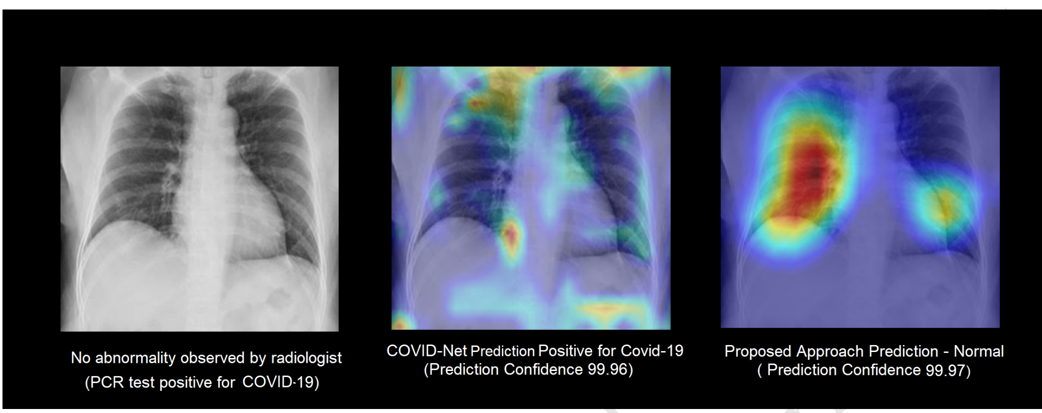
Towards Building Responsible Artificial Intelligence Systems: The Role of Chest X-Ray imaging in Patient Management during the COVID-19 Pandemic

ANURAG PALKAR, Persistent Systems Ltd.
ARNAV JHINGRAN, Persistent Systems Inc.
CHINMAY SAVADIKAR, Persistent Systems Ltd.
RAHUL KULHALLI, Persistent Systems Ltd.
BHUSHAN GARWARE, PH.D., Persistent Systems Ltd.
ARUN JAMKAR, M.B.B.S., M.S., PH.D., FICS, FMAS, FIAGES, FAIMER, Persistent Systems Ltd.



Ever since the onset of the COVID-19 pandemic, many countries are faced with the challenge of controlling large scale community spread while minimising the impact on their respective economies. Since the PCR test is the only reliable diagnostic test for the diagnosis of COVID-19, the lack of the availability and long turnaround time of these tests poses a need for efficient use of the same. Chest X-Rays, which are a comparatively cheap and widely available resource, can diagnose chest abnormalities like Pneumonia, which *may* indicate the presence of COVID-19. Leveraging Chest X-Rays might pave the path in reducing the community spread of the disease by effectively isolating patients with Pneumonia-like symptoms to be later tested using regular COVID-19 diagnostic tools. In this work we propose a two step process to effectively screen patients using AI-based automated tools for diagnosing Pneumonia from chest X-Rays. Further, we show that explanations for models trained to classify COVID-19 just from Chest X-Ray images are not consistent across images, which indicates that building such models may not be reliable to use in practice.

CCS Concepts: • **Computing methodologies** → **Neural networks**; *Supervised learning*; **Computer vision**; • **Applied computing** → *Life and medical sciences.*

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Additional Key Words and Phrases: COVID-19, neural networks, management system, explainable AI

ACM Reference Format:
Anurag Palkar, Arnav Jhingran, Chinmay Savadikar, Rahul Kulhalli, Bhushan Garware, Ph.D., and Arun Jamkar, M.B.B.S., M.S., Ph.D., FICS, FMAS, FIAGES, FAIMER. 2018. Towards Building Responsible Artificial Intelligence Systems: The Role of Chest X-Ray imaging in Patient Management during the COVID-19 Pandemic. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Mysterious deaths of a few people in Wuhan, Hubei, a province of China, due to an unknown respiratory disorder akin to Pneumonia late in December 2019 rapidly turned into a global pandemic disease named as COVID-19. As per World Health Organization (WHO) [10] as of today, over 13 million people around the world have been diagnosed with COVID-19, claiming lives of more than 585000 individuals. Controlling large scale community spread while minimising the impact on economy is the biggest challenge that most of the countries are currently facing. Artificial Intelligence (AI) plays a key role in understating and addressing crises for global surveillance to mitigate risk of infectious disease such as early detection, measuring dispersion, anticipating disruption, drug discovery, intelligent diagnosis, virtual care and vaccine development.

