2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC) | 978-1-6654-4415-6/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICSCCC51823.2021.9478152

Explainable Deep Neural Models for COVID-19 Prediction from Chest X-Rays with Region of Interest Visualization

Ishan Mathew Nedumkunnel
Department of Information Technology
National Institute of Technology
Surathkal, India
0000-0003-3144-6438

Linu Elizabeth George
Department of Information Technology
National Institute of Technology
Surathkal, India
0000-0001-6184-5291

Neil Abraham Rosh
Department of Information Technology
National Institute of Technology
Surathkal, India
0000-0003-3390-8392

Sowmya Kamath S
Department of Information Technology
National Institute of Technology
Surathkal, India
0000-0002-0888-7238

Veena Mayya
Department of Information Technology
National Institute of Technology
Surathkal, India
0000-0002-8091-5053

***Abstract*—COVID-19 has been designated as a once-in-a-century pandemic, and its impact is still being felt severely in many countries, due to the extensive human and green casualties. While several vaccines are under various stage of development, effective screening procedures that help detect the disease at early stages in a non-invasive and resource-optimized manner are the need of the hour. X-ray imaging is fairly accessible in most healthcare institutions and can prove useful in diagnosing this respiratory disease. Although a chest X-ray scan is a viable method to detect the presence of this disease, the scans must be analyzed by trained experts accurately and quickly if large numbers of tests are to be processed. In this paper, a benchmarking study of different preprocessing techniques and state-of-the-art deep learning models is presented to provide comprehensive insights into both the objective and subjective evaluation of their performance. To analyze and prevent possible sources of bias, we preprocessed the dataset in two ways - first, we segmented the lungs alone, and secondly, we formed a bounding box around the lung and used only this area to train. Among the models chosen to benchmark, which were DenseNet201, EfficientNetB7, and VGG-16, DenseNet201 performed better for all three datasets.**

***Index Terms*—COVID-19 Diagnosis, Deep neural networks, Disease prediction, Medical imaging, Region of Interest**

I. Introduction

The first recorded case of COVID-19, caused by the coronavirus named SARS-CoV-2, was back in November 2019, in Wuhan, China. Since then, due to its unique attributes, it has proven to be an extremely infectious virus, infecting 67.6 million people as of December 2020. COVID-19, like other coronaviruses, affects respiratory organs significantly, causing shortness of breath and other similar symptoms. In severe cases, it can evolve into pneumonia. The COVID-19 virus's presence is detected most accurately using the RT-PCR (Real-Time reverse transcription-polymerase Chain Reaction) test. PCR test results may take a few hours to days, depending on the testing site. PCR sample collection and analysis is a highly manual and laborious process, and reagent contamination can cause a high false-positive rate [1].

In Chest X-rays and Computed Tomography (CT) Scans, COVID-19 appears as peripheral and bilateral nodular ground-glass opacities and consolidation [2]. Therefore, these tools are also used, albeit not as the primary diagnostic tool, to detect the virus's presence. Using computer vision to aid diagnosing in the medical industry helps save medical professionals' precious time, especially in a pandemic, where the healthcare services are overwhelmed with a flood of patients coming in. In this regard, deep neural models like Convolutional Neural Networks (CNNs) have shown to be very useful for medical image classification purposes. Transfer learning, in particular, helps overcome challenges like insufficient training data while still attaining reliable results. Our study aims to compare the performance of different state-of-the-art model architectures for multi-class prediction of two lung anomalies, COVID-19 & Pneumonia, and normal (healthy) lungs with different preprocessing techniques. Based on our analysis of existing literature on detecting COVID-19 and Pneumonia from Chest X-rays, we shortlisted three models to benchmark. Furthermore, we implemented various preprocessing methods like segmenting the lung and forming the bounding box to avoid bias. We have also used visualisation techniques to compare the features learned by these models.

This paper is organized as follows: II presents a detailed discussion on the existing research in the area of interest. In Section III, we describe the different models that were chosen for the benchmarking study, the datasets and different preprocessing methods adopted to eliminate bias. Section IV presents the observations with reference to the comparative

performance of the models, followed by conclusion and future directions of research.

