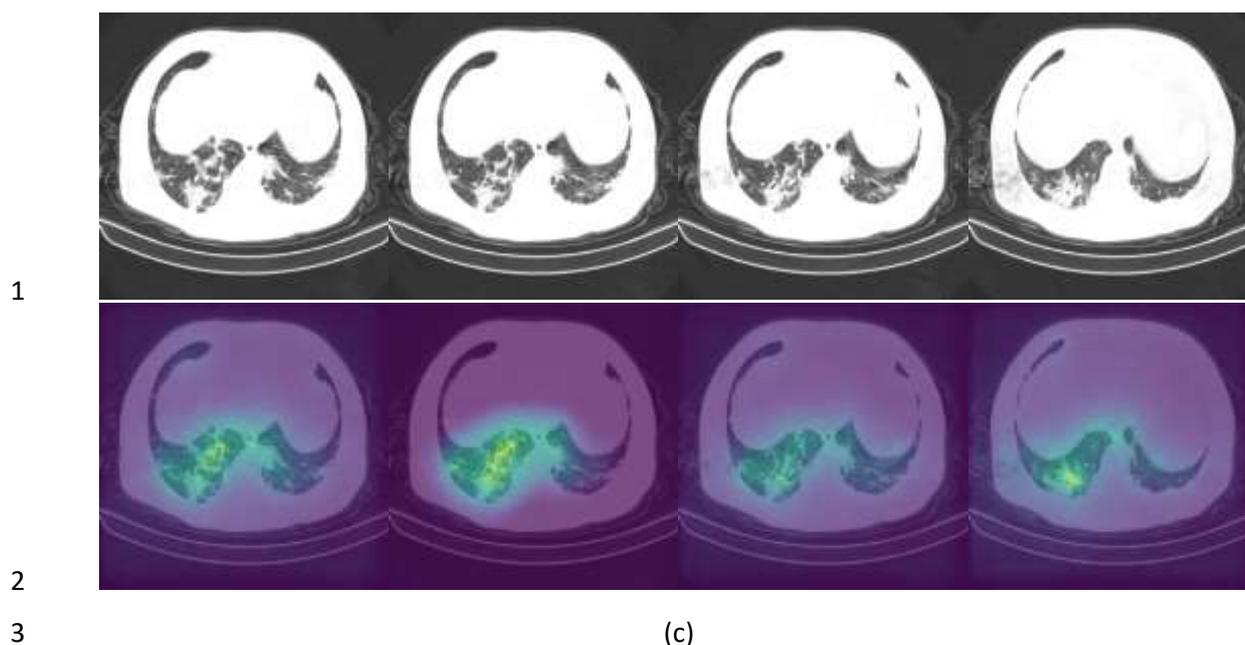
It is vital to understand how the algorithm makes decisions and to check that it identifies the correct features within the image. Saliency maps can be used as a visual check to see what features the algorithm is learning. For example, the saliency maps in Figure 9 show that the model correctly identifies the diseased areas of the scans. We used 100 samples with a smoothing noise of 0.05 to create these saliency maps.



(a)



(b)

