IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 July 2022, accepted 18 July 2022, date of publication 25 July 2022, date of current version 1 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3193700

APPLIED RESEARCH

Achieving Multisite Generalization for CNN-Based Disease Diagnosis Models by Mitigating Shortcut Learning

KAOUTAR BEN AHMED, LAWRENCE O. HALL, (Fellow, IEEE), DMITRY B. GOLDGOF, (Fellow, IEEE), AND RYAN FOGARTY, (Graduate Student Member, IEEE)
Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620, USA
Corresponding author: Kaoutar Ben Ahmed (kbenahmed@usf.edu)

ABSTRACT When a learned model has high accuracy under familiar settings (internal testing) and a big drop in accuracy under slightly different circumstances (external testing) we suspect it is using shortcuts to make decisions. This problem is known as shortcut learning. In medical imaging, shortcuts are undesired and unintended features that the model relies on to perform diagnosis. Shortcut-based solutions using medical images could lead to false diagnoses and have dangerous implications for patients. In the current COVID-19 era, a large set of papers have been published proposing the use of deep convolutional neural networks to perform diagnosis or triage of COVID-19 from chest X-rays (CXRs). These studies are reporting high accuracies which could be misleading and overestimated. To our knowledge, none of the currently published papers with high performance reported testing on samples from truly unseen data sources. Studies which did, have noticed a significant performance drop when testing on unseen sources indicating a failure to generalize. In this paper, we elucidate the generalization challenge of deep learning based models trained for disease diagnosis. We use the example of COVID-19 diagnosis from CXRs. Solutions that mitigate shortcut learning are introduced and experimentally shown to be effective. Our proposed methods enable the models to have a statistically significantly reduced performance drop-off on unseen data sources. Thus, lowering the performance drop to only 9% instead of 20%. The issues with convolutional neural networks addressed here generally apply to other imaging modalities and recognition problems, as shown.

INDEX TERMS Deep learning, CNN, shortcut learning, medical imaging, confounding features, COVID-19, X-ray.

I. INTRODUCTION

Deep learning (DL) has emerged as the leading machine learning tool in the domain of image analysis/understanding. Deep neural networks are now the state-of-the-art machine learning models across a variety of areas, from image analysis to natural language processing, and widely deployed in academia and industry. The ability of convolutional layers, for example, to extract features in an automated way has accelerated their adoption. The medical imaging community has begun a debate about whether deep learning would be applicable in medical imaging [1]. DL technology has been recently applied to healthcare problems, including computer-aided detection/diagnosis, disease prediction, image segmentation, etc. However, translation of deep learning technology from research to actual clinical use is highly challenging. The use of machine learning in general and deep learning in particular within healthcare is still in its infancy due to several factors. Mainly, labeled training data, which is necessary to train a deep learning model, is both expensive and difficult to produce. Medical imaging data are acquired in nonstandard ways and settings across sites. In addition, due to patient privacy and other concerns, having a centralized open source dataset of medical images is very rare and images

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

are distributed among different hospitals and imaging centers. Additionally, system robustness and generalization across acquisition protocols, machines, and hospitals is another core challenge, which is the focus of this paper. Prior work has found that clinical deep learning models face significant performance degradation when tested on external sites never seen during model training for different imaging modalities including chest X-rays [2]–[7].

Shortcut learning occurs when a deep learning model finds and follows a ''shortcut'' strategy to achieve excellent results under familiar circumstances but fails under slightly different settings such as real-world scenarios [8]. For instance, cows in familiar backgrounds (grass landscape) are detected and classified correctly, while cows with uncommon backgrounds (beach) are not detected or classified poorly [9]. Shortcuts are simple characteristics of a dataset that a model relies on to solve a given problem instead of learning the intended decision rules. Consequently the model suffers from generalization failures under slightly different circumstances. In medical imaging, when a diagnosis model relies on undesired shortcut features, it can fail even at the same site, when tested on a slightly shifted data distribution (for instance, a change of scanning equipment). To make sure that a model is using intended medically relevant features for disease diagnosis, it has to work well not only on internal test sets, but also perform as intended on out-of-distribution test sets (such as you get from unseen data sources). As concluded by Geirhos *et al.* [8], performance on a dataset does not necessarily indicate that the relevant concept has been learned. Recently, multiple studies in medical imaging show undesired model behavior relying on unintentional dataset bias to make decisions. Authors in [2] demonstrated that a model, trained for the purpose of identifying pneumonia from X-rays, had instead unexpectedly learned to identify particular hospital systems with near-perfect accuracy using a hospital-specific metal token on the X-ray images and other confounders. It achieved high performance on seen sources (93%) without really learning much about pneumonia. Similarly, a paper [10] studied shortcuts/confounders in the ISIC dataset (widely used for skin cancer classification), namely the presence of colored patches in benign skin images whereas the malignant lesion images contain none. Authors in [11] showed that 69.5% of the malignant images that were initially correctly classified, after manually inserting colored patches, were misclassified as benign. There is a corresponding large drop in the classifier sensitivity from 0.886 to 0.191.