1.1 Responsible AI

There is a rapid technical progress and a widespread adoption of Artificial Intelligence (AI) based products and workflows influencing many aspects of human life, especially in healthcare. Although the accuracy of AI models is undoubtedly the most important factor considered while deploying AI-based products, there is an urgent need to understand how AI can be designed to operate responsibly. Responsible AI is a framework that each software developing organization needs to adopt in order to build customer trust through the transparency, accountability, and fairness of deployed AI solutions. The principles of *Responsible AI* are well defined by Arrieta et at [2] and Dignum [7]. In this paper, we present an in-depth analysis of AI-enabled automated detection models available in the literature from the perspective of *Responsible AI* and propose a novel patient management approach for leveraging AI technology for automated analysis of X-ray imaging responsibly. One of the open research challenges is the non-availability of a metric to measure the quality of explanations[5]. The prime goal of AI enabled CXR diagnosis is to assist the radiologist by automating the most obvious decisions, allowing them to focus on critical cases where expert's intervention is needed. In this work we keep expert radiologists in the loop to make sure that the AI model has learned the right features while minimising bias.

1.2 Importance of Chest X-ray (CXR) imaging

COVID-19 pneumonia manifests with chest CT imaging abnormalities, even in asymptomatic patients [17]. Comparatively, the sensitivity of CT is more relevant in the setting of a public health approach that requires isolation of all infected patients where the availability of COVID-19 testing is limited and turnaround times are long [15]. However, there are challenges of widespread use of CTs due to higher risk of COVID-19 infection transmission along the route to a CT scanner room, particularly in environments lacking PPE (Personal Protective Equipment). Even though the sensitivity of a CXR is lesser compared to CTs, there are some advantages which makes CXR-based diagnosis one of the critical steps in patient management during crisis[14]. One of the biggest advantage of CXR is Equipment Portability, with imaging performed within an infected patient's isolation room or in ICU. In hospitalized patients, CXR can be useful for assessing disease progression and alternative diagnoses. Adam Jacobi et al. [8] highlighted the role of portable CXR due to widespread availability and infection control issues that currently limit the CT utilization. Salehi et al. [26] found that mobile X-rays provide adequate image quality for diagnosing pneumonia and it is a simple but reliable alternative in critically ill patients who cannot not undergo chest CT.

2 PRIOR WORK

Non-availability of publicly accessible CXRs of COVID-19 infected patients is one of the biggest challenge towards the development of a responsible AI-based diagnostic solution. Wong et al. [24] proposed computer-aided Covid-19 severity assessment system using deep neural network for predicting the geographic extent and opacity extent scores. The ground truth scores are generated for 130 CXRs of SARS-CoV-2 positive patient cases from the dataset curated by Cohen et al [6], which were annotated by expert radiologists. They report an optimum R2 of 0.865 and 0.746 between predicted scores and radiologist scores for geographic extent and opacity extent respectively, with minor modification of COVID-Net architecture [23]. However, the predicted scores are not explainable.

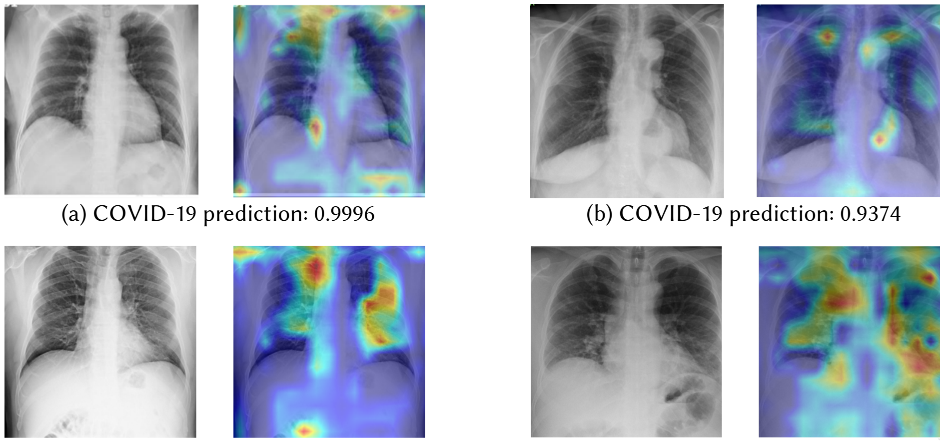
Ozturk et al. [11] use a dataset curated by Cohen et al. and design the DarkCovidNet model, which draws inspiration from the Darknet-19 [13] model. They train two variants of the model - a binary classification variant framed as COVID-19 vs. No-Findings, and a multi-class classification variant framed as COVID-19 vs. No-Findings vs. Pneumonia - using 5-fold cross-validation. For the three-class-classifier, they obtain an average sensitivity, specificity, and F1-score of 85.35%, 92.18%, and 87.37% respectively. For the binary classifier, they obtain an average sensitivity, specificity, and F1-score values of 95.13%, 95.3%, and 96.51%, respectively. They also generate heatmaps and evaluate the model outputs via an expert radiologist.

