Lung Disease Clasification - Applied AI Progress Report

Group-Q

Abstract

The ABSTRACT is to be in fully justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length. Look at previous CVPR abstracts to get a feel for style and length. Look at previous CVPR abstracts to get a feel for style and length. Look at previous CVPR abstracts to get a feel for style and length. Look at previous CVPR abstracts to get a feel for style and length. Look at previous CVPR abstracts to get a feel for style and length. Look at previous CVPR abstracts.

1. Introduction-1.5

Early diagnosis of respiratory diseases like pneumonia and COVID-19 leads to decreased mortality rate [7] and is a powerful way to manage a pandemic [38]. These diseases can be diagnosed using a variety of tests like pulse oximetry, chest x-ray, CT scan [29], PCR [1] however chest X-rays are by far the most accessible [9] to low and middle income countries. Furthermore, the scan is available in minutes making it one of the fastest ways of diagnosis [30]. However, the bottleneck with this method is the need for an expert radiologists to evaluate the scan [20]. Many researchers have tried to solve this problem by creating a deep learning based lung disease classification system [34] but haven't been able to come up with models that can replace radiologists. Small [12] and highly imbalanced data [34], along with varying specifications of X-ray scanners leading to low inter-hospital accuracy [26] are the biggest problems that researchers have faced. Another issue with using deep neural networks in medical settings is its black-box nature [3], doctors and patients will not trust a model that cannot explain its results [21].

This project is an attempt to compare three CNN backbone architectures namely, ResNet-34, MobileNet V3 Large and EfficientNet B1 along with three lung disease datasets to identify the type of architecture that works best for lung disease classification. Two of the datasets used presented a multiclass classification problem with 3 classes while the third dataset presented a multiclass, multilabel classification problem. A total of 12 models were trained in this study, four for each of the three datasets. The first three models for each dataset was trained from scratch and the fourth model was trained using transfer learning. Transfer learning was performed by deep-tuning ImageNet weights and the performance was evaluated to check improvement over the models trained from scratch. The small dataset problem and the issue of different radiographic contrast [22] is mitigated using data augmentation. Imbalanced data problem is handled by undersampling the majority class. The hyperparameters were fixed across models and the F1 scores and cross entropy loss have been used to compare models and select the best overall model. All the models were optimized using the Adam optimizer [18] with default parameters and the cosine annealing [19] learning rate scheduler was used to decrease the learning rate as training progressed. Further, an ablation study was performed to find the best learning rate for the selected model. Finally, GradCAM [11] and T-SNE were used to visualize the trained models and understand model predictions better. An F1 score of 0.8 and 0.98 was achieved for the two multiclass datasets, whereas the maximum F1 for the multilabel dataset with 7 classes was 0.46.

Related Works: Li *et al.* [] were among the first to use CNNs in a medical setting. They used a single convolutional layer to classify interstitial lung diseases using CT scans, achieving better performance than existing approaches. Since then there has been a dramtic increase in application of CNNs in healthcare, deep neural networks have been used to perform various tasks like segmenting regions of interest in MRI [,], classifying X-Ray [], MRI [], and CT [] scans. Further, GANs have been used to generate high quality scans [] when there is a lack of available data due to either privacy reasons or availability of subjects. GANs have also been used to generate high quality CT scans from MRI scans []. Apart from radiographic scans, deep CNNs have also been used to detect malarial parasite in blood smear images [] with an accuracy of 99.96%. Another interesting application is the use of 1-D convolutions to detect heart anomalies using ECG data []. Researchers have also used architectures like the Inception V3 to

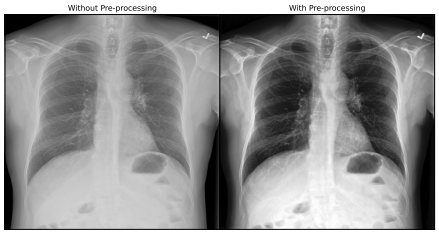


Figure 1. Effect of pre-processing on Chest X-ray images.

