COSC 2673/2793 - Assigment 4

Jigar Mangukiya(s3715807)

I. Introduction

Since Dec 2019, Corona Virus Infectious Diseases (Covid-19) pandemic has wreaked havoc globally by wide-spread infections and large number of deaths with no end in sight. While researchers around the world are looking for a vaccine, it has become equally important to be able to identify the infection in a patient as early as possible. As early detection of covid-19 infection can lead to quick isolation and ultimately helping in slowing down the spread of disease. Fair amount of attention has been given to AI based disease detection methods in the research community with multiple attempts at detecting the disease from medical images like x-ray & computed tomography.

The report outlines a proposal to use a machine learning approach to classify covid-19 infected patients, pneumonia (not caused by SARS-CoV-2) & healthy patients using chest X-ray. Report consists a critical literature review of published papers analyzing suitability of using medical images for identifying covid-19 or similar respiratory diseases, the approach taken by the authors, compares the result achieved, and challenges faced in developing such solutions. Second part of the report proposes an approach that uses AI algorithms to provide an end-to-end solution for using x-ray images in identifying covid-19, pneumonia & non-infected healthy patients.

II. Literature Review

Purpose of reviewing published literature is to gain an understanding of how researchers have approached the problem of using x-ray images of lungs to identify respiratory disease in patients, which methods were effective, what were the results and the challenges faced while creating the solution.

A. *Testing methods & significance of chest x-rays*

While the “gold standard” for diagnosing covid-19 infection is real-time reverse transcription polymerase chain reaction (RT-PCR) due to its high sensitivity & specificity, it is vastly resource consuming and takes long time [1]. However, covid-19 has observed exponential spreading rate through community transmissions in several countries and it is turning out to be challenging to perform RT-PCR tests in mass due to testing kit availability in countries with large population and poor health infrastructure. This prompts a requirement for having a method that can perform diagnosis quickly with readily available equipment. Also, studies [2], [3] have shown false-positive & false-negative deficiencies with RT-PCR test, with the concerning false-negatives being cases where the test could not diagnose the cases with existing clinical characteristics & radiographical evidences, With extreme case being when an isolated patient was tested negative for RT-PCR test 3 times for 3 weeks and was later tested positive [3]. The high false-negative rate can adversely affect the strategies of controlling the spread of the virus. A supplementary method, that can provide quicker diagnosis to support, flag RT-PCR test or can be used as a screening measure will help in controlling the devastation caused by this pandemic. Gross [4] revealed a clear evidence in radiographic in unenhanced chest computed tomography (CT) images, observing a patchy bilateral ground glass opacities (GGO) & consolidations in lungs. CT offers high sensitive and but it is less likely to be available in regions with poor health infrastructures due to high cost and operational expertise required. Chest x-ray (CXR) radiography, while being less sensitive than CT, is relatively cheaper and is vastly more available and used medical practice. It can serve well as the first-line of diagnosis and as a supplementary measure to other more sophisticated diagnosis methods like RT-PCR & next generation sequencing (NGS) if and when available. Bell [5] & Wong [8] have provided evidence that conventional CXRs of covid-19 were abnormal for at least 60%-80% patients at the time of hospitalization, which advocates the use of CXR as preliminary diagnosis methods. There are recorded use-cases of using portable x-ray equipment from behind a glass to reduce the risk of exposure to medical staff and requirement of personal protection equipment [5], [6]. [5], [6], [7] provides evidence of viability of using CXRs to diagnose covid-19.

B. *Application of AI/ML in detecting diseases from x-rays*

Over the years, there has been ample attention given to using machine learning to detect diseases from various types of medical images. There are mainly two generations of machine learning algorithms that were used to solve the problem of detecting the disease from radiography. Conventional ML classifiers like K-nearest neighbor classifier (KNN) & support vector machine (SVM) and artificial neural network-based classifiers like convolutional neural networks (CNN).

Traditional machine learning algorithms require a feature extraction framework to be employed to carve out important features from the images to be fed to classifier. It is problematic as the manual extraction of important features from the images with even the most sophisticated feature extraction process will still be lossy. Study [9] compares results from multiple published studies and states that the hand-crafted features are inherently less suitable for handling the variations in data like quality of images & demographics of patients.