2.1 COVID-Net

There are several recently-published bodies of work which attempt to classify COVID-19 from XRay images. We refer to a representative paper – COVID-Net [23]. This model is trained on an open-access benchmark dataset called COVIDx which the authors have developed by aggregating data from multiple sources. The complete dataset contains 13,975 CXR images across 13,870 patient cases. The authors have also open-sourced several variants of their model. COVID-Net is a domain-specific model that is modeled using generative synthesis [25]. The model is designed to make one of the following three predictions: A) no infection (normal), B) non-COVID-19 infection (e.g., viral, bacterial, etc.), and C) COVID-19 viral infection. The resultant model architecture uses selective long-range connections and 'PEPX' modules, where each module consists of multiple 1x1 convolutions in conjunction with a depth-wise convolution so as to achieve high representational capacity and improve ease of training while still maintaining computational and memory efficiency. The authors have also provided visual explanations using GSInquire [12, 25].

Of the several model variants, we choose COVIDNet-CXR4-C because of its high sensitivity (96%) towards the COVID-19 class The goal of the current work is not to have a direct comparison of our model with COVID-Net because we propose only two classes – Normal and Abnormal. Instead, using COVID-Net, we show that the explanations, and in some cases, the predictions for the COVID-19 class based on only CXR-images are not consistent with a medical expert's opinion about the image. To demonstrate our hypothesis, we generate explanations for COVID-Net and compare them with those generated by our model. The authors, however, have not open-sourced GSInquire.

Furthermore, Adebayo et al. [1] show that saliency methods (eg. Integrated Gradients [21], Gradient*Input [18], Guided Backprop [20], Smoothgrad [19]) do not necessarily reflect the model parameters or the data generating process. Hence, we use GradCAM [16], a well-known technique for generating our explanations.



(a) COVID-19 prediction: 0.9996

(b) COVID-19 prediction: 0.9374

(c) COVID-19 prediction: 0.8592

(d) COVID-19 prediction: 0.999

Fig. 1. Excerpts of explanations obtained from COVID-Net, where the radiologist diagnosed the CXR to be normal

Figures 1a, 1b, 1c, and 1d represent the scenarios where the ground-truth label was COVID-19, the COVID-Net model predicted the images to be COVID-19, but the radiologist diagnosed the images as being normal. The heatmaps focus on portions outside the lung area, which may indicate that there may be some bias captured from the data. Also, it is difficult to find similar and consistent features being highlighted via heatmaps across all the images predicted as COVID-19 by the COVID-Net model. Hence, some scepticism about models which classify CXR images as COVID-19 seems reasonable.

3 PROBLEM FORMULATION

As per the interpretations shown in Figure 1, it is difficult to justify the reason behind diagnosing the CXR images as the COVID-19 cases. It is almost impossible, even for experienced radiologists, to differentiate between Pneumonia (bacterial or viral) and COVID-19 by just looking at a single CXR image. As shown in the Fig 2, we formulate the role of chest X-ray images in the AI-enabled COVID-19 patient management framework with a diagnosis model and a prognosis model. At the community screening level, it is important to separate-out the suspected individuals. At this step, it is sufficient to determine whether the CXR is normal or whether it shows Pneumonia-like symptoms. Only those individuals who show Pneumonia-like symptoms need to be isolated for observation and/or undergo the PCR test. If the patient's PCR test turns out to be positive, the patient should undergo the required treatment for COVID-19. Here, AI can help tackle the next challenge of forecasting need for intensive care like oxygen support or mechanical ventilators. The details about the implementation of the diagnosis model are given in the next section.