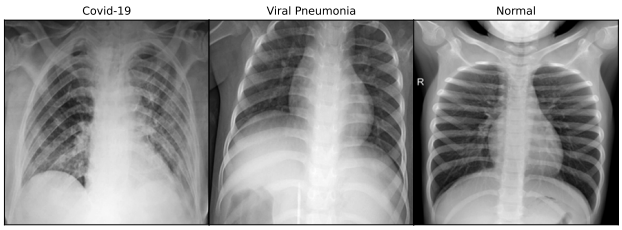
perform dermatologist level skin-cancer detection using skin lesion images [] using transfer learning.

In the recent years, many researchers have tried to predict lung diseases using deep CNNs, Wang *et al.* [35] used state of the art backbone architectures to train a lung disease classifier for multilabel data by training only the prediction and transition layers from scratch and leaving pre-trained ImageNet weights freezed while training. They achieved a high AUC of over 0.6 for most of the classes in the dataset with this technique. Rajpurkar *et al.* [] created a 121 layer deep CNN - CheXNet to detect pneumonia using chest X-rays with radiologist level accuracy. Labhane *et al.* [] used transfer learning with state of the art backbone architectures like VGG16, VGG19 and InceptionV3 to predict pneumonia in pediatric patients and achieved an F1 score of 0.97. Islam *et al.* combined CNN and LSTM to create a COVID-19 detector []. The CNN extracted complex features from scans and the LSTM was used as a classifier. This method resulted in an improvement over a vanilla CNN network and an F1 score of 98.9% was achieved. Abbas *et al.* [] created the DeTraC network to detect COVID in chest X-rays that improved performance of existing backbone models significantly with the highest accuracy of 98.23% using the VGG19 architecture. Guefrechi *et al.* [] on the other hand used data augmentation techniques like random rotation, flipping and noise with transfer learning on backbone architectures like ResNet50, InceptionV3 and VGG16 to achieve a high accuracy of 98.3%.

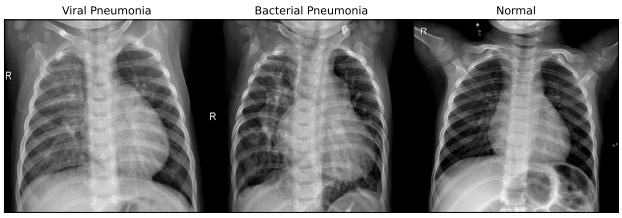
In the following sections methodology of the approach and the results will be discussed.

2. Methodology-2

Datasets: (Tab. 1) with varying disease types were chosen to ensure model robustness. Other criteria included the *number of images per class* and *image quality* as noisy scans can lead to mis-diagnosis [28]. The **COVID** dataset was created using 43 [8] different publications. [5, 6, 25] X-rays with widespread, hazy, and irregular ground glass opacities are of the COVID-19 class [16]. Whereas, the ones with haziness only in the



(a) COVID Dataset



(b) Pneumonia Dataset

(c) Chest X-Ray 8 Dataset

Figure 2. Sample Chest X-rays from the datasets used.

Dataset	No. of Images	Classes	Size
COVID [5, 6, 25]	10000k:3.6k:1.3k	3	299^2
Pneumonia [17, 33]	3k:1.5k:1.5k	3	224^2
Chest X-Ray8 [35, 36]	7.2k:7k:7k:4.1k :3.9k:3.5k:2.9k	7	1024^2

Table 1. Shortlisted Datasets.

lower regions [39] are viral pneumonia cases as shown in Fig. 2. The **Pneumonia**, dataset contains scans from pediatric patients of one to five year olds collected as part of patients' routine clinical care. [17,33] Scans with one white condensed area affecting only one side of the lungs are tagged as bacterial pneumonia [2]. Whereas, X-rays which show bilateral patchy areas of consolidation are classified as viral pneumonia [13]. The **Chest X-ray 8** dataset was released by NIH [37] with over 100k chest X-ray images and their radiological reports which Wang *et al.* [35] used to create disease labels through NLP. [36] It contains 15 classes but only 7 were chosen for this study. This dataset is significantly different from the other two as it is a multilabel dataset. Classes were iteratively removed, ensuring that classes are not highly imbalanced to finally reach the 7 classes. With over 29,000 images of size 1024 x 1024, this dataset was the biggest and thus had to be resized down to 384 x 384 to reduce training time.