The SARS COV-2 viral infection (COVID-19) can have a devastating impact on the respiratory system. It has caused an enormous number of deaths since late 2019. Radiologist examination of chest X-rays has been argued to be an effective way to screen for COVID-19, in case of lung involvement, since machines are ubiquitous and their cost and cleaning complexity is low. In our recent work [6], we reviewed a large number of published papers claiming the discovery of high performing deep neural network models capable of accurate diagnosis of COVID-19 from chest X-rays. Many approaches show high accuracy (over 90%) in differentiating from pneumonia or other non-COVID-19 classes. However, we identified multiple flaws in the suggested solutions which lead us to question their clinical utility. For instance, multiple papers used a pneumonia set which was of children and studies [12] have shown that convolutional neural networks can detect the size of an object. So models will associate anatomical features of age with the diagnosis (could tell pneumonia by smaller lungs). Additionally, other studies ([13], [14]) showed that CNNs can learn and rely on age-related shortcuts to make a decision. Furthermore, the majority of suggested works ([15], [16]) were done using cross validation or random splits (train/val/test) using all data, as COVID-19 data was hard to get. Thus, no tests on unseen sources were performed. In addition, some publicly available large databases of COVID-19 CXRs are actually a collection of other smaller open source data sources. For example the Covid-19 Radiography database [17] includes the Dr. Joseph Cohen Dataset [18]. Therefore, some papers [19], [20] which claimed generalization to unseen data sources did not realize that their training dataset is a subset of the claimed unseen source. Our experimental results in [6] showed that models which have achieved high internal testing results (AUC = 1.00), in the worst case, only scored an AUC of 0.38 externally. This suggests that generalization and robustness assessment is necessary before such models can be clinically adopted. To our knowledge, there has been no success at creating a generalized model/solution capable of performing well when tested using data from several external data sites.

The contributions of this paper are as follows.

1) We draw attention to the spreading problem of overestimated performance results of CNN-based disease diagnosis models that were only tested internally.
2) We emphasize the importance of the performance comparison between internal testing vs external testing to detect if a model is suffering from shortcut learning.
3) Shortcuts exist in most data and rarely disappear by adding more data [8]. Hence, we demonstrate that modifying the training data to block specific shortcuts is a potential solution.
4) We propose a pipeline to reduce a deep model's performance gap between internal and external sites.

The remaining sections are outlined as follows. Section II describes details of the data sources used. Section III covers the methodology and overall workflow of this work. Section IV presents the experimental setup and interprets results. Section V discusses this work's findings and compares to other methods. Finally, Section VI concludes the paper and Section VII suggests some future research directions.

II. DATA SOURCES

In this section, we describe all the data sources used and discussed in this paper for both COVID-19 positive and negative classes. All data was deidentified. For the

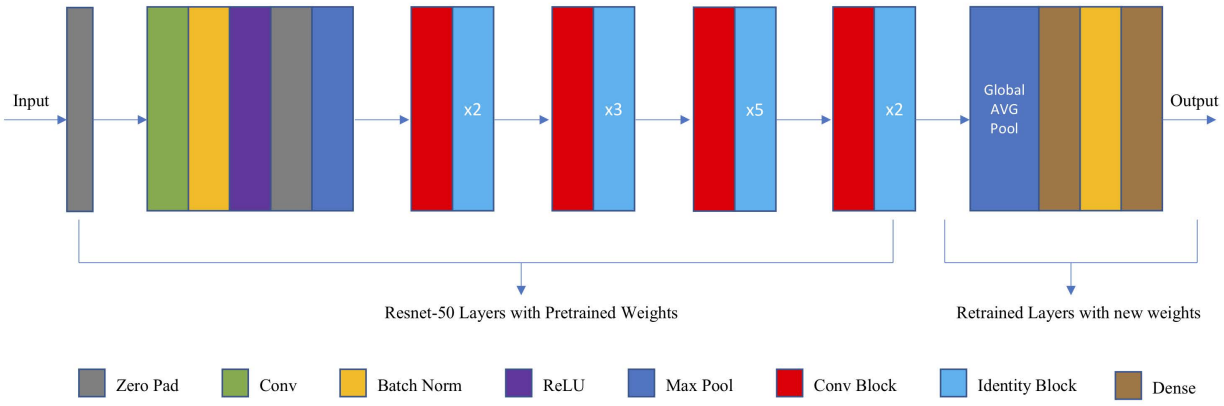


FIGURE 1. The customized ResNet50 architecture deployed in the proposed approach. Each convolution block has 3 convolution layers and each identity block also has 3 convolution layers.

COVID-19 positive class, the sources used in this work are, BIMCV-COVID-19+ [21], COVID-19-AR [22], CC-CXRI-P [23], V2-COV19-NII [24] and COVIDGR [25].

BIMCV COVID-19 is a large open-source dataset of chest X-ray images CXR (Computed Radiography (CR), Digital X-ray (DX)) and Computed Tomography (CT) images. It includes a subset of confirmed COVID-19 positive cases from the Valencian Region Medical ImageBank (BIMCV COVID-19(+)) and another subset for negative cases (BIMCV COVID-19(-)). Radiological reports are also included along with Polymerase chain reaction (PCR) results and other data. COVID-19-AR is a publicly available imaging dataset of chest X-rays and CT scans of COVID-19 patients with positive PCR tests. This dataset is available from The Cancer Imaging Archive (TCIA) and contains images donated by the radiology department and other sites at the University of Arkansas. CC-CXRI-P is a dataset of the chest X-Ray images (CXRs) constructed from cohorts from the China Consortium of Chest X-ray Image Investigation (CC-CXRI). All CXRs are classified into COVID-19 pneumonia due to SARS-CoV-2 virus infection, other viral pneumonia, bacterial pneumonia, other lung disorders, and normal controls. V2-COV19-NII is a data repository consisting of de-identified radiological imaging and clinical data of positive COVID-19 patients published by the Institute for Diagnostic and Interventional Radiology, Hannover Medical School, Hannover, Germany. X-ray images did not undergo any pre-processing and were saved in Niftii format. COVIDGR is a set of X-ray images to assist in the diagnosis of the COVID-19 disease, built with the close collaboration of expert radiologists in Spain. All the images were obtained from the same equipment and under the same X-ray regime. Only the PosteriorAnterior (PA) view is available.