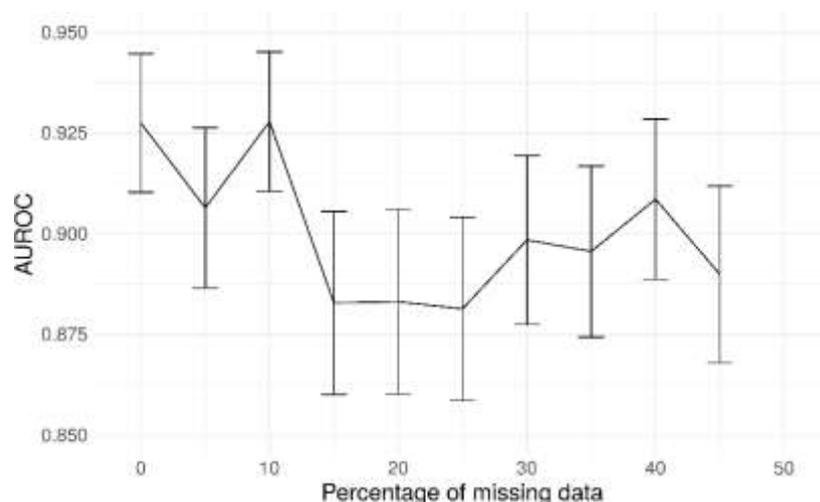


7 **Figure 9:** Example of original images and saliency maps showing highlighted regions on four patients
8 in the Zhang et al.⁴⁵ dataset. Four consecutive images (a, b, c and d) display how the diseased areas
9 differ between slices. All images are taken from the external validation set.

10 Sensitivity analysis

11 Mixed-effects models are capable of accounting for missing data. However, only one image had less
12 than 20 slices; hence, we could not adequately assess if our model can indeed maintain good
13 performance with missing data. Here, we rerun the analysis using the same dataset, using the same
14 model and weights; however, we reduce the number of slices available as testing data inputs to
15 simulate missing data. Blank images replace these slices. We uniformly sampled the slices choosing
16 between 10 and 19 slices; this equates to between 5 and 50% missing data for the model. We ran
17 inference at each level of missingness and briefly show the AUROC to determine at which point the
18 predictive performance is significantly reduced.

1 The plot of AUROCs at different levels of missingness is shown in Figure 10, along with 95%
2 confidence intervals. We can see that at 20% missingness, there is a statistically significant decrease
3 in predictive performance. Although, even at 50% missingness, the model still performs relatively
4 well, with an AUROC of 0.890 (95% CI: 0.868, 0.912). It should be noted that this does not mean that
5 there is no reduction in performance at 5-15% missingness, only that the reduction was not
6 statistically significant at the 95% confidence level.



7
8 **Figure 10:** AUROC values at different levels of missingness. At 20% missingness, the loss in
9 performance becomes statistically significant; however, even with 50% missing images, the model
10 still has a reasonably high AUROC.

11 Deep learning models can be susceptible to adversarial attacks⁴⁶, where minor artefacts or noise on
12 an image can cause the image to be misclassified, even when the image does not look significantly
13 different to a human observer. Here, we perform a brief sensitivity analysis by adding a small
14 Gaussian noise to the image. We tested the model performance on the external dataset, with each
15 image having a random Gaussian noise added. Experiments were conducted with standard
16 deviations of 0 up to 0.005 in increments of 0.001 added to the normalised image. We did not add
17 Gaussian noise in the data augmentation so that the model is not explicitly trained to deal with this
18 kind of attack.

19 When using a variance of 0, the images are unchanged, and the results are the same as the standard
20 results above. We present results on the Zhang et al.⁴⁵ dataset. Example images for each level of
21 variance are shown in Figure 11, and a graph showing the reduction in AUROC is shown in Figure 12.

22



23

24

(a)

(b)

(c)

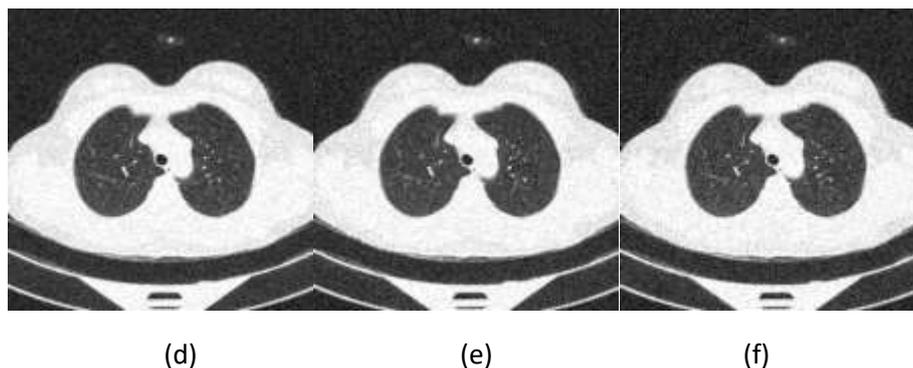


Figure 11: Example images showing the effect of increasing the amount of noise in the image input. (a) deviation=0; (b) deviation=0.001; (c) deviation=0.002; (d) deviation=0.003; (e) deviation=0.004; (f) deviation=0.005.