II. Related Work

Image classification techniques have provided significant results for diagnosis and prognosis in the medical field. This method is applied to images of X-rays, CT scans, and even MRI scans. Poddar et al [3] proposed a hybrid model called VDSNet which combines CNN, VGG, data augmentation and spatial networks (STN) to detect lung diseases and reported a validation accuracy of 73%. The dataset used consisted of chest X-ray images and some additional information such as gender as well as age. Kieu et al [4] attempted to detect abnormal regions from a minimal set of features through a model that uses multiple Convolutional Neural Networks. ConvnetJS library was used to build the proposed mutli-CNN model. The classifiers of the model finally provided the normal/abnormal density of each image and Fusion rule was employed to compute the results.

There exist several studies that examine the performance and effectiveness of transfer learning on natural image datasets. One such study by Wu et al [5], implemented and compared ResNet50, VGG-19, VGG-16, and Inception-v3, and found that the accuracy of flower recognition is improved with transfer learning when compared to traditional methods. Hon et al [6] used VGG and Inception Models for transfer learning to overcome the major challenge that required a large dataset for training, as well as to optimize the architecture of the deep learning model. Hashmi et al [7] put forward a weighted classifier for the task of penumonia detection. This paper uses a number of state-of-the-art models namely InceptionV3, Xception, ResNet18, MobileNetV3, and, DenseNet121 and combines the predictions made by the same in an optimal way. Ozturk et al [8] builds upon the Darknet-19 model termed as 'You Only Look Once', which performs object detection in real-time, to classify images of X-rays. DarkCovidNet, the model proposed, had been trained on three labels: Pneumonia, COVID-19 and No-Findings for which it attained an average classification accuracy of 87.02%.

Rahimzadeh et al. [9] presented a concatenated neural network and achieved an overall accuracy of 91.4% between five folds to classify the chest X-rays into three classes: normal, COVID-19, and pneumonia. The features extracted from ResNet50V2 and Xception models were used to build the concatenated model and finally connected to a convolutional layer for classification. Bai et al. [10] attempted to classify between COVID-19 viral pneumonia contagion and other viral pneumonia from chest CT scans. CovaidAID [11] used the pre-trained model of CheXNet to classify CXR images into the No-Findings, COVID-19 and the various types of pneumonia. Though the size of the dataset used was small, the proposed model achieves an accuracy of 90% and also a sensitivity of 100% for the COVID-19 infection. Jaiswal et al [12] utilized DenseNet for COVID-19 infection classification, however a binary classification problem was tackled, COVID-19 vs Normal. Zebin et al [13] used the Cohen dataset for COVID-19 positive X-rays used VGG16, ResNet50, and EfficientNetB0 to predict whether a given X-ray is COVID-19 positive, Pneumonia positive or normal. They implemented CycleGAN architecture for increasing the under-represented COVID-19 class images.

Based on the extensive literature review conducted, several potential directions for further research were identified. Several works employ state-of-the-art deep learning models like variations of VGG, ResNet, DenseNet, MobileNet, Xception for these prediction tasks. These models have performed well in different image classification scenarios, and are therefore are well suited for COVID-19 detection as well. A few researchers have experimented with different preprocessing techniques to ensure the identification of the Region of Interest. Most works have reported their results using standard metrics, without showcasing the features that the models have picked up. This opens up potential new avenues to explore for visualization of learned features, for more intuitive predictions.

III. Proposed Methodology

A. Datasets

The frontal chest X-ray images have been gathered from a variety of sources that are available for public use, collected by Cohen et al [14], Daniel et al [15] and Linda et al [16]. The dataset contains X-ray image scans with 3 labels - COVID-19, NORMAL (healthy), and PNEUMONIA. There are a total of 6432 frontal chest X-ray images, out of which around 20% of them have been used for testing. Table I shows the train/test data split.