Artificial neural network-based classifiers on the other hand do not require feature extraction but directly learn it from the images. In recent years, deep learning neural networks have been proven to be very efficient at detecting diseases from chest radiography. Study [10] in 2014, developed a deep CNN to detect lymph nodes in chest from radiography images which achieved (at the time of publishing) substantial results for the challenging problem. They used random rotations, translations and scaling to introduce variance in the dataset for robust model development. At the time of publishing the state-of-the-art methods had sensitivity of 52.9% at 3.1 FP/volume. The deep CNN proposed by [10] achieved a significantly better performance of 83% sensitivity (max) at 3.1 FP/volume compared to then state-of-the-art.

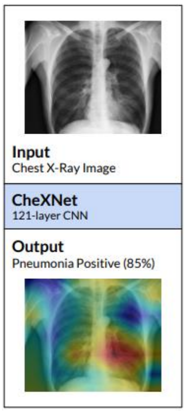

Fig. 1. CheXNet detecting pneumonia and outputing heatmap with indicative regions. Source – [11]

This exhibited the capabilities of using deep learning networks in detecting diseases from CSRs.

Several subsequent studies achieved even better performance on much larger datasets. A particularly significant study was [11]. It proposed a novel algorithm, named CheXNet, which was 121 layers deep CNN trained on Chest-X-ray-14 to detect pneumonia from CXRs. Research used over 100,000 examples with 14 chest/lungs disease labels to train, validate and test the performance of the algorithm. Significant aspect about [11] was they acquired a annotated test set from 4 practicing (academic) radiologists and found that model achieved higher F1 score than average radiologist. Apart from providing the probability of pneumonia, model also outputted a heatmap marking the area of image that were indicative of pneumonia. See Figure 1 for example. Their model used DenseNet architecture [12], using weights from a pre-trained ImageNet model [13]. While training, the authors have used image normalization (centering) and horizontal augmentation to make the model robust against human-variance introduced while images were captured.

The substantial results achieved by study [11] for identifying pneumonia using a deep CNN network paves way for the hypothesis of usability of AI/ML frameworks to identify covid-19 SARS-Cov-2. Both pneumonia and covid-19 cause somewhat similar abnormalities in CXRs like Ground-glass opacities (GGO) of various forms & Bronchial and Vascular wall thickening [14]-[16]. However, crazy paving, & consolidation is vastly more apparent in covid-19 infection while pneumonia infections shows reticular opacities [14]-[16]. These subtle differences can allow classification of CXRs having no-infection (no abnormalities), covid-19 (vastly spread GGO and consolidation) & pneumonia (unilateral GGO & reticular opacities) using deep learning CNNs due to their ability of identifying low level features like positions of GGO & consolidation in the image.

The problem of interest for this proposal is AI/ML based approach to diagnose covid-19 and pneumonia from a dataset of CXRs. For further analysis of deep learning models for multi-class classification [16] proposes a deep learning CNN based on, DarkNet-19 [17] model, named DarkCovidNet. DarkNet-19 [17] is considered state-of-the-art deep learning model that can identify more than 9000 different categories in real-time. Study [16] uses DarkCovidNet in two configurations, the binary classification configuration, classifying in classes “covid-19 infected” and “non-infected” and multi-class configuration, classifying “covid-19”, “pneumonia” & “non-infected” classes. Proposed model in [16] has 17 convolution layers with batch normalization & LeakyReLU operations after each layer, respectively for, increasing training efficiency and preventing neurons from dying. Dying neurons is a major problem in moderate to very deep neural networks with large filters. The proposed model had a healthy 1.164 million parameters. Researchers use two different datasets to create dataset for multi-class classifier, which consisted 127 CXR images of diagnosed covid-19 patients [18], 500 no-finding & 500 pneumonia CXRs from [19]. Dataset used here is unbalanced, which can be due to recency of disease. A 5-fold cross-validation was performed to validate the results and prevent over-fitting. Overall, for 3 class classification, model acquired a healthy average of (85.35%, 92.18%) for sensitivity & specificity respectively. Which can be considered in line with sensitivity of methods like PCR & NGS. However, researchers have not performed any augmentations in images, which could have presented a more realistic scenario where CXRs taken by different equipment and by different operators will have certain variations, presenting a more challenging query to the model. Augmentation would have also been useful to nullify the imbalanced nature of dataset.