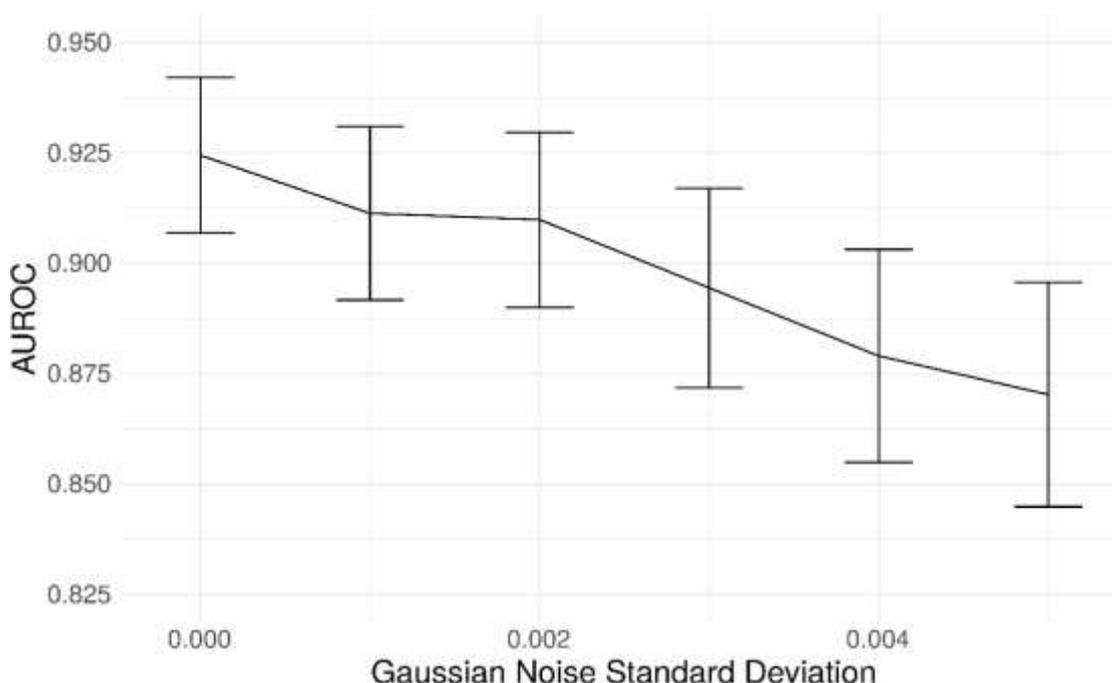


Figure 12: Graph showing the drop in AUROC as the amount of noise in the image input increases. The AUROC falls steadily with increased noise in the image.

10 Discussion

11 Artificial intelligence is set to revolutionise healthcare, allowing large amounts of data to be
12 processed and analysed automatically, reducing pressure on stretched healthcare services. These
13 tools can aid clinicians in monitoring and managing both common conditions and outbreaks of novel
14 diseases. However, these tools must be assessed adequately, and best practice guidelines for
15 reporting and development must be followed closely to increase reproducibility and reduce bias. We
16 have developed a deep learning model to classify CT scans as healthy or COVID-19 using a novel
17 mixed-effects model. Following best practice guidelines, we have externally validated the model. In
18 addition, we robustly externally geographically validated the developed model in several
19 performance areas, which are not routinely reported. For example, discriminative performance
20 measures show that the model can discriminate between healthy and COVID-19 CT scans well,

1 calibration shows that the model is not clinically harmful. Finally, the clinical usefulness measures
2 show that the model may be useful in a clinical setting. From the results presented here, it would
3 seem that our deep learning model outperforms the RT-PCR tests as shown in the review by Watson
4 et al.³; however, those results are conservative estimates and were conducted under real-world
5 clinical settings. A prospective study is required to determine if this is the case.