4 DIAGNOSIS MODEL

The intention of the proposed model is to detect abnormal CXR images from the normal ones, which is further narrowed down to the detection of Pneumonia (viral, bacterial, etc.). The model is trained on Pneumonia and Normal Chest X-ray images; no COVID-19 samples were included in the training. It aims at detecting pneumonia characteristics on frontal CXRs. COVID-19 samples were further checked for Pneumonia.

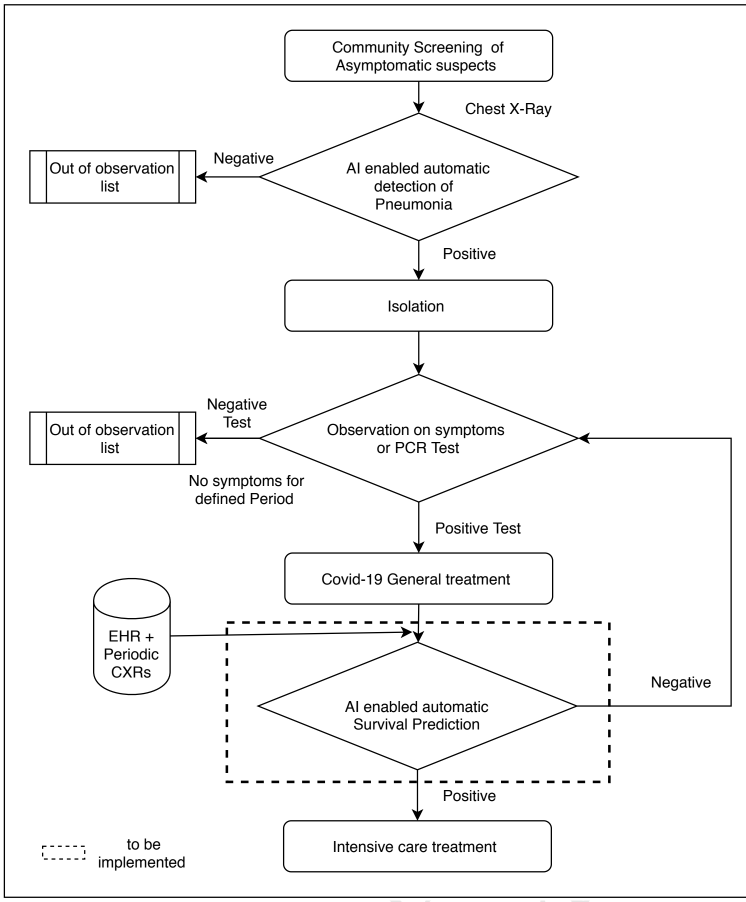


Fig. 2. Proposed role of AI enabled Chest X-ray imaging for COVID-19 patient management

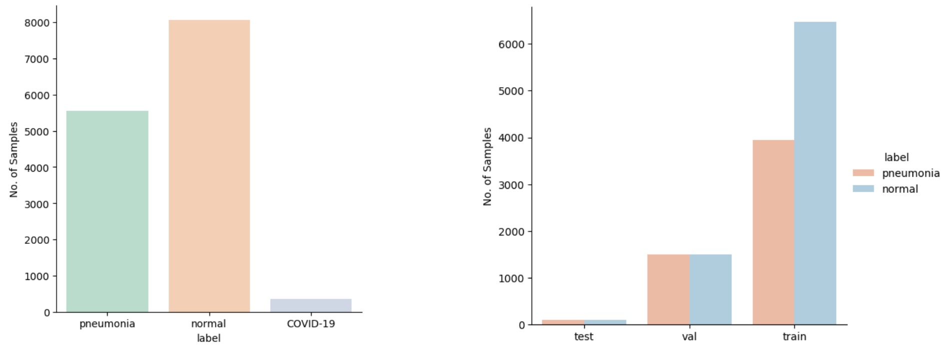
4.1 Dataset Details

The dataset we have used to train the model is same as used to train and evaluate COVID-Net, called as COVIDx dataset which comprises of 5 different publicly available data repositories 1) COVID-19 Image Data Collection, 2) Figure 1 COVID-19 Chest X-ray Dataset Initiative, 3) ActualMed COVID-19 Chest X-ray Dataset Initiative, 4) RSNA Pneumonia Detection Challenge dataset, which used publicly available CXR data, and 5) COVID-19 radiography database [23]. Total available CXR images were 13970, of which 5551 were pneumonia, 8066 were normal and 353 were COVID-19 as in the Fig 3a.