Several other studies were conducted based on using pre-trained model as part of transfer learning, which gave very competitive results for certain models [19], [20]. Studies [19], [20] evaluates the performance of various transfer-learning models which are considered state-of-the-art in object detection and image classification task. Study [19] used intensity normalization as well as Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve the input data quality. Study [20] used manual rescaling of images. Both studies have noted mentions of imbalance in covid-19 infection class. Neither studies used any resampling methods to rectify the imbalance as well as no data augmentation techniques were used to make the model robust. Table 1 mentions the top performing models from [19] & [20].

As can be seen from table, performance of multitude of models are closely similar, with mobile net v2 achieving best results. A good similarity in high performance models and the criticality of false negatives in the spread of disease pushes the proposal in direction of using a model ensemble to make predictions.

C. *Ethical considerations*

Any automated AI system for medical diagnosis project should, very seriously, make ethical considerations due to nature of impact of the system on human lives [22]. Data privacy & anonymization of datasets should be ensured. Authors in [22] argues that since CAD (computer aided decisions) systems to AI systems, it is always preferred to keep humans in loop, to ensure certain sample bias & algorithm bias is not affecting the patient’s treatment.

TABLE 1- Performance of different pre-trained models in [19],[20]

Ref.	**Pre-trained Model**	**Acc.**	**Sensi.**	**Specifi.**
[19]	Inception_Resnet_V2	92.18	92.11	96.06
[19]	DensNet201	88.09	87.99	94.00
[20]	MobileNet V2	94.72	98.66	96.46
[20]	VGG19	93.48	92.85	98.75

III. Project Proposal

From the extensive literature review, some of which was not included in previous section due to size limitations, it is evident that conventional machine learning algorithms like KNN & SVM are not the most effective algorithms to be used with medical image data. State-of-the-art models are mostly using pre-trained CNN models to classify the data. It was also evident that the many of the pre-trained algorithms like Inception Resnet V2, MobineNet V2 & VGG 19 have high sensitivity and specificity. Considering the cost of false negatives are too high in given scenario, with potential of starting new clusters of disease spread in community, it is not the most ideal to use prediction evidence from a single model to make a diagnosis or even provide supplementary diagnosis. These pointers have guided this proposal to use a ensemble of pre-trained CNN models to make multi-class classification of CXRs. The proposal will walk through detailed steps to train such an ensemble and guidelines on how deploying along with ethical consideration wherever deemed necessary.

A. *Problem Statement*

Constructing a ML model to classify chest x-ray images into three classes, covid-19 infected, pneumonia infected & non-infected.

B. *Dataset*

It is assumed in the problem description that the project will use a dataset that has 200 images for covid-19 infected patients and 300 images each for rest of the two classes. Dataset is provided by a public hospital in Victoria. Here there is chance of induction of selection bias. Considering covid-19 is global pandemic with virtually every country is affected with it, and assuming that the model developed here will be used regardless of any geographic information in query, a more complete view of population would've been a sample acquired from different geographies. Avoiding selection bias should be an ethical consideration for the project but with the limited scope of this proposal, it will be left with this note.

C. *Date preprocessing & Sampling*

Initially a stratified sample of 80:20 can be acquired for primary train test split. A stratified sample will ensure that all classes are represented in both train and test split in same proportion as they were in population. The project will be using k-fold cross validation so a separate validation set will not be required.

The training sample will be relatively small and slightly imbalanced with 2:3:3 class distribution. To increase the size as well as balancing out the slight imbalance, new images should be generated using data augmentation techniques. The augmentations that can be used are contrast modification, Gaussian noise & Gaussian blur, which can be related to different types of equipment that produce different quality CXRs. Apart from these jittering, rotations and shears can be used for generating additional images. Do note that these are additive data augmentation that adds new images in training set. A further round of augmentation will be used while training the models using generators for controlling the over fitting of model and adding robustness.

D. *Model descriptions*

The project will use Inception Resnet V2 & MobileNet V2 pre-trained models, as they are the best performing models with high sensitivity & Specificity. This section will give a brief about each model. Full description of both models will be outside the scope of this proposal hence, the proposal will outline brief descriptions and move on to implementation.

1. *Inception resnet v2*
   Model is formed combining two models, inception v4 & Residual networks [23]. The network has 164 layers and is trained with more than a million images with classification ability of 1000 classes. Pre-trained model offers rich feature representation for wide range of images. Input image size is fixed at 299*299 [23].
2. *MobileNet v2*
   MobileNet v2 is state-of-the-art mobile based image classification, object detection & textual semantic segmentation model [24]. The model is 53 layers deep and is trained on images from ImageNet [25]. It uses depthwise separable convolutions layers. Input size is fixed at 224 * 224. Model is pre-trained on

E. *Model Hyperparameter Tuning & Freezing*

For tuning keras ImageGenerators with augmentation should be used. This will result in a more efficient loading of images as Generators load the images in batches instead of loading them in memory all at once.