6 Compared to previously proposed models, our model showed similar discriminative performance to
7 one existing method; however, our method generalised better to an external geographical validation
8 set and showed improved calibration performance. Interestingly, in both internal and external
9 validation, the sensitivity and NPV are similar in all models. However, the specificity and PPV are
10 statistically significantly improved for the mixed-effects model in the external validations dataset.
11 The performance of the proposed model in the external validation set is similar to that reported by
12 PCR testing³. However, a direct comparison should not be made as PCR testing on this exact dataset
13 is unavailable.

14 There are several limitations of the study that should be highlighted and improved in future work.
15 Firstly, we have only performed external geographical validation in a single dataset. Further external
16 validation, both geographical and temporal, is needed on many datasets to determine if the model is
17 correct in each intended setting. Although we performed a brief sensitivity analysis here, more
18 extensive work on adversarial attacks is needed. Future studies could consider following the method
19 proposed by Goodfellow et al.⁴⁶ to improve robustness against adversarial examples. Patient
20 demographic data were not available for this study, but future studies could incorporate this data
21 into the model to improve results. Finally, rules of thumb for assessing sample size calculations in
22 the validation set can lead to imprecise results⁴⁷. Simulating data is a better alternative; however, it
23 is difficult to anticipate the distribution of the model's linear predictor. Therefore, we were required
24 to revert to the rule of thumb using a minimum of 200 samples in each group²⁸.

25 Initial experiments used the Zhang et al.⁴⁵ dataset for training; this showed promising results on the
26 internal validation set; however, external validation showed random results. In addition, saliency
27 maps showed that the model was not using the features of COVID-19 to make the diagnosis and was
28 instead using the area around the image. We concluded that the images for each class were slightly
29 different, perhaps due to different imaging protocols, and the algorithm was learning the image
30 format rather than the disease. We then used the MosMed dataset for training and the Zhang et
31 al.⁴⁵ dataset for external validation. This highlights the need for good quality training data and
32 external validation and visualisation.

33 Future studies should validate models and follow reporting guidelines such as TRIPOD¹⁷ or the
34 upcoming QUADAD-AI⁴⁸ and TRIPOD-AI⁴⁹ to bring about clinically useful and deployable models.
35 Further research could look deeper into the areas of images identified by the algorithm as shown on
36 the saliency maps; this could potentially identify new features of COVID-19 which have gone
37 unnoticed. Before any model can be fully deployed, clinical trials are needed to study the full impact
38 of using such algorithms to diagnose COVID-19 and the exact situations in which such a model may
39 be used. In-clinic prospective studies comparing the performance deep learning models with RT-PCR
40 and lateral flow tests should be carried out to determine how deep learning compares; this will show
41 whether deep learning could be used as an automated alternative to RT-PCR testing.

42 This study indicates that deep learning could be suitable for screening and monitoring of COVID-19
43 in a clinical setting; however, validation in the intended setting is vital, and models should not be
44 adopted without this. It has been observed that the quality of reporting of deep learning prediction
45 models is usually very poor; however, with a bit of extra work and by following best practice

1 guidelines, this problem can be overcome. This study highlights the importance of robust analysis
2 and reporting of models with external validation.