Our model was trained for binary classification of Pneumonia versus Normal cases. The training set comprised of 3951 pneumonia and 6466 normal cases, validation set of 1500 pneumonia cases and 1500 normal cases and test set of 100 pneumonia and 100 normal samples as shown in the Fig 3b.

4.2 Architecture

The idea of the combination of lower parameter count and additional regularization with batch-normalized auxiliary classifiers and label-smoothing allows for training high quality networks



(a) label types (b) train-test-val split

Fig. 3. Distribution of CXR images

on relatively modest sized training sets, inspired us to use the InceptionV3 model, comprising of Inception modules[22]. The architecture aims to use multiple smaller models which are then integrated to get required features. The model was built and evaluated using the Keras deep learning library with a TensorFlow backend. The model was then trained by replacing the last layer of the model for 2 class classification with pooling followed by softmax layer. The other implementation details include Adam optimizer with initial learning rate of 2e-4, categorical crossentropy to monitor the learning of the model.

4.3 Quality and Accuracy

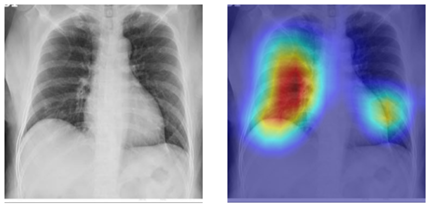
To analyse the performance of the model, sensitivity and positive predictive value (PPV) were calculated along with the test accuracy. The test accuracy for the model was 94.5% with normal sensitivity = 0.94 and pneumonia sensitivity = 0.95. The PPV for normal was 0.95 and for pneumonia was 0.94. The model was also tested for COVID-19 cases which were 353 in total, of these 304 were predicted as pneumonia and 49 as normal having sensitivity = 0.86.

4.4 Trust and Explanations

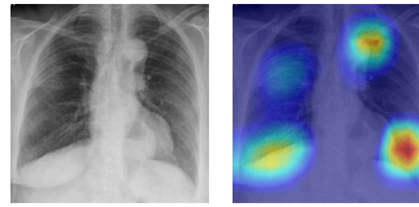
Our model classifies images 4a, 4b, 4c as normal and 4d as pneumonia. The heatmaps in figures 4a and 4c align with the region which a radiologist looks at to diagnose as a normal image. The heatmap in figure 4b place some importance on the right lung. However, some weight is also placed in the top-left part of the lung, which is clinically irrelevant. Our model classifies figure 4d as pneumonia and the heatmaps are more localized in the areas which may be potential indicators of an abnormality.

5 PROGNOSIS MODEL

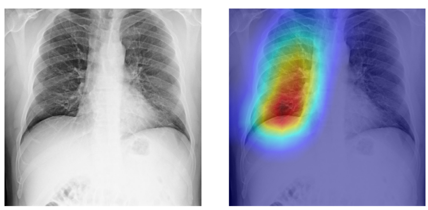
Radiologists use the Brixia score[4][3], a chest X-ray scoring system for COVID-19 to predict patient in-hospital mortality and disease severity. In the Prognosis model, we propose to use CXRs of an already diagnosed patient to determine the patient survivability or need for intensive care. Radiologists consider CXRs to be highly indicative of the progression of the disease and its impact on lung function.



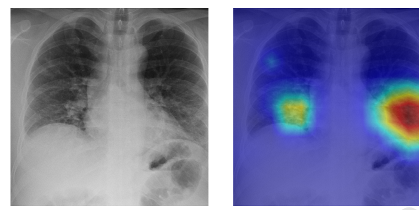
(a) Prediction: Normal (0.9974)



(b) Prediction: Normal (0.9910)



(c) Prediction: Normal (0.9902)



(d) Prediction: Pneumonia (0.9999)

Fig. 4. Figure

5.1 Dataset details

Cohen et al[6] shared CXRs from multiple publicly available sources along with clinical notes on survivability. The dataset includes 811 CXRs from hospitals in over 26 countries with 540 images labeled as COVID-19. Each image includes the patient metadata and the prognosis labels and notes which are used to determine the survivability label.