Arch.	Params (Mil.)	Layers	FLOPS (Bil.)	Imagenet Acc.
MobileNet	5.5	18	8.7	92.6
EfficientNet	7.8	25	25.8	94.9
Resnet	21.8	34	153.9	91.4

Table 2. Shortlisted Backbone Architectures.

Furthermore, normal class images were undersampled to choose only 7000 scans. The data consisted of multiple scans from the same subject which would lead to data leakage between the train, val and test sets. This was prevented with the use of GroupShuffleSplit from the scikit library.

Before training, all the images were pre-processed using histogram equalization and Gaussian blur with a 5x5 filter as Giełczyk *et al.* [10] showed that this improved the F1 score by about 4% for chest X-ray classification. Visually, the contrast of the scan improved and allowed irregularities to stand out as shown in Fig. 1. Next, the scans were divided into train, validation and test with the 70:15:15 split. During training, the scans were augmented using RandomHorizontalFlip, RandomAdjustSharpness, and RandomAutocontrast in Pytorch [24] to increase the number of images the model gets to learn from and ensure that the model is robust to scans from different machines.

Backbone Architectures: (Tab. 2) of various configuration and blocks were chosen. Other selection criteria were the *number of trainable parameters*, important as total training time and hardware resources are limited for this project and the *top 5 classification accuracy* on the ImageNet 1K benchmark dataset. **ResNet 34**: residual learning network with 34 layers that are made possible by skip connections. The 34 layer variant was chosen to decrease training time for this study. [14] **MobileNet V3 Large**: uses depthwise separable convolution from MobileNet V2 [27] along with squeeze-excitation blocks in residual layers from MnasNet [31]. Howard *et al.* [15] also used network architecture search to find the most effective model. The large configuration was chosen to not compromise on the prediction accuracy. **EfficientNet B1**: uses compound scaling to scale the model by depth, width and resolution. The B1 version was chosen to have faster training without compromising on the accuracy. [32]

Optimization Algorithm: The Adam optimizer [18] was chosen as the algorithm of choice as it converges faster on image classification tasks and does not requires little tuning. It integrates benefits of RMSProp and Adagrad to produce robust results on a wide range of problems.

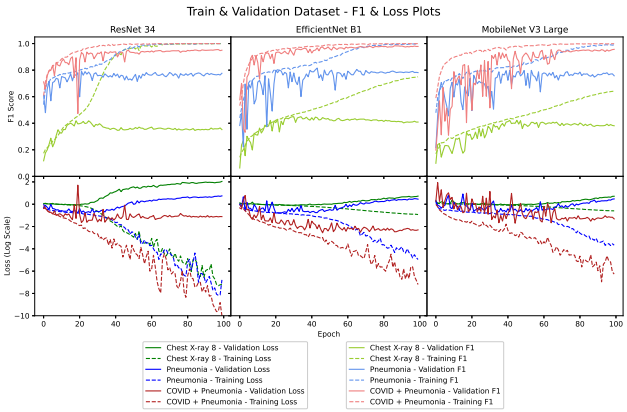


Figure 3. Train & Val F1 & Loss plots for the 9 models.

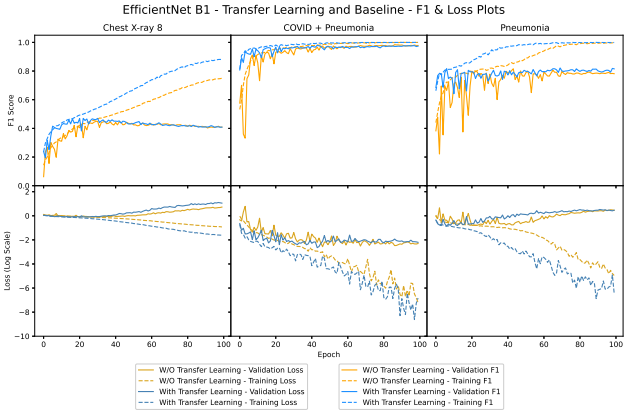


Figure 4. Train & Val, F1 & Loss plots for EfficientNet trained from scratch and with ImageNet weights.