3

4 References

- 5 1. Coronavirus disease 2019 (COVID-19) - Symptoms, diagnosis and treatment | BMJ Best Practice:
6 BMJ Publishing Group; 2020 [Available from: [https://bestpractice.bmj.com/topics/en-](https://bestpractice.bmj.com/topics/en-gb/3000201)
7 [gb/3000201](https://bestpractice.bmj.com/topics/en-gb/3000201).
- 8 2. Torjesen I. Covid-19 will become endemic but with decreased potency over time, scientists
9 believe. *BMJ* 2021;372:n494. doi: 10.1136/bmj.n494
- 10 3. Watson J, Whiting PF, Brush JE. Interpreting a covid-19 test result. *BMJ* 2020;369:m1808. doi:
11 10.1136/bmj.m1808
- 12 4. Mallett S, Allen AJ, Graziadio S, et al. At what times during infection is SARS-CoV-2 detectable and
13 no longer detectable using RT-PCR-based tests? A systematic review of individual participant
14 data. *BMC Medicine* 2020;18(1):346. doi: 10.1186/s12916-020-01810-8
- 15 5. Ruan Z-R, Gong P, Han W, et al. A case of coronavirus disease 2019 with twice negative nucleic
16 acid testing within 8 days. *Chinese Medical Journal* 2020;133(12)
- 17 6. Pontone G, Scafuri S, Mancini ME, et al. Role of computed tomography in COVID-19. *J Cardiovasc*
18 *Comput Tomogr* 2020;S1934-5925(20)30436-6. doi: 10.1016/j.jcct.2020.08.013
- 19 7. Li L, Qin L, Xu Z, et al. Using Artificial Intelligence to Detect COVID-19 and Community-acquired
20 Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. *Radiology*
21 2020;296(2):E65-E71. doi: 10.1148/radiol.2020200905
- 22 8. Bai HX, Wang R, Xiong Z, et al. Artificial Intelligence Augmentation of Radiologist Performance in
23 Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT. *Radiology*
24 2020;296(3):E156-E65. doi: 10.1148/radiol.2020201491
- 25 9. Mittal B, Oh J. CoviNet: Covid-19 diagnosis using machine learning analyses for computerized
26 tomography images. *Thirteenth International Conference on Digital Image Processing (ICDIP*
27 *2021)* 2021;11878:1187816.
- 28 10. MacCormick IJC, Zheng Y, Czanner S, et al. Spatial statistical modelling of capillary non-perfusion
29 in the retina. *Scientific Reports* 2017;7(1):16792. doi: 10.1038/s41598-017-16620-x
- 30 11. Zhu W, Ku JY, Zheng Y, et al. Spatial linear mixed effects modelling for OCT images: SLME model.
31 *Journal of Imaging* 2020;6(6):44.
- 32 12. Albtoush OM, Al-Shdefat RB, Al-Akaileh A. Chest CT scan features from 302 patients with COVID-
33 19 in Jordan. *European Journal of Radiology Open* 2020;7:100295. doi:
34 <https://doi.org/10.1016/j.ejro.2020.100295>
- 35 13. Haseli S, Khalili N, Bakhshayeshkaram M, et al. Lobar distribution of COVID-19 pneumonia based
36 on chest computed tomography findings; A retrospective study. *Arch Acad Emerg Med*
37 2020;8(1):e55-e55.
- 38 14. Xiang C, Lu J, Zhou J, et al. CT findings in a novel coronavirus disease (COVID-19) pneumonia at
39 initial presentation. *BioMed Research International* 2020;2020:5436025. doi:
40 10.1155/2020/5436025
- 41 15. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of
42 covid-19: systematic review and critical appraisal. *BMJ* 2020;369:m1328. doi:
43 10.1136/bmj.m1328
- 44 16. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine
45 learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans.
46 *Nature Machine Intelligence* 2021;3(3):199-217. doi: 10.1038/s42256-021-00307-0
- 47 17. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction
48 model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ : British*
49 *Medical Journal* 2015;350:g7594. doi: 10.1136/bmj.g7594