TABLE I. number of images per class after data-split

Split/Label	COVID-19	Normal	Pneumonia	Total
Train	460	1266	3418	5144
Test	116	317	855	1288
Total (class-wise)	576	1583	4273	6432

We used online (real-time) data augmentation, leveraging the *ImageDataGenerator* library of Keras [17]. This library performs data augmentation in real-time on each batch of images, as it is being fed to the model for training. Augmentation techniques like horizontal flipping, shifting, altering brightness and random zoom have been used. To remove potential bias in the data and identify the ROI within each scan (which is around the lung area), we used two different approaches - *Segmentation* and *bounding box*, which are described below.

1) Segmentation: The first approach is to segment the lungs, for removal of any bias sources like the presence of different texts on the X-ray, or parts of the body apart from the lungs that might be visible and so on. For this task, the standard U-Net [18] architecture has been used, which was trained on Shenzhen Hospital X-ray and Montgomery County X-ray Set [19]. The proposed model [20] for medical image segmentation was built for the RSNA Pneumonia Detection Challenge on Kaggle. The U-Net model was specifically built for biomedical image segmentation, and therefore serves our purpose well.

2) Bounding Box: To prevent the model from learning features from the intricate mask edges, another approach was implemented. We generated bounding boxes around the contours of the lung mask to create clean edges. This also ensures that minimal information is lost because of potential error while generating the mask. Fig. 1 and 2 show illustrate the process.

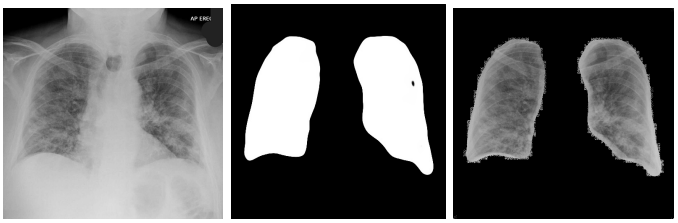

Fig. 1. Original X-ray along with generated mask and final overlay image

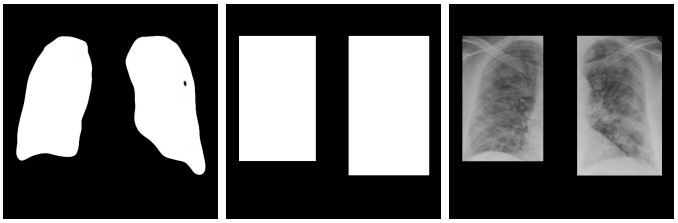

Fig. 2. Generated mask from Segmentation process, mask with bounding box, and final overlay image

B. Deep Neural Models

Based on the extensive review of existing works and their use in other similar classification scenarios, the following models were chosen for benchmarking experiments. Each of these models have been trained and tested on all 3 types of datasets (mentioned above). Transfer learning is used to compensate for the relatively small dataset that is currently available to us - this improves training speed as the model is already able to discern some basic low-level features beforehand, and needs to be trained to recognise more relevant and high-level features. The models are described below.

1) DenseNet201: - In the DenseNet model, features learned by all previous layers are used by the classifier due the densely connected nature of the model's convolutional layers. The decision boundaries obtained from this model tend to be smoother. Rich patterns and more diverse features can be learned by this model. It does not face the vanishing gradient problem either, and takes less time to train due to its relatively low number of parameters (both achieved by dense connections across convolutional layers). The variant of DenseNet used is DenseNet201, which is 201 layers deep. For our experimental purposes, only the layers from the fifth convolutional block onwards (including) are made trainable. The total number of trainable parameters are 8,088,579 and the number of non-trainable parameters are 11,339,584.

2) EfficientNetB7: - EfficientNet [21] in comparison to most CNN models, achieves both a greater accuracy and is more efficient, while reducing parameter size and FLOPS by a significant margin. EfficientNet is particularly useful for using deep learning on the edge, as it reduces compute cost, battery usage, and also training and inference speeds. It has also been shown that the latest version of EfficientNet, i.e., EfficientNet-B7, has the highest accuracy among all, with less number of parameters. For our experimental purposes, only the layers from the seventh convolutional block onwards (including) are made trainable. The model has a total of 24,551,235 trainable parameters and 41,021,264 non-trainable parameters.