3. Results-2.5

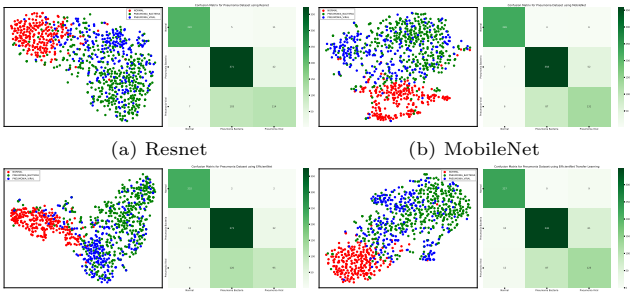
Experiment Setup:

Two datasets in this study had a very small number of samples which caused the models to overfit early. To mitigate this, random contrast and sharpness adjustment [23] data augmentation techniques were used. Some scans in the datasets were anterior-posterior while some others were posterior-anterior and using the horizontal flip data augmentation would make the model invariant to these differences [4]. Inception was the first model trained and each epoch took over 1 hour. To reduce the training time, the X-ray images were resized, pre-processed and split into train, test and validation sets separately. Furthermore, EfficientNet, MobileNet and ResNet 34 were chosen as they have a considerably low number of learnable parameters. Now each epoch is taking less than 4 minutes.

Nine models were trained from scratch and the train-

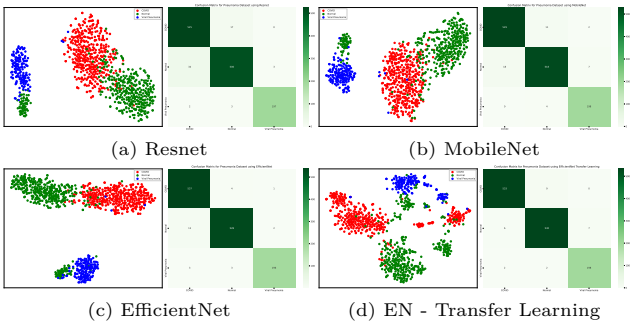
Model	ResNet			MobileNet			EfficientNet			EN - Transfer Learning		
Dataset	*F1*	*Time*	*Epoch*	*F1*	*Time*	*Epoch*	*F1*	*Time*	*Epoch*	*F1*	*Time*	*Epoch*
Pneumonia	0.784	82	**22**	**0.804**	**75**	42	0.768	110	44	0.782	114	70
COVID	0.963	68	71	0.959	**45**	82	0.970	80	89	**0.978**	56	**46**
X-Ray 8	0.411	11,502	**19**	0.406	**7,275**	42	0.445	13,820	31	**0.457**	13,813	29

Table 3. F1 (higher is better), time per epoch in seconds (lower is better), and number of epochs to reach the best validation loss (lower is better) for the 12 models that were trained.



(a) Resnet (b) MobileNet (c) EfficientNet (d) EN - Transfer Learning

Figure 5. T-SNE and Confusion matrics of the Pneumonia dataset.

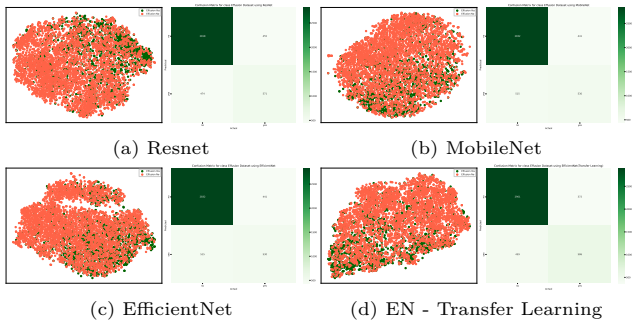


(a) Resnet (b) MobileNet (c) EfficientNet (d) EN - Transfer Learning

Figure 6. T-SNE and Confusion matrics of the COVID dataset.

ing, validation F1 score and loss can be seen in Fig. 3. From the plots it is clear that going from a smaller architecture to a bigger architecture, makes the model start to overfit earlier. Another interesting observation is that cosine annealing impacted the loss of MobileNet the most every 10 epochs due to warm restarts. From the graphs it can be seen that all three datasets had similar performance across models when trained for a high number of epochs. The X-ray 8 dataset performed the worst among the three datasets which could be due to the high number of classes as compared to the other datasets. Surprisingly, the pneumonia dataset performed worse than the COVID + pneumonia dataset which indicates that COVID cases are easier to distinguish from pneumonia cases.