The COVID-19 negative class includes cases with non-COVID-19 lung diseases as well as normal cases. We used data from 3 sources: the National Institutes of Health (NIH) chest X-ray dataset [26], Chexpert [27] and Padchest [28].

TABLE 1. Details of data source used in this work.

Class	Data Source	Number of Images	Image Type
COVID-19(+)	BIMCV-COVID-19(+)	468	PNG
	V2-COV19-NII	243	NIFTII
	CC-CXRI-P(+)	513	JPG
	COVID-19-AR	77	DICOM
	COVIDGR	426	JPG
COVID-19(-)	NIH	607	PNG
	Chexpert	420	JPG
	Padchest	310	PNG
	BIMCV-COVID-19(-)	500	PNG
	CC-CXRI-P(-)	542	JPG

Table 1 summarizes all the data sources used in this paper. Note that ground truth labels were extracted from the metadata and radiology reports included with all of the datasets.

III. METHODS

A. BASELINE MODEL

Here, we used the built model in our previous work [6] as a baseline. It consists of a ResNet50 pre-trained on ImageNet as a base model. We removed the fully connected layers of the base model. Then, we applied global average pooling after the last convolutional layer of the base model and we added a new fully connected layer of 64 units with random weight initialization. Finally, we added an output layer with Sigmoid activation. The resultant model was fine-tuned using the X-ray imaging data. All the layers of the base model were frozen and only the weights of the newly added layers were learned. The total number of trainable parameters was 184K. Fig. 1 illustrates the architecture of the customized Resnet-50 used as baseline model. In this experiment, the model was trained using a cosine annealing cyclic learning rate with a maximum learning rate of 10^{-4}. Training was

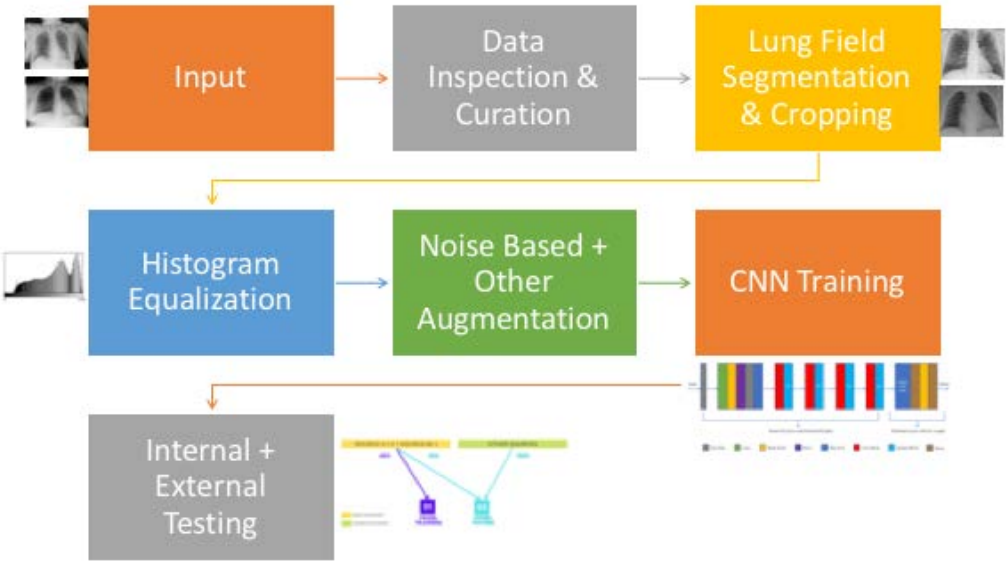


FIGURE 2. Overall workflow of suggested approach to mitigate shortcut learning.

done for 1000 epochs and model snapshots were saved at the end of each of 20 cycles, which were 50 epochs long. Hyper-parameter tuning was not needed in this scenario since high internal performance results were obtained. As a first step to prepare the data for training, we normalized all the images to 8 bit PNG format in the [0-255] range. The images were originally 1 grayscale channel, we duplicated them to 3 channels. The reason behind this is that Resnet50, our base model, was pretrained on 8 bit color (3 channel) images of the Imagenet corpus. To augment the relatively small training data we applied rotations of 2, 4, −2, and −4 degrees and horizontal flipping. We chose a small rotation angle as X-rays are typically not rotated much. The training data was standardized by subtracting the mean of all data and dividing by its standard deviation. Code and data are available on Github.[1]

B. OVERALL WORKFLOW

The workflow starts with multiple sets of data coming from separate sources. The first step is the visual inspection of the images and the associated ground truth and metadata. Data curation is explained in more detail in Section IV-A. Then, we suggest lung field segmentation and cropping the inputs as a first pre-processing step to get rid of noise signals outside the lungs. Section IV-B provides more experimental results of the impact of this step. Histogram equalization is a second pre-processing step that we recommend and results of this step can be found in Section IV-C. To augment the size of the training set, we suggest using Gaussian noise augmentation in addition to the other traditional augmentations (rotation, flipping, etc.). Details of this type of augmentation are experimentally studied in Section IV-D. Once training data is prepared, we can send it as input to the CNN of choice to perform training. In this study we used a modified Resnet-50 architecture, but any CNN model can be used to perform this task. Finally, a generalization test, similar to the one we suggest in Section III-C, needs to be performed to compare internal vs external source testing performance.