For both of these models are pre-trained with parameter weights pre-loaded in all distributions. There is a major distinction in depths of models which will make feature representation pretty discrete. Output layer for both models will be replaced with 3 outputs and a softmax activation. This activation results in probability division across the classes that sums to 1, which is a convenient way for making multiclass classification. Adam optimizer should be used, which can be considered as a combination of stochastic gradient descent (SGD) & RMSProp optimizers providing more finer control over learning rate using auxiliary parameters beta1 & 2.

The model themselves don't require hyperparameter tuning, as they are tuned on much larger ImageNet dataset which is much more versatile then the small dataset this project will be working with. Therefor it is recommended to freeze the bottom layers. Freezing the bottom layers will ensure that the model doesn't overfit the new dataset.

For tuning the hyperparameters like epochs, learning rate, momentum and decay can be tuned using methods like GridSearchCV or RandomSearchCV. Both methods take as input a parameter grid. They differ in terms of how they search the permuted parameter space. GridSearchCV will iterate through every possible permutation of parameters while RandomSearchCV will take an extra argument specifying how many iterations to be tested. It will then, randomly select that many parameter combinations and run the training & validation cycles to identify best set of parameters. Both methods perform cross validations and the number should be kept from around 3 to 5 for balancing the protection against overfitting and unnecessary iterations. For the most part GridSearchCV becomes impractical as it takes much more time with exponentially large parameter space. RandomSearchCV makes a dumb search on the parameter space and can miss the best choice. The ideal case will be to use GridSearchCV with small number of sets of fewer parameter combinations. This will convert the exponential parameter combination space into linear one. Using evidence from multiple iterations of small grid searches, best parameters can be identified, and model can be re-trained using the best parameters.

F. *Ensemble & evaluation*

Choosing right metrics evaluation is a critical thing for any project, especially for medical projects where the consequences of a bad prediction can be devastating. For the theoretical evaluation and comparing models etc. multiple parameters should be used. Here in this scenario, False negatives are a major problem. False negative cases are which termed as non-infected while in fact they were covid-19 infected. Consequences of such incorrect diagnosis can be catastrophic in high population density areas. Hence, instead of relying on a simple measure, evaluation should be performed using multiple metrics as it often is. The most important metrics here will be sensitivity & specificity. Sensitivity is a measure of how many infected patients are identified and specificity is a measure of how many actual negatives are identified from total negatives. This metrics pair should be supplemented with accuracy, that is number of correct predictions divided by total observations & F1 Score. F1 score is balance between precision and recall.

Once both models are trained with best parameters and the performance should be tested on test set. Here it should be done in 2 ways. Individual performance of each model on testing set should be evaluated using the metrics discussed earlier to see if one model performs significantly worse than the other. If that is the case, a decision can be made to either try to retrain the model or the drop the model from the project, which might sound a lot of work in vain, but is a real possibility here. If a model is dropped then the other model becomes primary model and predictions from that models will be the final predictions.

However, if both models perform somewhat similar, a model ensemble, can be created manually. But instead of using default bagging ensembles which take classification vote, here normalized probabilistic output from both models can be used to specify the prediction confidence.

G. *Conclusion*

It is evident from the study that there is indeed a need for quicker and reliable diagnosis method that can help either in screening and supplementary measure to doctors to reduce their workload and help better manage the fight against pandemic covid-19. AI based approaches are latest pinnacle in technology and can be leveraged to work as such a testing/screening method. However, any method that involves automated medical decision making, should be thoroughly scrutinized by experts, doctors & researchers. This proposal outlines one such method that holds potential based on the success of the previous works done by researchers. The ensemble approach can be grown with addition of multiple models and larger more representative datasets and can serve as first line for covid-19 diagnosis.

IV. References

[1] A. Tahamtan and A. Ardebili, “Real-time RT-PCR in COVID-19 detection: issues affecting the results,” Expert Review of Molecular Diagnostics, vol. 20, no. 5, pp. 453–454, Apr. 2020.