Main Results:



(a) Resnet (b) MobileNet (c) EfficientNet (d) EN - Transfer Learning

Figure 7. T-SNE and Confusion matrics of the Chest X-ray 8 dataset.

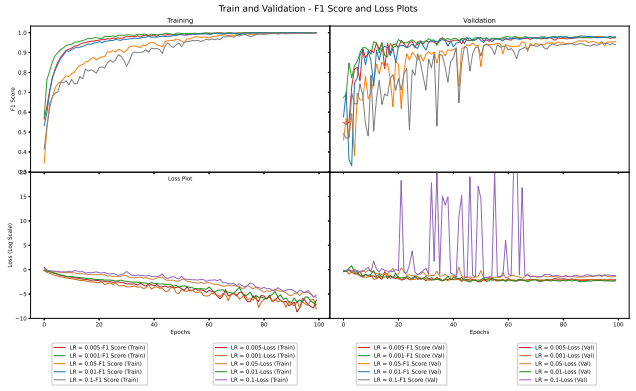


Figure 8. Train & Val, F1 & Loss plots for ablative study models.

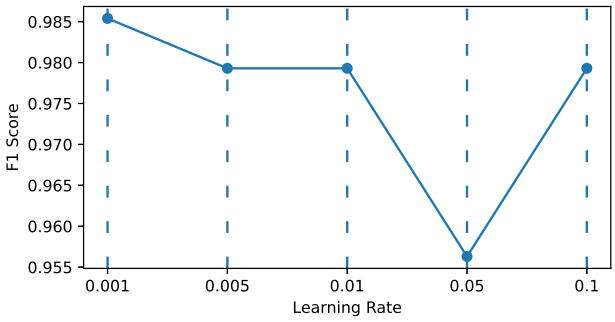


Figure 9. Ablative Study F1 scores (Higher is better).

Ablative Study:

References

[1] Noorullah Akhtar, Jiyuan Ni, Claire Langston, Gail J Demmler, and Jeffrey A Towbin. Pcr diagnosis of viral pneumonitis from fixed-lung tissue in children. *Biochemical and molecular medicine*, 58(1):66–76, 1996. 1

[2] How Drugs are Made and Product List. Viral vs. bacterial pneumonia: Understanding the difference. https://www.pfizer.com/news/articles/viral_vs_bacterial_pneumonia_understanding_the_difference, 2020. 2

[3] Paul J. Blazek. Why we will never open deep learning's black box. https://towardsdatascience.com/why-we-will-never-open-deep-learnings-black-box-4c27cd335118, 2022. 1

[4] Aleksander Botev, Matthias Bauer, and Soham De. Regularising for invariance to data augmentation improves supervised learning. *arXiv preprint arXiv:2203.03304*, 2022. 3

[5] Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul Islam. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020. 2

[6] Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul Islam. Covid-19 radiography database. https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database, 2021. 2

[7] Priya Daniel, Chamira Rodrigo, Tricia M Mckeever, Mark Woodhead, Sally Welham, and Wei Shen Lim. Time to first antibiotic and mortality in adults hospitalised with community-acquired pneumonia: a matched-propensity analysis. *Thorax*, 71(6):568–570, 2016. 1

[8] Società Italiana di Radiologia. Covid pneumonia dataset. https://sirm.org/category/senza-categoria/covid-19/, 2020. 2

[9] Guy Frija, Ivana Blažić, Donald P Frush, Monika Hierath, Michael Kawooya, Lluis Donoso-Bach, and Boris Brkljačić. How to improve access to medical imaging in low-and middle-income countries? *EClinicalMedicine*, 38:101034, 2021. 1