[1]https://github.com/COVID-19-Diagnosis

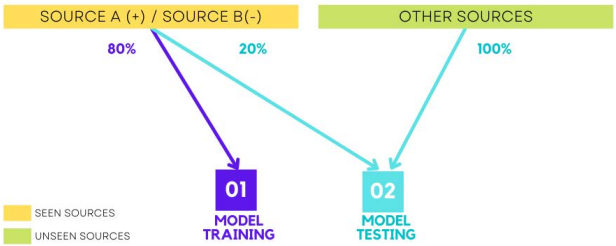


FIGURE 3. Generalization test experiment.

C. GENERALIZATION TEST

In order to investigate the generalization ability of deep learned models (which is the main focus of this paper), evaluation was performed on external data sources from which there were no examples in the training data. Experiments were done with training data from just one source per class (for instance, source A for COVID-19-positive class and source B for COVID-19-negative class). To compare internal vs external testing, We split the data sources into seen (internal) and unseen (external). A subset (80%) from the seen data sources is used for model training. Then, we compare model testing performance using 1) the randomly chosen held-out subset (20%) from the seen sources versus 2) testing samples from unseen external data sources (see Fig. 3).

D. METRICS

In the following sections we present the metrics: *accuracy*, *precision*, *recall* (aka sensitivity), F_1-score, and area under the curve (*AUC*). The classes represent the binomial case, COVID versus non-COVID. If we define a confusion matrix with true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), then we may derive the following metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

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

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 score = \frac{2}{precision^{-1} + recall^{-1}} \tag{4}$$

$$Specificity = \frac{TN}{FP + TN} \tag{5}$$

$$AUC = \frac{recall + specificity}{2} \tag{6}$$

Additionally, as mentioned is Section III-C, we use the difference between the model's performance on seen sources P_{seen} vs on unseen sources P_{unseen} as a metric of generalization. We call it a generalization drop-off.

$$DropOff = P_{seen} - P_{unseen} \tag{7}$$

IV. EXPERIMENTS AND RESULTS

A. DATA INSPECTION AND CURATION

One of the fundamental aspects to achieve a reliable contribution of AI in detecting COVID-19 from CXRs is the compilation of an adequate set of images in terms of both quality and quantity. To assure the validity of the ground truth, we made sure not to rely only on a positive RT-PCR but also, when possible, the associated CXR report confirming and supporting the test results. Visual assessment of the imaging data allowed us to remove multiple cases with a lateral view even though the corresponding metadata says it's frontal. Other metadata mistakes include X-ray projection and position, for example an AP Portable X-ray can be mislabeled as PA. As DICOM information on the type of projection may not always be included, some data sources [28] used pre-trained models for automatic X-ray view determination. Thus, some incorrect labeling can occur. Additionally, unlike the usual negative X-ray mode (''bones white''), some images were in a positive mode (''bones black''). To avoid any problems these images may cause to the model's performance, we inverted them to the conventional mode (''bones white''). To avoid creating confounders based on the CXR view, we used frontal view (from either AP or PA) CXRs in both classes.

Currently, due to data security and patient privacy, there is a limited availability of open source datasets with the desired quality and quantity to build a diagnostic system with clinical value. Recently, researchers have begun to realize that convolutional neural networks trained to identify COVID-19 from CXRs may not be learning underlying diagnostic features, but also exploit confounding information. For example, it was shown that CNNs were able to predict where a radiograph was acquired (hospital system, hospital department) with extremely high accuracy [29]. In this section, we show the impact of some situational dataset-related shortcuts, we experimentally identified, to facilitate risk analysis and enable mitigation strategies while preparing training datasets.

1) X-RAY POSITION

the PA erect view is the preferred imaging view in general, but if the patient is not able to stand up it is common to do an AP Portable view image. Most of the open source COVID-19 datasets are in AP Portable X-ray view due to the severe sickness of the patients which makes them unable to leave the bed to perform an upright PA scan. The use of bedside mobile CXR apparatus is frequently associated with more severe disease (due to lack of patient mobility). If the training data in the majority or all of the negative cases consist of PA views (patient standing), the model can easily learn to make the classification based on the patient's position and use apparatus portability labels as a signal indicating the disease and its severity.

Results in [29] show that CNNs can ignore disease related features and separate portable radiographs from the inpatient wards and emergency department with 100% accuracy using distinctive text indicating laterality and use of a portable scanner.

TABLE 2. External testing results of the model trained on CC-CXRI-P dataset.

