This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Explainable and Uncertainty-Aware Chest X-ray Classification with Selective Referral

Moinul Hossain[1] and Md Kishor Morol[2]

[1] American International University-Bangladesh (AIUB), Dhaka 1229, Bangladesh
mhshihab2001@gmail.com
[2] ELITE Research Lab, New York, USA
kishormorol@ieee.org

Abstract. The essential criteria of efficient CXR triaging system is reliable spatial cues and proper uncertainty calibration. We performed retrospective selective-referral pipeline testing on the 30,194 4-class COVID-QU-Ex CXR image database. The selective-referral pipeline utilized DenseNet-121, Grad-CAM++, and MC Dropout. We examined the entire pipeline using only one test set, unlike other works using different approaches, based on classification, localization of spatial cues, estimation of uncertainty, selective-referral strategy and global results. The system showed 97.81% of classification accuracy on the 4-classification task using 4,530 test images, with the macro F1 of 97.62% and macro-AUC of 0.9979. Concerning localizations for the 502 COVID-19 CXRs, Grad-CAM++ achieved the Pointing Game score of 82.05% with IoU equal to 0.63.MC Dropout produced an ECE value of 0.018, and an Error-AUROC of 0.941. The proposed process was able to refer 90 out of 99 wrongly classified cases (90.9%), based on referral criteria generated from validation. The system managed to retain 96.09% of the overall data with a classification accuracy rate of 99.79%. The above tests performed have been done retrospectively. They haven't been clinically tested or human-based trials have not been done yet.

Keywords: Explainable AI · Uncertainty Quantification · Chest X-ray · Selective Referral · Grad-CAM++ · Monte Carlo Dropout · Calibration

1 Introduction

Deep learning models have been able to achieve remarkable results on benchmark datasets of CXR [1,2]. Yet, mere classification results do not suffice to warrant the clinicians' trust since false classifications made due to lack of robust uncertainty measures constitute a type of silent errors that may harm subsequent decision-making processes [3]. In addition, spatial evidence can indicate the focus that the model is using, whereas predictive uncertainty will be able to pinpoint those cases where the model does not know what it is doing.

A selection of reader studies as well as implementation within the clinic, and also retrospective simulations have examined the use of AI in the triaging of CXR images [16,17]. Sadly, the areas of probability calibration, the examination of an

explanation, inter-institutional generalizability and efficient selectivism referral systems have yet to be sufficiently researched [15]. Grad-CAM++ is an example of an algorithm which allows for the creation of saliency maps by class [5], while MC Dropout offers a way to estimate the level of uncertainty in the prediction without re-training the model [6].

The selection of a prediction makes use of the trade-off between the levels of prediction accuracy and coverage [8]

In our work, these techniques have been considered collectively within one pipeline, on a single fixed test dataset, which is not the case in any of the prior works.

Problem Statement. Does a joint CXR pipeline which consists of DenseNet-121 classification, Grad-CAM++ explanation and MC Dropout uncertainty estimation yield a comparable or superior error-aware selective referral performance than deterministic confidence based techniques on the COVID-QU-Ex dataset, while also ensuring robust classification and localization?

Contributions.

(1) An offline selective referral pipeline based on DenseNet-121 classification, Grad-CAM++ visualization, and MC dropout uncertainty estimation was created.

(2) The testing of classification, calibration, uncertainty, selective referral, and localization was conducted for one fixed test dataset.

(3) The referrals' threshold values were obtained from validation data only and then used to compute coverage, error detection, false referral rate, retained accuracy and silent error rate metrics.

(4) We made a direct comparison between different confidence estimation techniques (entropy, max probability of softmax, deterministic entropy, temperature scaling of entropy) with equaling all other parameters to focus solely on the contribution of confidence estimation.

2 Related Work

Classification and Workflow for Chest X-rays.

State of the art approaches for pneumonia classification or multi-label classification of CXRs include CheXNet and CheXpert [1,2], while COVID-Net focuses on COVID-19 detection [23]. In addition to classification labels, the COVID-QU-Ex dataset also provides segmentation masks of lesions [24]. Classification performance has also been shown to be improved with CLAHE on COVID-19 CXRs [27].

While Sridharan et al. employed readers to experiment on an AI-assisted triage workflow, Yoon et al. analyzed retrospective workflow reduction [16,17]. We differentiate our work by only considering the simulation of a retrospective workflow without taking into account the radiologist's opinion, the time to read, or the decision process.