[10] Agata Giełczyk, Anna Marciniak, Martyna Tarczewska, and Zbigniew Lutowski. Pre-processing methods in chest x-ray image classification. *Plos one*, 17(4):e0265949, 2022. 3

[11] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 1

[12] Sarra Guefrechi, Marwa Ben Jabra, Adel Ammar, Anis Koubaa, and Habib Hamam. Deep learning based detection of covid-19 from chest x-ray images. *Multimedia Tools and Applications*, 80(21):31803–31820, 2021. 1

[13] W Guo, J Wang, M Sheng, M Zhou, and L Fang. Radiological findings in 210 paediatric patients with viral pneumonia: a retrospective case study. *The British journal of radiology*, 85(1018):1385–1389, 2012. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[15] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 3

[16] Adam Jacobi, Michael Chung, Adam Bernheim, and Corey Eber. Portable chest x-ray in coronavirus disease-19 (covid-19): A pictorial review. *Clinical imaging*, 64:35–42, 2020. 2

[17] Daniel Kermany, Kang Zhang, Michael Goldbaum, et al. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2(2), 2018. 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1, 3

[19] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[20] P Mehrotra, V Bosemani, and J Cox. Do radiologists still need to report chest x rays? *Postgraduate medical journal*, 85(1005):339–341, 2009. 1

[21] Aleksandra Mojsilovic. Introducing ai explainability 360. https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/, 2019. 1

[22] Andrew Murphy. Radiographic contrast. https://radiopaedia.org/articles/radiographic-contrast, 2022. 1

[23] Loris Nanni, Michelangelo Paci, Sheryl Brahnam, and Alessandra Lumini. Comparison of different image data augmentation approaches. *Journal of Imaging*, 7(12):254, 2021. 3

[24] PyTorch. Transforming and augmenting images. https://pytorch.org/vision/stable/transforms.html. 3

[25] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M. Zughaier, Muhammad Salman Khan, and Muhammad E.H. Chowdhury. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in Biology and Medicine*, 132:104319, 2021. 2

[26] Melissa Rohman. Ai performs poorly when tested on data from multiple health systems. https://healthimaging.com/topics/artificial-intelligence/ai-poorly-detects-pneumonia-chest-x-rays, 2018. 1

[27] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3

[28] Janaki Sivakumar, K Thangavel, and P Saravanan. Computed radiography skull image enhancement using wiener filter. In *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)*, pages 307–311. IEEE, 2012. 2

[29] Matt Smith. Common lung diagnostic tests. https://www.webmd.com/lung/breathing-diagnostic-tests, 2022. 1

[30] Healthwise Staff. Chest x-ray. https://www.healthlinkbc.ca/tests-treatments-medications/medical-tests/chest-x-ray, 2021. 1

[31] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019. 3

[32] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3

[33] Tolga. Chest x-ray images. https://www.kaggle.com/datasets/tolgadincer/labeled-chest-xray-images, 2020. 2

[34] Guangyu Wang, Xiaohong Liu, Jun Shen, Chengdi Wang, Zhihuan Li, Linsen Ye, Xingwang Wu, Ting Chen, Kai Wang, Xuan Zhang, et al. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and covid-19 pneumonia from chest x-ray images. *Nature biomedical engineering*, 5(6):509–521, 2021. 1

[35] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 2

[36] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Nih chest x-rays. https://www.kaggle.com/datasets/nih-chest-xrays/data, 2017. 2

[37] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, and Summers RM. Nih clinical center provides one of the largest publicly available chest x-ray datasets to scientific community. https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community, 2017. 2

[38] Lizhou Xu, Danyang Li, Sami Ramadan, Yanbin Li, and Norbert Klein. Facile biosensors for rapid detection of covid-19. *Biosensors and Bioelectronics*, 170:112673, 2020. 1

[39] Na Zhan, Yingyun Guo, Shan Tian, Binglu Huang, Xiaoli Tian, Jinjing Zou, Qiutang Xiong, Dongling Tang, Liang Zhang, and Weiguo Dong. Clinical characteristics of covid-19 complicated with pleural effusion. *BMC Infectious Diseases*, 21(1):1–10, 2021. 2