Class	Data Source	Presence of Devices	Size	Accuracy
COVID-19(+)	V2-COV19-NII	Majority	142	0.96
	COVIDGR(+)	None	326	0.38
	COVID-19-AR	Majority	69	0.94
COVID-19(-)	Chexpert	Majority	205	0.23
	NIH	Some	197	0.6
	COVIDGR(-)	None	236	0.79

2) DEVICES AND TUBES

Another confounding factor might be the presence of medical devices like ventilation equipment or ECG cables. Many of the positive images for COVID-19 present intubated patients, with electrodes and their cables, pacemakers, among other potential markers. If there is absence of medical devices in most or all the training samples in the negative class, the model can associate images with patient treatment instead of disease status. Experimentally, we used the CC-CXRI-P dataset [23] to train a model to differentiate COVID-19 cases from normal cases. The model achieved an internal testing accuracy of 99% on unseen samples from the same dataset. The extremely high performance led us to further investigate. A visual inspection of the subset of the CC-CXRI-P dataset

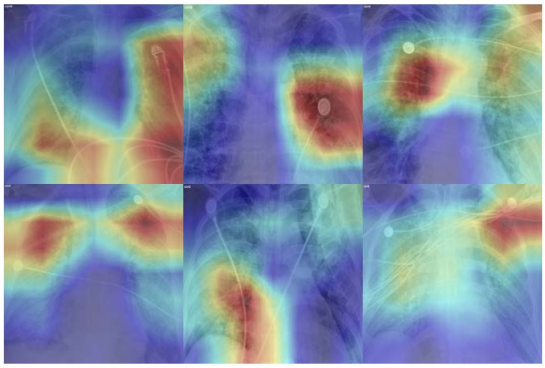


FIGURE 4. Gradcam showing the model focusing on medical devices.

used for this experiment shows that all patients in the positive class have medical devices visible and none of the patients in the negative class have any tubes or devices. Thus, suggesting a potential shortcut learning situation. To confirm the issue, we tested the model on data from unseen sites with and without presence of devices.

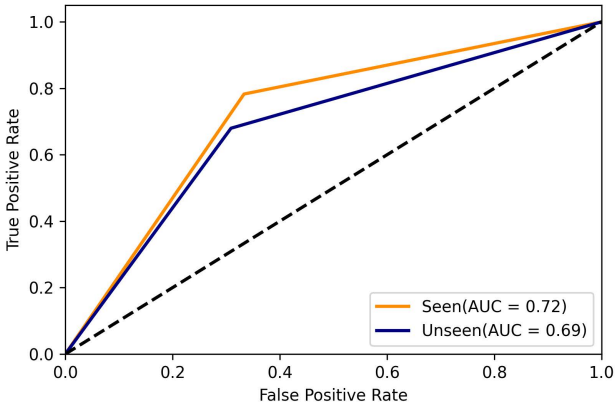
Table. 2 shows the external testing results of the model by source. The results suggest that the model is associating COVID-19 positive infection with the presence of medical devices. The accuracy was high on COVID-19 positive sources where the majority of its patients had devices (V2-COV19-NII: 96% and COVID-19-AR: 94%). Whereas the accuracy was very low on a source where the patient had no visible devices (COVIDGR+: 38%). Similarly, it can be seen that the model is associating COVID-19 negative infection to the absence of devices. The accuracy was good on negative sources that only had few or no devices visible (NIH: 60% and COVIDGR-: 79%) but very low on sources with devices (Chexpert: 23%).

To visually explain the decision of a CNN, Gradient-weighted Class Activation Mapping (GradCAM) [30] can be used. The GradCam visualization shown in Fig. 4 demonstrates that the highly activated areas (redish and yellow) are located around tubes and devices.

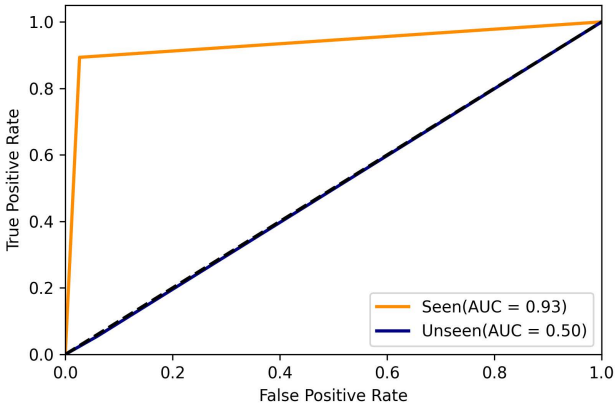
Datasets which provide additional annotations on the presence of medical devices are needed. Alternatively, novel algorithms to automatically detect (and perhaps remove) the devices are required.

3) RIB SUPPRESSION

In addition to the confounders discussed above, another potential source of bias are the ribs. Ribs on some CXR datasets can generally appear to be more prominent than in other datasets. Therefore the appearance of the ribs can be a potential confounder that the model can rely on to make decisions. A study [31] shows that rib suppression improves model generalization when identifying lung nodules from X-rays. Thus, it may improve the COVID-19 classifier generalization as well.



(a) both Classes from Same Source



(b) both Classes from Different Sources

FIGURE 5. Comparison of internal(seen) vs external(unseen) test results when training with data where (a) both classes from same source vs (b) from different sources.

4) BOTH CLASSES FROM THE SAME SOURCE

The heterogeneity of the images coming from different sources makes the CNN learn characteristics that are not themselves truly indicative of COVID-19. As a solution, we suggest to design the training data using both the COVID-19 positive and negative class from one single source. In our attempt to experimentally demonstrate the impact of this solution, we built two training sets that correspond to two scenarios. The first is the case where both classes are from the same source and the second is the scenario where both classes come from different sources.