Uncertainty and Selective Prediction.

MC Dropout is a common way to approximate Bayesian Model Averaging for estimating predictive uncertainty [6] and has also been used in classifying

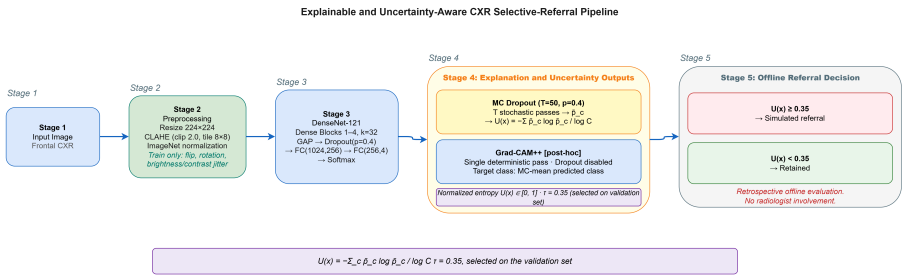

Fig. 1: Offline five-stage pipeline: preprocessing, DenseNet-121 prediction, MC Dropout uncertainty, post-hoc Grad-CAM++, and validation-selected simulated referral.

COVID-19 CXRs [9]. Since deep learning models tend to be overconfident, it is standard practice to apply temperature scaling for probability calibration [7]. Selective prediction allows for error reduction by choosing not to perform a classification when the sample is too uncertain [8]. Here, a few confidence scores have been calculated using the fixed trained model in order to see only the contribution of uncertainty estimation to the selective prediction.

Evaluation of Explainable AI.

Post-hoc explanations like Grad-CAM, Grad-CAM++, Score-CAM and Eigen-CAM have been developed [4,5,12,13]. Pointing Game, Intersection over Union (IoU), insertion and deletion are commonly used measures to define explanation quality [10,11]. Significant difficulties with the evaluation of explainable AI, specifically regarding generalization and clinical utility, were documented in a recent systematic review [15] which prompted this work.

3 Methodology

3.1 Offline Selective-Referral Workflow

Our model's performance is evaluated retrospectively on the out-of-sample test set; since we do not have radiologist evaluators involved, referral and retention rates are simulated deployment results. After CLAHE and normalization, we pass the chest X-rays to the classifier DenseNet-121; during inference we perform $T = 50$ stochastic forward passes with dropout turned on to compute the mean predictive distribution $\bar{\mathbf{p}}$ and normalize its entropy $\mathcal{U}(\mathbf{x})$ according to Eqn. 3. We compute Grad-CAM++ by using the mean class' prediction, and the case is referred when $\mathcal{U}(\mathbf{x}) \geq \tau$. When $\mathcal{U}(\mathbf{x}) < \tau$, the case is kept and automatically classified. An interface that shows predictions, saliency maps, and uncertainty estimates can be a valuable tool to help radiologists perform this classification task better, but that is outside the scope of the current experiment.

Table 1: Stratified dataset split; totals sum to 30,194 images.

Split	COVID	Normal	Opacity	Pneumonia	Total
Train	8,369	7,491	4,208	1,068	21,136
Validation	1,793	1,605	902	228	4,528
Test	1,794	1,605	902	229	4,530

3.2 Dataset and Preprocessing

The data used for our research consisted of 30,194 front CXRs of COVID-QU-Ex Dataset [24] which consisted of 4 classes i.e., 11,956 COVID-19, 10,701 Normal, 6,012 Lung Opacity, and 1,525 Viral pneumonia, respectively. The COVID-QU-Ex Dataset consists of segmentation of lung mask for each of the CXR, the segmentation of lung-infected region for a fraction of 2,913 cases that have diagnosis of COVID-19. We have found out that after the interception of list of infected masks and the set list of static classtest split, we get 502 CXRs diagnosed with COVID-19 to use for lesion location.

Masks of lung-section of lung opacity has used for a secondary evaluation for localization at lung sections.

However, these masks cannot be considered as the lesions Annotation. Patient independence was not checked by us as notoughinformationwasfoundinthemeta-dataatourdispose.