- 1 18. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): A guide
2 for authors and reviewers. *Radiology: Artificial Intelligence* 2020;2(2):e200029. doi:
3 10.1148/ryai.2020200029
- 4 19. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability
5 of prediction model studies. *Annals of Internal Medicine* 2019;170(1):51-58. doi:
6 10.7326/M18-1376
- 7 20. Jiang J, Nguyen T. Linear and generalized linear mixed models and their applications: Springer
8 2007.
- 9 21. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision.
10 *Proceedings of the IEEE conference on computer vision and pattern recognition* 2016:2818-
11 26.
- 12 22. Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. *2009 IEEE*
13 *Conference on Computer Vision and Pattern Recognition* 2009:248-55. doi:
14 10.1109/CVPR.2009.5206848
- 15 23. Xiong Y, Kim HJ, Singh V. Mixed effects neural networks (MeNets) with applications to gaze
16 estimation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
17 2019:7735-44. doi: 10.1109/CVPR.2019.00793
- 18 24. Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly*
19 *Weather Review* 1950;78(1):1-3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
- 20 25. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized
21 likelihood methods. *Advances in large margin classifiers* 1999;10(3):61-74.
- 22 26. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: A
23 framework for traditional and novel measures. *Epidemiology* 2010;21(1)
- 24 27. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive
25 analytics. *BMC Medicine* 2019;17(1):230. doi: 10.1186/s12916-019-1466-7
- 26 28. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk models was defined:
27 from utopia to empirical data. *Journal of Clinical Epidemiology* 2016;74:167-76. doi:
28 <https://doi.org/10.1016/j.jclinepi.2015.12.005>
- 29 29. R: A Language and Environment for Statistical Computing [program], 2021.
- 30 30. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated
31 Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*
32 1988;44(3):837-45. doi: 10.2307/2531595
- 33 31. Du Z, Hao Y. reportROC: an easy way to report ROC analysis. R package version 3.5, 2020.
- 34 32. rms: Regression Modeling Strategies [program], 2021.
- 35 33. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models.
36 *Med Decis Making* 2006;26(6):565-74. doi: 10.1177/0272989X06295361
- 37 34. Smilkov D, Thorat N, Kim B, et al. Smoothgrad: Removing noise by adding noise. arXiv 2017. *arXiv*
38 *preprint arXiv:170603825*
- 39 35. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. *Proceedings of the IEEE*
40 *conference on computer vision and pattern recognition* 2016:770-78.
- 41 36. Efficientnet: Rethinking model scaling for convolutional neural networks. International
42 Conference on Machine Learning; 2019. PMLR.
- 43 37. Mittal B, Oh J. CoviNet: Covid-19 diagnosis using machine learning analyses for computerized
44 tomography images. *ProcSPIE* 2021;11878 doi: 10.1117/12.2601065
- 45 38. Abadi M, Agarwal A, Barham P, et al. Tensorflow: Large-scale machine learning on
46 heterogeneous distributed systems. *arXiv preprint arXiv:160304467* 2016
- 47 39. Tensorflow: A system for large-scale machine learning. 12th {USENIX} symposium on operating
48 systems design and implementation ({OSDI} 16); 2016.
- 49 40. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and
50 compare ROC curves. *BMC Bioinformatics* 2011;12(1):77. doi: 10.1186/1471-2105-12-77
- 51 41. Wickham H. ggplot2: Elegant Graphics for Data Analysis.: Springer-Verlag New York 2016.

- 1 42. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:14126980*
2 2014
- 3 43. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction
4 model: PART II - binary and time-to-event outcomes. *Statistics in Medicine* 2019;38(7):1276-
5 96. doi: <https://doi.org/10.1002/sim.7992>
- 6 44. Morozov SP, Andreychenko AE, Blokhin IA, et al. MosMedData: Data set of 1110 chest CT scans
7 performed during the COVID-19 epidemic. *DD* 2020;1(1):49-59. doi: 10.17816/dd46826
8 [published Online First: 2020-12-30]
- 9 45. Zhang K, Liu X, Shen J, et al. Clinically applicable AI system for accurate diagnosis, quantitative
10 measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell*
11 2020;181(6):1423-33.e11. doi: 10.1016/j.cell.2020.04.045
- 12 46. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. *arXiv preprint*
13 *arXiv:14126572* 2014
- 14 47. Snell KIE, Archer L, Ensor J, et al. External validation of clinical prediction models: simulation-
15 based sample size calculations were more reliable than rules-of-thumb. *Journal of Clinical*
16 *Epidemiology* 2021;135:79-89. doi: 10.1016/j.jclinepi.2021.02.011
- 17 48. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-
18 centered diagnostic test accuracy studies: QUADAS-AI. *Nature Medicine* 2021;27(10):1663-
19 65. doi: 10.1038/s41591-021-01517-0
- 20 49. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline
21 (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction
22 model studies based on artificial intelligence. *BMJ Open* 2021;11(7):e048008. doi:
23 10.1136/bmjopen-2020-048008

24