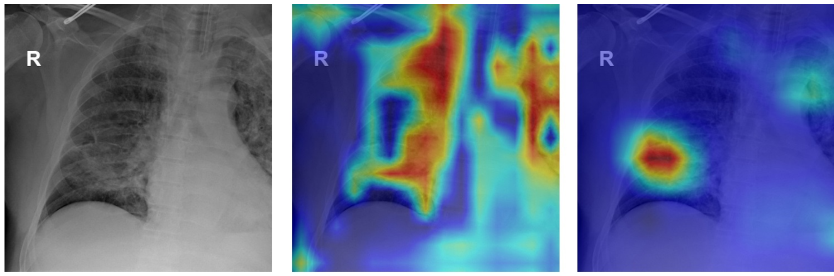
5.2 Model Architecture

A CNN model was built to provide a binary classification to predict survivability (1 = survive and 0 = not survive). The dataset was further augmented with a flip transform of the images of the original dataset to increase the dataset size for training. The model consists of 2 convolution layers followed by 2 fully connected linear layers. Each convolutional layer consists of a 2D convolution, followed by a Batch Normalization (BN) step, a rectified linear activation (ReLU) filter and a maxpool layer. The result of the convolution steps consist of 64k features. The final 2 layers of the model consist of a fully connected linear network followed by a binary output classification step. the false positive rate (predicting survival when the label was not survive) to be 0% and false negative rate to be 11%, hence maintaining high specificity.

5.3 Reliability and Trust

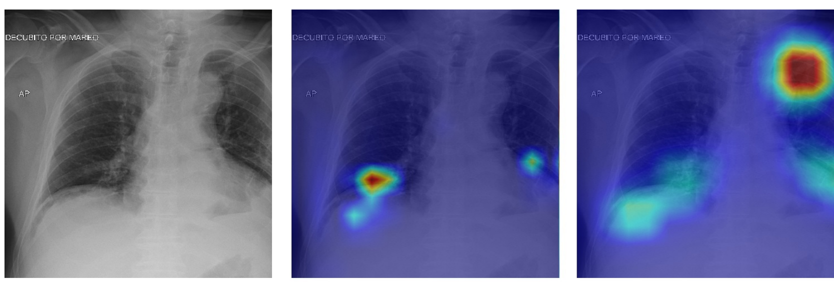
Even though the accuracy of the proposed prognosis model is high we do not claim that it passes the principles of *Responsible AI*. In reality, there are many factors like breathing hold time, oxygen concentration etc. that play a crucial role in determining the survivability. The dataset shared by Cohen et al[6] has lot of sparsity for the clinical measures mentioned above. Hence, in order to develop reliable and trusted prognosis model we need to have access to the detailed information like electronic health record (EHR) about the patient data as shown in 2.

A series of chest films of admitted patients are shared by Dr. Edgar Lorente [9]. If such information is made available, a reliable and trustworthy model can be developed. This has been left for future scope due to non-availability of publicly available data.



(a) Input Image (b) Heatmap for COVID-Net (Prediction: COVID19) (c) Heatmap for our model (Prediction: Pneumonia)

Fig. 5. Explanations generated by COVID-Net (b) are clinically more relevant



(a) Input Image (b) Heatmap for COVID-Net (Prediction: Normal) (c) Heatmap for our model (Prediction: Pneumonia)

Fig. 6. Heatmap generated by or model (c) shows high value in the top-left part of the lung, which is clinically irrelevant

6 DISCUSSION

There are some instances where the explanations generated by COVID-Net are more clinically relevant than the ones generated by our models. For example, in Figure 5, the heatmap generated by COVID-Net for class COVID-19 aligns better with the clinical features. Although our model classifies it as pneumonia, its heatmap focuses on a subset of the COVID-Net heatmap. Furthermore, Figure 6 shows that our model also focuses on to some irrelevant features (top lobe of the left lung), possibly indicating that our model might also be basing its output on some bias in data. Such discrepancies may me resolved by using multiple models for making a diagnosis and using the most trustworthy ones.