The images are downscaled to 224×224. The CLAHE operation is performed with the clip limit set to 2.0 and the 8×8 grid size along with ImageNet normalization. The augmentation used during the training process is horizontal flipping ($p = 0.5$), rotation ($\pm 15°$), and brightness and contrast adjustment (± 0.2). Cross-entropy with class weights inversely proportional to the number of samples in each class deals with the problem of class imbalance in the Viral Pneumonia dataset (training sample size = 1,068 images, 5.1%).

3.3 Model, Training, and Evaluation

Localised classification system - replace the classifier (which used DenseNet121, trained on ImageNet): GlobalAveragePooling $\rightarrow$ Dropout(0.4) $\rightarrow$ FC$(1024, 256)$ $\rightarrow$ FC$(256, 4)$. Train model using AdamW [29] with the learning rate set to 10^{-4}, weight decay 5×10^{-4}, for 60 epochs with cosine annealing scheduler on batches of 32 with 10 patients of early stop. Experiment performed in PyTorch 2.1 with random seed 42 and one GPU; test-set uncertainties are presented by bootstrap intervals rather than running the experiment multiple times.

All the base architectures used the same training/test sets, normalization, loss function, optimizer and a budget to early stop on patient number.

The entire workflow with CLAHE, class weighting, MC averaging, and data augmentations was used. Architectural comparisons and sequential ablation tests question separately.

Evaluation is done based on classification metrics: accuracy, precision (macro), recall (macro), F1 (macro), and AUC (one-vs-rest). Calibration metrics are 15-bin

ECE based on the most confident prediction, multiclass Brier score

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}(p_{nc}-y_{nc})^2$$

together with NLL. For AUROC/AUPRC error on false identified image, misclassified test/numtest.

Selectivity performance metrics are coverage, false referrals/numtest, acccovered, recall of errors on false positives, and lowconfidenceerror/numtest. For XAI models the metrics are Pointing Game, IoU with threshold of 0.5 using prespecified saliency-threshold bin, and insertion/deletion AUC over 100 steps. Pointing Game is performed ignoring the IoU constraint. Confidence intervals are computed using 1k pairwise bootstrap resampling.

3.4 Grad-CAM++ and MC Dropout

Following Chattopadhay et al. [5], the Grad-CAM++ channel weight and final saliency map are

$$w_k^c=\sum_{i,j}\alpha_{ij}^{kc}\,\mathrm{ReLU}\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right),\qquad L_{++}^c=\mathrm{ReLU}\left(\sum_k w_k^c A^k\right),\tag{1}$$

where Y^c is the pre-softmax class score and A^k is a final-layer feature map. The map is bilinearly upsampled to input resolution. Grad-CAM++ is post-hoc and does not change training or logits. MC Dropout was selected because the classifier already contains a dropout layer, it yields a predictive distribution without retraining, and it can be applied to pretrained weights. Classifier-only dropout may underestimate epistemic uncertainty, and repeated passes increase latency. For $C=4$ classes,

$$\bar{p}_c=\frac{1}{T}\sum_{t=1}^{T}p_c^{(t)},\qquad \mathcal{U}(\mathbf{x})=-\frac{\sum_{c=1}^{C}\bar{p}_c\log\bar{p}_c}{\log C}.\tag{2}$$

For each $T\in\{5,10,20,30,50\}$, the validation threshold attaining at least 90% error recall was identified, and the corresponding coverage was compared. $T=50$ achieved the highest validation coverage under this constraint. The resulting $\tau=0.35$ was fixed before test evaluation.

4 Results

4.1 Classification and Ablation

Table 2 shows comparison between the basic-protocol architecture and the full enhanced pipeline. The full enhanced pipeline had the best accuracy (97.81%,

Table 2: Test-set classification. Accuracy is overall; precision, recall, F1, and one-vs-rest AUC are macro-averaged.

Method	Acc.%	Prec.%	Rec.%	F1%	AUC
VGG-16	91.24	90.87	91.04	90.95	.971
ResNet-50	93.67	93.41	93.55	93.48	.981
InceptionV3	94.12	93.88	93.96	93.92	.984
EfficientNet-B3	95.78	95.54	95.69	95.61	.991
COVID-Net	96.31	96.07	96.22	96.14	.993
DenseNet-121 basic	96.54	96.28	96.43	96.35	.994
Full pipeline	**97.81**	**97.64**	**97.61**	**97.62**	**.9979**

Table 3: Sequential ablation; each row adds to the preceding row.