3) VGG-16: - The VGG-16 model is one of the most versatile and widely used pre-trained models for image classification, and has hence been used for benchmarking. For the purposes of this paper, layers after the fifth convolutional block (including) are made trainable. The number of trainable parameters is 7,677,827, while the number of non-trainable parameters is 7,635,264.

For the task of classifying the chest X-ray images, we adopted transfer learning techniques. Transfer learning performs well in training DNNs with comparatively little data. All the models have been pretrained on the Imagenet database [22], which is widely used for research. The model is loaded using a generic and well-trained image classification network for feature extraction, and then by adding a few layers to suit the classification task at hand. Fig. 3 shows the layers that have been added:

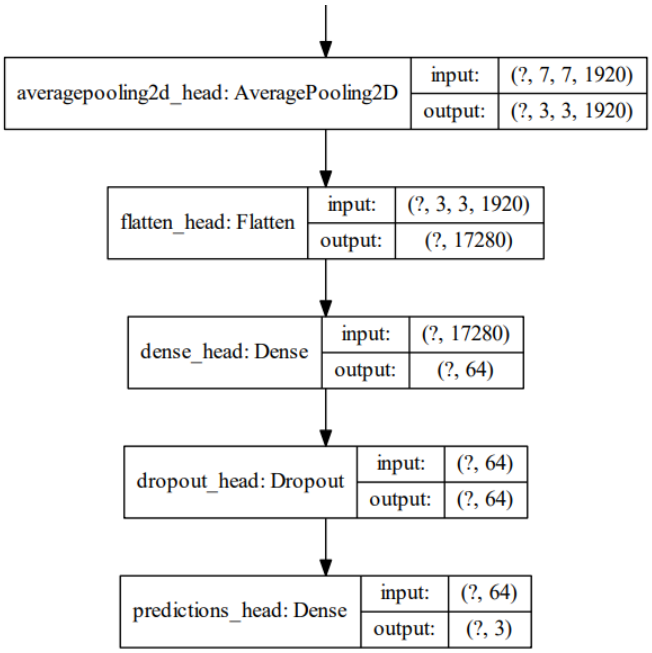

Fig. 3. Classification Task - Layers added

For visualisation, heatmaps are generated using the Gradient-weighted Class Activation Mapping (Grad-CAM) technique [23]. This technique shows how important each pixel (location) is for the considered class by producing heatmaps of 2D class activation over the input image. Grad-CAM performs visualisation in the following manner: Given an input image,

the output feature map of a particular convolution layer is taken. Each channel in this feature map is weighed by the gradient of the corresponding output class with respect to that particular feature map. Using Grad-CAM, we can validate visually where our network is learning from, verifying that the correct patterns in the image are being looked at and that the model is indeed picking up those patterns.

IV. Experimental Results and Discussion

For the experimental validation, we experimented with all 3 models, training them for 20 epochs on each of the 3 datasets created as described earlier. The images were resized to a shape of 224*224*3, and a batch size of 16 (pre-augmentation) was used. The model was trained using the Adam optimizer (lr=0.0001), and categorical cross-entropy was used as the loss function.

For each dataset, the training accuracy and loss at each epoch for all models have been plotted. This allows us to compare the training performance of the models on the same dataset. The metrics used are - recall (1), precision (2), F1-score (3) and accuracy (4). Recall for a particular class refers to the portion of correctly predicted images for that class out of all the images in that class. Precision for a particular class refers to how many images are truly of that class, out of all those predicted to be in that class. F1-score for a particular class is the harmonic mean of the recall and precision of that class. Accuracy is the number of accurately predicted images out of all the images.

$$recall = \frac{TP}{FN + TP} \tag{1}$$

$$precision = \frac{TP}{FP + TP} \tag{2}$$

$$F1 - score = 2. \frac{recall\ .\ precision}{recall + precision} \tag{3}$$

$$Accuracy = \frac{AccuratelyPredictedImages}{TotalNumberofImages} \tag{4}$$

where FP is the number of False Positives, FN is the number of False Negatives, TN is the number of True Negatives and TP is the number of True Positives.