[2] Y. Wang, H. Kang, X. Liu, and Z. Tong, “Combination of RT‐qPCR testing and clinical features for diagnosis of COVID‐19 facilitates management of SARS‐CoV‐2 outbreak,” Journal of Medical Virology, vol. 92, no. 6, pp. 538–539, Mar. 2020.

[3] “Real-time RT-PCR in COVID-19 detection: issues affecting the results,” Expert Review of Molecular Diagnostics, 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/14737159.2020.1757437. [Accessed: 10-Jun-2020].

[4] A. Gross, D. Thiemig, F.-W. Koch, M. Schwarz, S. Gläser, and T. Albrecht, “CT appearance of severe, laboratory-proven coronavirus disease 2019 (COVID-19) in a Caucasian patient in Berlin, Germany,” RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren, vol. 192, no. 05, pp. 476–477, Mar. 2020.

[5] D. J. Bell, “COVID-19 Radiology Reference | Radiopaedia,” Radiopaedia.org, 2019. [Online]. Available: https://radiopaedia.org/articles/covid-19-3. [Accessed: 11-Jun-2020].

[6] “Policies and Guidelines for COVID-19 Preparedness: Experiences from the University of Washington | Radiology,” Radiology, 2019. [Online]. Available: https://pubs.rsna.org/doi/10.1148/radiol.2020201326. [Accessed: 11-Jun-2020].

[7] S. Ianniello, C. L. Piccolo, G. L. Buquicchio, M. Trinci, and V. Miele, “First-line diagnosis of paediatric pneumonia in emergency: lung ultrasound (LUS) in addition to chest-X-ray (CXR) and its role in follow-up,” The British Journal of Radiology, vol. 89, no. 1061, p. 20150998, May 2016.

[8] H. Y. F. Wong et al., “Frequency and Distribution of Chest Radiographic Findings in COVID-19 Positive Patients,” Radiology, p. 201160, Mar. 2019.

[9] C. Qin, D. Yao, Y. Shi, and Z. Song, “Computer-aided detection in chest radiography based on artificial intelligence: a survey,” BioMedical Engineering OnLine, vol. 17, no. 1, Aug. 2018.

[10] H. R. Roth et al., “A New 2.5D Representation for Lymph Node Detection Using Random Sets of Deep Convolutional Neural Network Observations,” Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, pp. 520–527, 2014.

[11] P. Rajpurkar et al., “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” arXiv.org, 2017. [Online]. Available: https://arxiv.org/abs/1711.05225. [Accessed: 12-Jun-2020].

[12] G. Huang, Z. Liu, van, and Weinberger, Kilian Q, “Densely Connected Convolutional Networks,” arXiv.org, 2016. [Online]. Available: https://arxiv.org/abs/1608.06993. [Accessed: 12-Jun-2020].

[13] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[14] H. Shi et al., “Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study,” The Lancet Infectious Diseases, vol. 20, no. 4, pp. 425–434, Apr. 2020.

[15] H. X. Bai et al., “Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT,” Radiology, p. 200823, Mar. 2020.

[16] 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,” Computers in Biology and Medicine, vol. 121, p. 103792, Jun. 2020.

[17] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 6517-6525, doi: 10.1109/CVPR.2017.690.

[18] Cohen J.P. 2020. COVID-19 Image Data Collection.https://github.com/ieee8023/COVID-chestxray-dataset

[19] Wang X., Peng Y., Lu L., Lu Z., Bagheri M., Summers R.M. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases; pp. 2097–2106.

[20] K. El Asnaoui and Y. Chawki, “Using X-ray images and deep learning for automated detection of coronavirus disease,” Journal of Biomolecular Structure and Dynamics, pp. 1–12, May 2020.

[21] I. D. Apostolopoulos and T. A. Mpesiana, “Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks,” Physical and Engineering Sciences in Medicine, vol. 43, no. 2, pp. 635–640, Apr. 2020.

[22] “Ethics of Artificial Intelligence in Radiology: Summary of the Joint European and North American Multisociety Statement Radiology,” Radiology, 2019. [Online]. Available: https://pubs.rsna.org/doi/10.1148/radiol.2019191586
K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778.

[23] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” arXiv.org, 2016. [Online]. Available: https://arxiv.org/abs/1602.07261. [Accessed: 12-Jun-2020].

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” 2019.

[25] “ImageNet,” Image-net.org, 2017. [Online]. Available: http://www.image-net.org/. [Accessed: 12-Jun-2020].