Configuration	Acc.%	Macro-F1%
ResNet-50 baseline	93.67	93.48
+ DenseNet-121 backbone	96.54	96.35
+ augmentation and CLAHE	96.98	96.74
+ MC Dropout averaging ($T = 50$)	97.21	96.97
+ class-weighted loss (full)	**97.81**	**97.62**

4,431/4,530), macro F1-score (97.62%), and macro AUC (0.9979). Since preprocessing, loss and inference are also different, the current table cannot be considered a comparison of architectures only. The sequential ablation in Table 3 is order-dependent: DenseNet-121 gives the largest gain, followed by augmentation/CLAHE, MC averaging, and class weighting. Grad-CAM++ is omitted because it is post-hoc. The final checkpoint was selected using validation loss, and the locked test set was evaluated after model selection. Per-class recall was 98.8% for COVID-19 (1,772/1,794), 97.9% for Normal (1,571/1,605), 96.5% for Lung Opacity (870/902), and 95.2% for Viral Pneumonia (218/229). The diagonal sum is 4,431, consistent with 97.81% overall accuracy. Macro-AUC was 0.9979 and every class exceeded 0.996 (Fig. 2).

4.2 Saliency Localization and Faithfulness

After the intersection of the infection-mask manifest with the classification test split, there were 502 COVID-19 test images with infection masks. The COVID-19 PG and IoU results in Table 4 were all calculated from this subset. The PG and IoU for Grad-CAM++ were 82.05% and 0.63, respectively, outperforming all other approaches we evaluated. The Lung Opacity lung-mask IoU result was 0.57 and was considered solely as saliency containment within the lung. AUC results for insertion and deletion are calculated across all classes.

4.3 Calibration, Error Detection, and Selective Referral

All rows in Table 5 use the same full-pipeline trained weights. Deterministic rows disable dropout; MC rows activate dropout and average stochastic predictions.

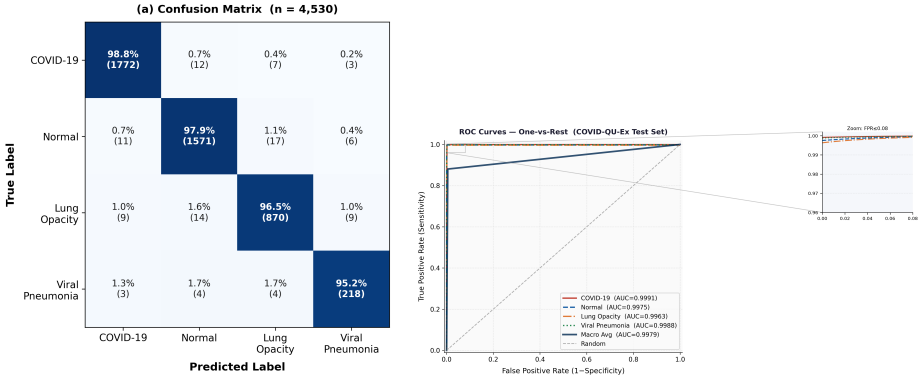

Fig. 2: Normalized confusion matrix and one-vs-rest ROC curves on the 4,530-image test set.

Table 4: XAI comparison. PG/IoU use the 502-image COVID-19 infection-mask subset; insertion/deletion AUC use all classes. IoU uses threshold 0.5; Pointing Game does not.

Method	PG↑	IoU↑	Del.AUC↓	Ins.AUC↑
Grad-CAM	73.10	.48	.312	.741
Grad-CAM++	**82.05**	**.63**	**.274**	**.803**
Score-CAM	80.14	.59	.281	.791
EigenCAM	76.38	.52	.298	.762

Temperature scaling is fitted on the validation set. MC Dropout $T = 50$ achieved the lowest ECE, Brier, and NLL and the highest Error-AUROC. Difference CIs from paired bootstrap resampling excluded zero: Error-AUROC improved over deterministic entropy by 0.138 [0.124, 0.152] and over $1 - \max p$ by 0.120 [0.107, 0.133]; ECE improved over raw softmax by 0.024 [0.014, 0.034]. At $\tau = 0.35$, $T = 50$ referred 177/4,530 images: 90 errors and 87 correct cases. It therefore captured 90.9% of 99 errors, retained 4,353 cases at 99.79% accuracy, and left 9 low-uncertainty errors. Table 6 contrasts this point with $T = 10$. A complete threshold-sensitivity analysis and risk–coverage curve were not retained in the original experiment pipeline; accordingly, we restrict quantitative reporting to the validation-selected operating point and state this limitation explicitly.