In the first scenario, we used the BIMCV-COVID-19 data source. This source published both positive and negative cases. We chose samples to include in the COVID-19 positive class if both a confirmed PCR test and radiological findings (visible in the X-ray and noted in the radiological report) are present. Whereas samples in the negative class are ones with a negative PCR test and no findings present in the CXR or documented in the reports. Fig. 5a compares the internal vs

external performance of a model trained with data where both classes belong to the same source. Note that the performance is decent and the testing results are consistent across sites, both seen and unseen.

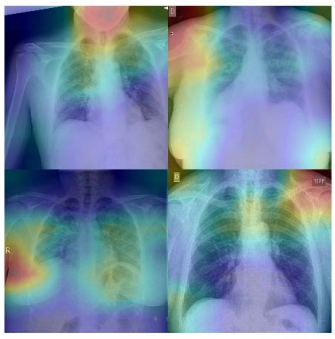
In the second scenario, for the positive class, we used the same samples from the experiment above (BIMCV-COVID-19+). Whereas for the negative class, we used samples from a different source (NIH). We picked images with the ''no findings'' label, which refers to having no pulmonary diseases, to be consistent with the negative class used in the first scenario. Fig. 5b compares the internal vs external performance of a model trained with data where both classes belong to different sources. As seen in the figure, there is a big accuracy gap when testing on data from seen sites vs unseen ones. The model shows almost perfect performance on data from seen sources while it fails at classifying data from unseen sites. Thus, suggesting shortcut learning.

The results suggest that COVID-19 positive and COVID-19 negative classes should be consistently sourced and go through consistent image acquisition and pre-processing pipelines. This will minimize systematic structural differences between the classes. Thus, eliminate bias. However, a shortcut could still be learned that will be ineffective when a local acquisition protocol changes. For the rest of this paper we focus on a typical case where having all classes from the same source is not possible.

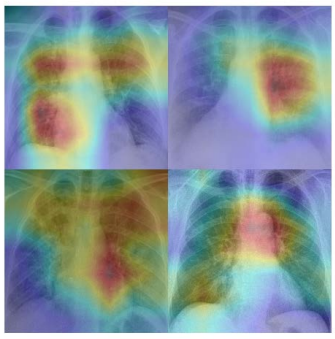
B. LUNG SEGMENTATION AND CROPPING

The first step to debiasing a deep learning model trained to identify COVID-19 from CXR images is the process of lung segmentation and cropping. The majority of the CXR images available online have markers present at the edges of the images (dates, projection labels, portability markers, arrows…etc). Deep learning models may easily rely on those markers for learning and completely or partially ignore the lung area, which contains the desired information in this context. Thus, the learned model will be unreliable for clinical decision support even though it can achieve very high classification results. Therefore, we recommend lung field segmentation and cropping as a first step in data pre-processing and preparation. A model trained using cropped CXRs will be forced to focus on information in the lungs, rather than the burned-in annotations, when making decisions. This will increase its generalization and reliability. Fig. 6a shows a GradCam visualization of our baseline model, trained using original uncropped images. Heatmaps that highlight the regions or pixels which had the highest impact on predicting the actual class are placed on top of the original images. Higher intensity in the heat-map indicates higher importance of the corresponding pixel in the decision. The heatmaps suggest that the model is activated by the noise areas and it is ignoring the lung field. Fig. 6b shows more focus on the lung area and will be discussed in the proceeding.

The utility of lung field segmentation and cropping in medical image classification is supported in previous studies [32].



(a) GradCam results when using Original CXRs



(b) GradCam results when using Cropped CXRs

FIGURE 6. An example of an input CXR before vs after lung cropping.

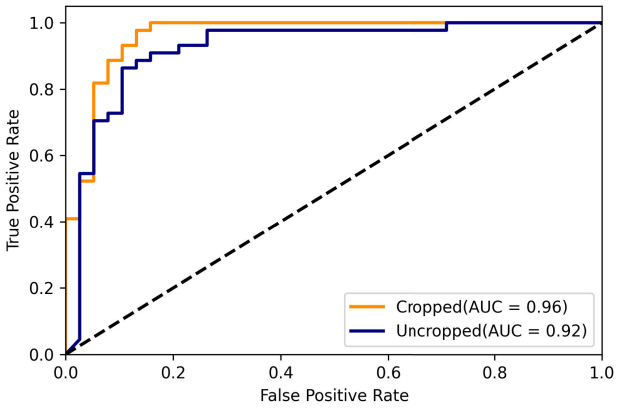
(a) Original CXR

(b) Cropped CXR

FIGURE 7. An example of an input CXR before vs after lung cropping.

A comprehensive review of deep learning based medical image segmentation techniques may be found in [33].

To assess the impact of lung segmentation and cropping we compare the performance of a model trained using uncropped images versus another model trained using cropped inputs. For the cropped inputs, we used a U-NET model, pre-trained



(a) Internal Testing

	Precision	Recall	F1-score	Accuracy	AUC
Cropped	0.53	0.69	0.56	0.57	0.59
Uncropped	0.5	0.55	0.52	0.54	0.53

(b) External Testing

FIGURE 8. Performance comparison of models trained with cropped vs uncropped images.

by [34] on a CXR dataset, to perform automatic lungs segmentation. Then, we cropped the lungs area based on the generated mask. Fig. 7 shows an example of an original CXR before and after cropping.