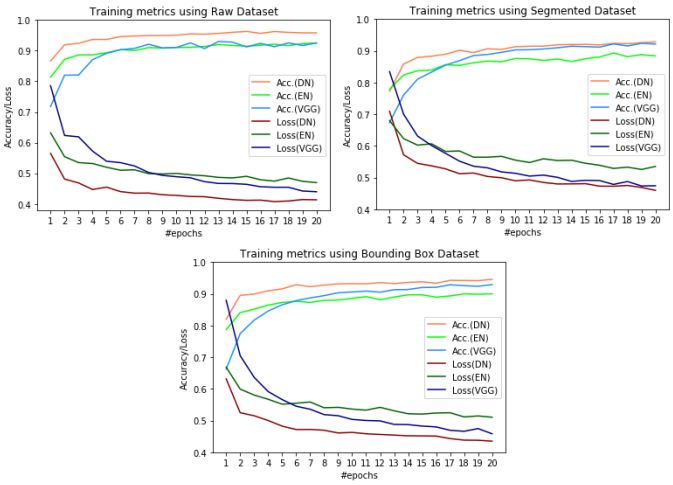

Fig. 4. Comparison of training metrics of all model-dataset combinations

Recall (1), Precision (2) and F1-Score (3) are computed for each of the labels - Pneumonia, COVID-19 and Normal, for all 3 types of datasets. The accuracies (4) are computed for each dataset as a whole. These values, reported in Table II, reflect the observation made in the training metrics, from which it is evident that all the models perform well for COVID-19 detection. From Fig. 4, it can be noted that DenseNet gives the highest accuracy and the lowest loss in terms of training metrics for all the datasets. However, this is not enough to conclusively say that the model is indeed better, as we need to compare the testing metrics of these models as well. Comparing these values while keeping the dataset constant shows us a clearer picture of the performance of each model.

The F1-Scores of each model, as seen from Table II show that the DenseNet model performed the best across all datasets. For this reason, we move forward by comparing the performance of each dataset within DenseNet. The training accuracy and loss after each epoch for each dataset with DenseNet is plotted in Fig. 5. The training accuracy is highest for the raw dataset, followed by bounding box dataset and finally the segmented dataset.

It is important to remember that the key to useful predictions

TABLE II. Performance Evaluation of the Deep Neural Models Using Standard Metrics

Model	Metric	Raw			Segmented			Bounding-Box		
		COV	NOR	PNEU	COV	NOR	PNEU	COV	NOR	PNEU
DenseNet	Precision	0.9909	0.8882	0.9693	1.0	0.8575	0.9687	0.9904	0.8575	0.9586
	Recall	0.9396	0.9274	0.9602	0.8965	0.9495	0.9438	0.8965	0.9116	0.9485
	F1-score	0.9646	0.9074	0.9647	0.9454	0.9011	0.9561	0.9411	0.8837	0.9535
	Accuracy		**0.9503**			**0.9409**			**0.9347**	
EfficientNet	Precision	1.0	0.8457	0.9630	0.9714	0.9029	0.9462	1.0	0.8479	0.9540
	Recall	0.8534	0.9337	0.9450	0.8793	0.8801	0.9672	0.8448	0.9148	0.9461
	F1-score	0.9209	0.8875	0.9539	0.9230	0.8913	0.9566	0.9158	0.8801	0.9500
	Accuracy		0.9340			0.9378			0.9293	
VGG-16	Precision	1.0	0.7883	.9725	0.9557	0.8463	0.9779	0.9814	0.8713	0.9413
	Recall	0.9310	0.9400	0.9122	0.9310	0.9558	0.9345	0.9137	0.8548	0.9567
	F1-score	0.9642	0.8575	0.9414	0.9432	0.8977	0.9557	0.9464	0.8630	0.9489
	Accuracy		0.9208			0.9394			0.9278	

is learning relevant features. Although the model trained on the raw dataset seems to be performing well, when we analyse the visualisation through GradCam, we can see that the model is learning irrelevant features outside of the lung. In Fig. 6, for Raw, the text on the top of the X-ray is where the network is most activated, whereas in the model that was trained on the bounding box dataset, the presence of ground glass opacities in the lung are picked up very well, as seen from the heatmap shown in Fig. 6 (third row, last image).