5 Discussion

Findings and Interpretation. Grad-CAM++ provided the best localization of the evaluated saliency methods. COVID-19 infection-mask results show spatial correspondence with the annotations, and Lung Opacity IoU only tests for saliency containment within the lungs. Results of the insertion/deletion experiments provide perturbation-based evidence of faithfulness but not causation or radiological validity, therefore, further investigation with clinician readers is needed. **Uncertainty and Referral.** MC Dropout outperformed deterministic

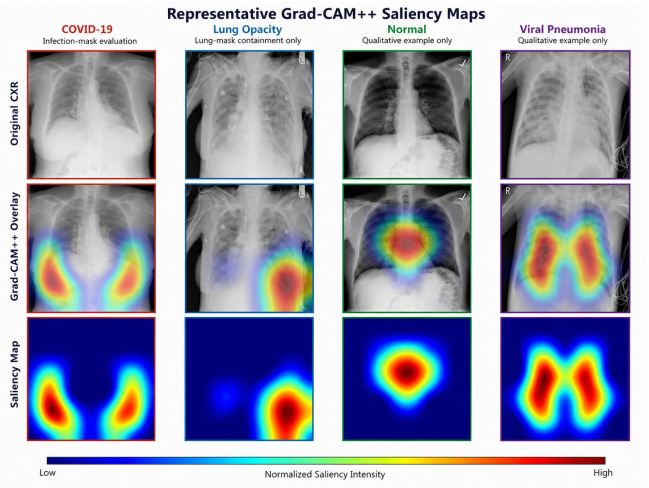

Fig. 3: Representative Grad-CAM++ maps. COVID-19 lesion metrics use infection-region masks; Lung Opacity lung masks are not lesion annotations.

Table 5: Calibration and error detection using identical trained weights; 95% bootstrap CIs are shown for the main metrics.

Inference/score	ECE	Brier	NLL	Err-AUROC	Err-AUPRC
Det. $1 - \max p$	.042	.071	.198	.821	–
Det. entropy	.042	.071	.198	.803	–
Temp.-scaled $1 - \max p$	.021	.052	.151	.791	–
MC Dropout $T = 10$	.024	.048	.141	.891	.742
MC Dropout $T = 50$	**.018**	**.041**	**.124**	**.941**	**.813**

confidence scores for calibration and error detection, without retraining. At the validation-selected operating point, 3.91% of the images were referred, while the retained image subset had 99.79% accuracy and 0.20% silent failure rate. This proves error concentration in this dataset retrospectively, and does not mean safety or any practical value for the workflow. **Scope and limitations of clinical practice.** These four categories are image-classification labels, not diagnoses; clinical diagnosis of COVID-19 requires history, laboratory results, symptoms, and further investigation. The work applies one open-source dataset and does not involve any external, prospective or clinician-based validation. Patient-level independence could not be verified, the findings are based on a single run and no subgroup fairness was checked. Further limitations include: infection mask generation for only part of COVID-19 images, lung region masks instead of lesion masks for Lung Opacity, classifier-only dropout, T=50 inference cost, dataset-specific threshold, no deep ensemble and no risk–coverage/AURC evaluation. MIMIC-CXR [25] requires ontology-aware external validation.

Table 6: Selective prediction at the validation-selected $\tau = 0.35$. False-referral and silent-failure denominators are all 4,530 test images.

Metric	$T = 10$	$T = 50$
Error recall (%)	78.8	**90.9**
False referral (%)	2.21	**1.92**
Referral rate (%)	3.93	**3.91**
Coverage (%)	96.07	**96.09**
Retained accuracy (%)	99.52	**99.79**
Silent failures (%)	0.46	**0.20**

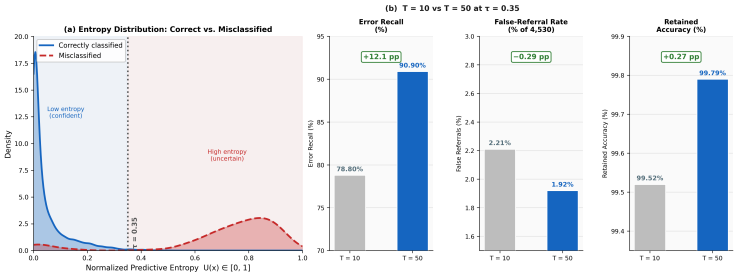

Fig. 4: Normalized predictive-entropy distributions and comparison of $T = 10$ with $T = 50$ at the validation-selected operating point.