For model training, BIMCV-COVID-19+ data was used as the source for COVID-19 positive class and Padchest for the COVID-19 negative class. A randomly chosen 20% of the data was held out for internal testing. For external testing of a model on unseen sites, we used samples from all the other sources not used in training.

We tested the last model's snapshot (20th or 1000th epoch). Fig. 8a compares the internal performance of models trained with uncropped vs cropped images. While Fig. 8b compares the external performance of the two models.

The results suggest that the model's accuracy is statistically significantly higher when images are cropped. Using the McNemar's test [35], we calculated a p-value of 0.02 which provides 98% confidence the difference between the model's performance on unseen sources before and after cropping is significant. Moreover, Fig. 6b shows GradCam results after lung field cropping. It can be seen that the model shifts focus to the lung area instead of the X-ray's corners where noise was present. However, the significant drop in AUC shows that even after lung segmentation, a present shortcut is still influencing the classification model. Thus, more ways to avoid confounders must be evaluated to design a proper COVID-19 diagnosis system using CXR images.

C. HISTOGRAM EQUALIZATION

Systematic differences in image contrast between the datasets can result in biased model training and negatively impact model generalization. As seen in Fig. 9, datasets are easily distinguished from each other due to the obvious difference in the brightness and contrast of the images. CXR images can have variable contrast based on technical calibration of the imaging acquisition equipment. Specifically, the energy of the primary beam and the possible application of techniques to reduce scatter radiation such as collimation, grids or air gaps [36].

Histogram equalization processing has proven to be effective for normalization [37], [38]. Therefore, in this study histogram equalization was used as a preprocessing step on all CXR sources. In Fig. 10, the result of equalizing shows consistent contrast between sources. All values above the 95% order statistic were clipped to 1.0 since some sources had bright annotations and other sources did not. This technique ensured that for every source the brightest CXR pixels were normalized to the brightest white.

In order to experimentally see the impact of histogram equalization, we compared training the baseline model with histogram equalized preprocessed inputs vs unprocessed images. We used the V2-COV19-NII for COVID-19 positive class and Padchest for COVID-19 negative class.

We tested the last model's snapshot (20th). Fig. 11a compares the internal performance of models trained with histogram equalized images vs with unprocessed images. While Fig. 11b compares the external performance of the two models. The results suggest that the model performs statistically significantly better when trained with histogram equalized images. Using the McNemar's test [35], we calculated a p-value of 1.7×10^{-4} which is lower than the significance threshold, $alpha = 0.01$. There is a significant difference between the model's performance on unseen sources before and after histogram equalization with 99% confidence.

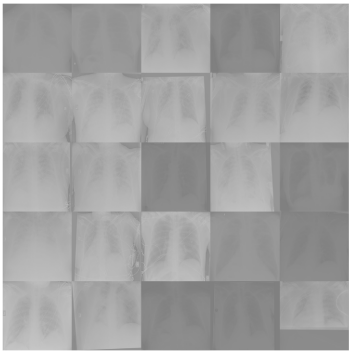
However, a high testing accuracy gap between external and internal sources still exists, which suggests the persistence of shortcut learning.

Similar techniques such as quantile normalization, and histogram matching were also explored, but determined to be no more effective.

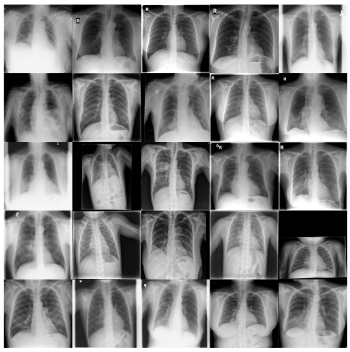
D. NOISE-BASED AUGMENTATION

In addition to image rotations, zero mean Gaussian noise with a standard deviation of 0.01 was applied to the original images. Over-fitting usually happens when the deep learning model tries to learn high frequency features (patterns that frequently occur) which may not be useful. Gaussian noise, which has zero mean, essentially has data points in all frequencies, effectively distorting the high frequency features and helping the model to look past them.

We applied this technique as a data augmentation procedure. This approach ensures that the DL network will be less sensitive to source noise level. Samples were augmented prior to training, while original fidelity images were also included during training. In this way, the trained network needed to classify across noisy and clean images.

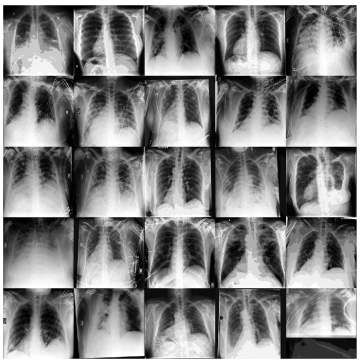


(a) V2-COV19-NII

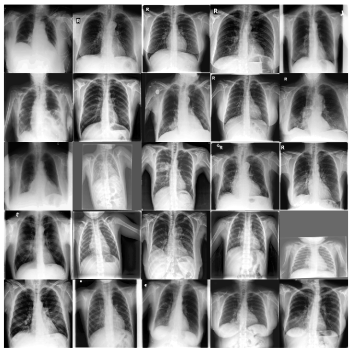


(b) Padchest dataset

FIGURE 9. Before histogram equalization.



(a) V2-COV19-NII



(b) Padchest dataset

FIGURE 10. After histogram equalization.