7 CONCLUSION AND FUTURE WORK

In this paper, we present an in-depth analysis of AI-enabled automated detection models available in the literature from the perspective of *Responsible AI* We focus on human interpretable explanations and conclude that even though a model may show very good predictive performance on small datasets, it may not be trustworthy. AI based systems should be used responsibly by formulating problems by consulting experts. In the future work, we aim to test the models proposed above on larger datasets and test the accuracy, interpretability, and interoperability across the images captured from different X-Ray scanning devices.

8 ACKNOWLEDGEMENTS

We would like to thank Dr. Abhijit Pawar (MBBS, DNB, DMRE; Professor, Department of Radiology, SKN Medical College, Pune) for his invaluable contribution in helping diagnose the CXRs and providing his valuable insights.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[3] Andrea Borghesi, Angelo Zigliani, Salvatore Golemi, Nicola Carapella, Patrizia Maculotti, Davide Farina, and Roberto Maroldi. Chest x-ray severity index as a predictor of in-hospital mortality in coronavirus disease 2019: A study of 302 patients from italy. *International Journal of Infectious Diseases*, 2020.

[4] Andrea Borghesi, Angelo Zigliani, Roberto Masciullo, Salvatore Golemi, Patrizia Maculotti, Davide Farina, and Roberto Maroldi. Radiographic severity index in covid-19 pneumonia: relationship to age and sex in 783 italian patients. *La radiologia medica*, page 1, 2020.

[5] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[6] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020.

[7] Virginia Dignum. Responsible artificial intelligence: designing ai for human values. 2017.

[8] Adam Jacobi, Michael Chung, Adam Bernheim, and Corey Eber. Portable chest x-ray in coronavirus disease-19 (covid-19): A pictorial review. *Clinical Imaging*, 2020.

[9] Dr. Edgar Lorente. Covid-19 imaging findings.

[10] World Health Organization et al. Coronavirus disease (covid-19): situation report, 179. 2020.

[11] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in Biology and Medicine*, page 103792, 2020.

[12] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St Jules, Xiao Yu Wang, and Alexander Wong. Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms. *arXiv*, pages arXiv–1910, 2019.

[13] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016.

[14] JCL Rodrigues, SS Hare, A Edey, A Devaraj, J Jacob, A Johnstone, R McStay, A Nair, and G Robinson. An update on covid-19 for the radiologist-a british society of thoracic imaging statement. *Clinical radiology*, 75(5):323–325, 2020.

[15] Geoffrey D Rubin, Christopher J Ryerson, Linda B Haramati, Nicola Sverzellati, Jeffrey P Kanne, Suhail Raoof, Neil W Schluger, Annalisa Volpi, Jae-Joon Yim, Ian BK Martin, et al. The role of chest imaging in patient management during the covid-19 pandemic: a multinational consensus statement from the fleischner society. *Chest*, 2020.

[16] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[17] Heshui Shi, Xiaoyu Han, Nanchuan Jiang, Yukun Cao, Osamah Alwalid, Jin Gu, Yanqing Fan, and Chuansheng Zheng. Radiological findings from 81 patients with covid-19 pneumonia in wuhan, china: a descriptive study. *The Lancet Infectious Diseases*, 2020.

[18] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[19] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[20] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[21] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *arXiv preprint arXiv:2003.09871*, 2020.

[24] Alexander Wong, Zhong Qiu Lin, Linda Wang, Audrey G Chung, Beiyi Shen, Almas Abbasi, Mahsa Hoshmand-Kochi, and Timothy Q Duong. Towards computer-aided severity assessment: training and validation of deep neural networks for geographic extent and opacity extent scoring of chest x-rays for sars-cov-2 lung disease severity. *arXiv preprint arXiv:2005.12855*, 2020.

[25] Alexander Wong, Mohammad Javad Shafiee, Brendan Chwyl, and Francis Li. Ferminets: Learning generative machines to generate efficient neural networks via generative synthesis. *arXiv preprint arXiv:1809.05989*, 2018.

[26] Gang Wu and Xiaoming Li. Mobile x-rays are highly valuable for critically ill covid patients. *European radiology*, page 1, 2020.