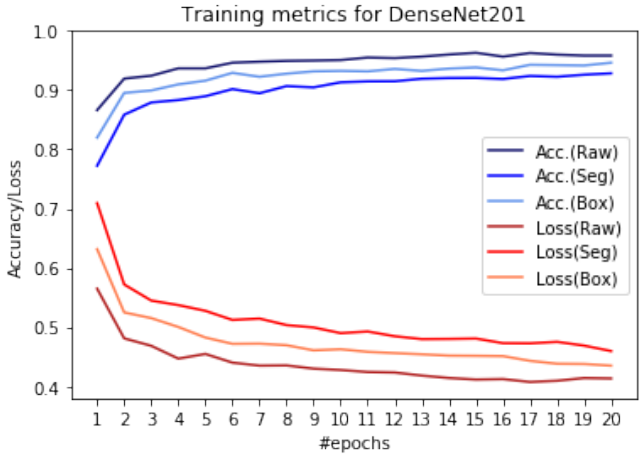

Fig. 5. Observed Accuracy and Loss performance for DenseNet

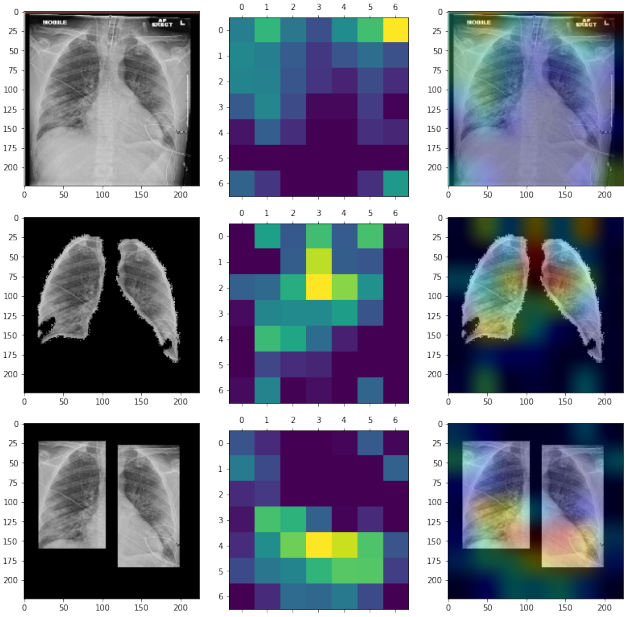

Fig. 6. Feature Learning of DenseNet for the (a) Raw dataset (b) Segmented dataset (c) Bounding Box dataset

V. Conclusion & Future Work

In this paper, an attempt to compare and benchmark the suitability of different preprocessing techniques on some state-of-the-art CNN models to classify lung diseases is presented. The problem of detecting COVID-19 from these X-rays and differentiating them from non-COVID-19 pneumonia is relatively new and often challenging. In the proposed work, different preprocessing methods were applied to the datasets, for training deep neural models, after which the classification performance was compared across all the models. Among the chosen models, DenseNet performed the best with best-in-class precision, f-score, and accuracy performance. Due to the limitation posed by dataset size and availability of COVID-19 X-rays, we have employed transfer learning to overcome this barrier and implement multi-class prediction with state of the art convolutional models. Visualisation techniques were incorporated for studying the bias introduced in the models due to extraneous factors like text on the X-ray etc. These experiments underscored the need for effective ROI identification for achieving best performance and also for scaling the performance for larger datasets. Our findings provide a basis for model selection for feature engineering, in cases where multimodal data/ensemble models needs to be used for classification. It can also be used to help select among different types of preprocessing techniques, depending on the use-case, which we intend to explore as part of future work. In addition to these models, other state-of-the-art models can be compared - after applying one of the proposed preprocessing techniques, by making more of their layers trainable to further improve performance.