6 Conclusion

The study demonstrates retrospective feasibility on COVID-QU-Ex: 97.81% accuracy, Error-AUROC 0.941, 90.9% error capture at 3.91% referral, and COVID-19 infection-region PG 82.05% and IoU 0.63. External, prospective, and reader studies are required before claims of clinical utility. **Prospects of Application.** *A future interface could present predictions, saliency, and uncertainty for radiologist prioritization; its workflow effects require prospective evaluation.*

References

1. Rajpurkar, P., et al.: CheXNet: Radiologist-level pneumonia detection. arXiv:1711.05225 (2017)
2. Irvin, J., et al.: CheXpert: A large CXR dataset with uncertainty labels. In: Proc. AAAI, vol. 33, pp. 590–597 (2019)
3. Kompa, B., Snoek, J., Beam, A.L.: Second opinion needed. npj Digital Medicine **4**(1), 1–6 (2021)
4. Selvaraju, R.R., et al.: Grad-CAM: Visual explanations. In: Proc. ICCV, pp. 618–626 (2017)
5. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++. In: Proc. WACV, pp. 839–847 (2018)
6. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation. In: Proc. ICML, pp. 1050–1059 (2016)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proc. ICML, pp. 1321–1330 (2017)
8. Geifman, Y., El-Yaniv, R.: Selective prediction in deep neural networks. In: Proc. NeurIPS, vol. 30 (2017)
9. Ghoshal, B., Tucker, A.: Estimating uncertainty in deep learning for COVID-19 detection. arXiv:2003.10769 (2020)
10. Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. IJCV **126**(10), 1084–1102 (2018)
11. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized input sampling for explanation of black-box models. In: Proc. BMVC (2018)
12. Wang, H., et al.: Score-CAM. In: Proc. CVPR Workshops (2020)
13. Muhammad, M.B., Yeasin, M.: Eigen-CAM. In: Proc. IJCNN (2020)
14. Topol, E.J.: High-performance medicine. Nature Medicine **25**(1), 44–56 (2019)
15. Ahmed, F., Naz, N.S., Khan, S., et al.: Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges. BMC Medical Imaging **26**, Article 37 (2026). `https://doi.org/10.1186/s12880-025-02118-w`
16. Sridharan, S., et al.: Real-World evaluation of an AI triaging system for chest X-rays: a prospective clinical study. European Journal of Radiology **181**, 111783 (2024). `https://doi.org/10.1016/j.ejrad.2024.111783`
17. Yoon, S.H., et al.: Use of artificial intelligence in triaging of chest radiographs to reduce radiologists' workload. European Radiology **34**(2), 1094–1103 (2024). `https://doi.org/10.1007/s00330-023-10124-1`
18. Huang, G., Liu, Z., Van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proc. CVPR, pp. 4700–4708 (2017)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning. In: Proc. CVPR, pp. 770–778 (2016)
20. Simonyan, K., Zisserman, A.: Very deep CNNs for large-scale image recognition. In: Proc. ICLR (2015)
21. Szegedy, C., et al.: Rethinking the inception architecture. In: Proc. CVPR, pp. 2818–2826 (2016)
22. Tan, M., Le, Q.: EfficientNet. In: Proc. ICML, pp. 6105–6114 (2019)
23. Wang, L., Lin, Z.Q., Wong, A.: COVID-Net. Scientific Reports **10**(1), 19549 (2020)
24. Tahir, A.M., et al.: COVID-QU-Ex dataset. Mendeley Data V1. `https://doi.org/10.17632/9xkhgts2s6.1` (2021)
25. Johnson, A.E., et al.: MIMIC-CXR. Scientific Data **6**(1), 317 (2019)

26. Chowdhury, M.E.H., et al.: Can AI help in screening COVID-19? IEEE Access **8**, 132665–132676 (2020)
27. Rahman, T., et al.: Image enhancement for COVID-19 detection. Comput. Biol. Med. **132**, 104319 (2021)
28. Adadi, A., Berrada, M.: Peeking inside the black-box. IEEE Access **6**, 52138–52160 (2018)
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. ICLR (2019)