References

[1] D. Willman, "Contamination at cdc lab delayed rollout of coronavirus tests," *The Washington Post*, Apr 2020. [Online]. Available: https://www.washingtonpost.com/

[2] W. Kong and P. P. Agarwal, "Chest imaging appearance of covid-19 infection," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, p. e200028, 2020. [Online]. Available: https://doi.org/10.1148/ryct.2020200028

[3] S. Bharati, P. Podder, and M. R. H. Mondal, "Hybrid deep learning for detecting lung diseases from x-ray images," *Informatics in Medicine Unlocked*, vol. 20, p. 100391, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352914820300290

[4] P. N. Kieu, H. S. Tran, T. H. Le, T. Le, and T. T. Nguyen, "Applying multi-cnns model for detecting abnormal problem on chest x-ray images," in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, 2018, pp. 300–305.

[5] Y. Wu, X. Qin, Y. Pan, and C. Yuan, "Convolution neural network based transfer learning for classification of flowers," in *2018 IEEE 3rd International Conference on Signal and Image Processing (ICSIP)*, 2018, pp. 562–566.

[6] M. Hon and N. M. Khan, "Towards alzheimer's disease classification through transfer learning," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017, pp. 1166–1169.

[7] M. F. Hashmi, S. Katiyar, A. G. Keskar, N. D. Bokde, and Z. W. Geem, "Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning," *Diagnostics (Basel)*, vol. 10, no. 6, Jun 2020.

[8] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Comput. Biol. Med.*, vol. 121, p. 103792, 06 2020.

[9] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2," *Inform Med Unlocked*, vol. 19, p. 100360, 2020.

[10] H. X. Bai, R. Wang, Z. Xiong *et al.*, "Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT," *Radiology*, vol. 296, no. 3, pp. E156–E165, 09 2020.

[11] A. Mangal, S. Kalia, H. Rajgopal, K. Rangarajan, V. Namboodiri, S. Banerjee, and C. Arora, "Covidaid: Covid-19 detection using chest x-ray," 2020.
[12] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, and M. Kaur, "Classification of the covid-19 infected patients using densenet201 based deep transfer learning," *Journal of Biomolecular Structure and Dynamics*, vol. 0, no. 0, pp. 1–8, 2020, pMID: 32619398.
[13] T. Zebin and S. Rezvy, "Covid-19 detection and disease progression visualization: Deep learning on chest x-rays for classification and coarse localization," *Applied Intelligence*, Sep 2020.
[14] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," https://github.com/ieee8023/covid-chestxray-dataset, 2020, accessed: 2020-10-01.
[15] D. S. Kermany, K. Zhang, and M. Goldbaum, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," 2018.
[16] L. Wang, A. Wong, Z. Q. Lin, P. McInnis, A. Chung, and H. Gunraj, "Covid-19 chest x-ray dataset initiative," https://github.com/agchung/Actualmed-COVID-chestxray-dataset, 2020.
[17] "Keras preprocessing," Nov 2020. [Online]. Available: https://github.com/keras-team/keras-preprocessing/blob/master/keras_preprocessing/image/image_data_generator.py
[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.
[19] S. K. Antani, "Tuberculosis chest x-ray image data sets — national library of medicine." [Online]. Available: https://lhncbc.nlm.nih.gov/publication/pub9931
[20] Eduardomineo, "U-net lung segmentation (montgomery shenzhen)," Oct 2018. [Online]. Available: https://www.kaggle.com/eduardomineo/u-net-lung-segmentation-montgomery-shenzhennotebook-container
[21] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020.
[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
[23] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016. [Online]. Available: http://arxiv.org/abs/1610.02391