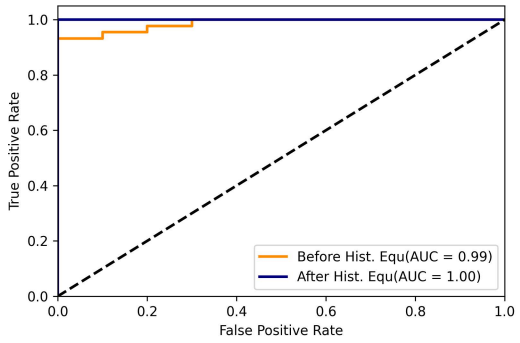
The Euclidean distance between the original image and its noisy augmentation was 0.014. Fig. 12 shows the comparison of the performance gap between seen and unseen sources before adding noise based augmentation (solid marker) and after (dotted marker).

The test accuracy of the last snapshot (20th) when using only rotation and flip augmentation for seen sources was 92% and 72% for unseen sources. When adding Gaussian noise based augmentation, the accuracy for seen was 90% and 81% for unseen sources. After adding the Gaussian noise based augmentation, the accuracy gap was improved to only 9% instead of 20%.

The test accuracy on unseen sources is statistically significantly higher when adding Gaussian noise augmentation (81%) than when only using flip and rotate augmentation (72%). Using the McNemar's test [35], we calculated a p-value of 8.13×10^{-5} which is lower than the significance threshold, $alpha = 0.01$. There is a significant difference between the model's performance on unseen sources before and after adding Gaussian noise based data augmentation with 99% confidence.

The results suggest that making features that are not medically relevant, but good discriminators, noisy can result in less focus on them. Medically relevant features are naturally expected to have some "noise" and be less affected. To further assess the effect of the amount of noise added, we experimented with multiple values of Gaussian noise standard deviation {0.005, 0.01, 0.05}. Fig. 13 shows the results of testing the 20th model snapshot for different values of Gaussian noise standard deviation. Based on the obtained results, it doesn't seem to hurt the accuracy much on testing data from seen sources. However, it allowed significant improvement to the performance on data from unseen sources. This suggests that the model has learned to ignore confounding features and focus on more medically meaningfully differences between classes.

To answer the question of how much noise is good enough, based on our experimental results, adding more noise



(a) Internal Testing

	Precision	Recall	F1-score	Accuracy	AUC
With Hist. Equ.	0.54	0.5	0.52	0.61	0.66
Unprocessed	0.52	0.27	0.36	0.58	0.59

(b) External Testing

FIGURE 11. Performance comparison of models trained with histogram equalized pre-processed images vs with unprocessed images.

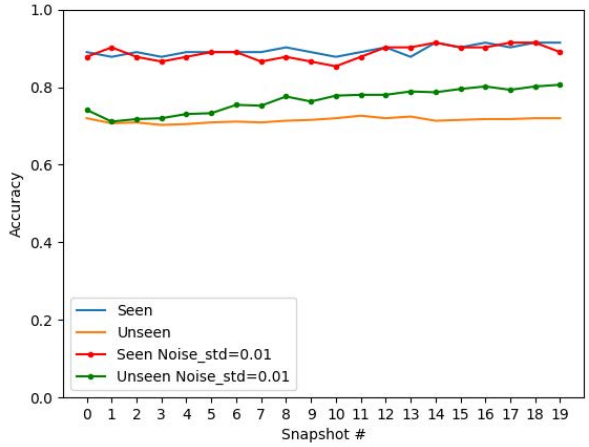


FIGURE 12. Test accuracy of the saved snapshots on seen vs unseen data sources after the Gaussian noise based training data augmentation.

Accuracy	baseline	std = 0.01	std=0.05	std=0.005
Seen	0.92	0.9	0.88	0.94
Unseen	0.72	0.81	0.78	0.78
Difference	0.2	0.09	0.1	0.16

FIGURE 13. Comparison of the performance gap when testing the 20th snapshot on seen vs unseen data sources for multiple Gaussian noise standard deviation values.

(std = 0.05) has little to no effect on the accuracy performance for seen sources. However it makes the accuracy of unseen sources slightly lower than a noise application of std = 0.01. Less noise (std = 0.005) increased the accuracy performance on both the seen and unseen sources at early stages but at the end the performance gap becomes larger. Generally, we noticed that the application of noise based data augmentation had no effect (or a slight improvement) on the performance for seen sources. However, it has significantly improved the model's performance on unseen sources.

TABLE 3. Testing accuracy performance of DeepCOVID-XR on unseen COVID-19 sources.

Data Source (Number of Samples)	Sensitivity
BIMCV COVID-19+ (38)	66%
V2-COV19-NII (163)	53%
COVID-19-AR (70)	63%
Overall (average)	58%

V. DISCUSSION

As we mentioned before, in this study we used a customized Resnet-50 as a CNN model. However, the problem of shortcut learning is present in all deep learning based models and our solution is applicable to any CNN regardless of the architecture.

To experimentally show that shortcut learning exists in other architectures, we tested a COVID-19 diagnosis system called DeepCOVID-XR [39]. It is claimed that it is capable of generalizing to a held-out test data set of 2214 images (1192 positive for COVID-19) from a single institution that the algorithm was not exposed to during model development. Their method achieved a sensitivity of 75% (898 of 1192). Authors claim that DeepCOVID-XR was trained and tested on, to their knowledge, the largest clinical data set of chest radiographs from the COVID-19 era of any published AI solution to date. It included images from multiple institutions across a large U.S. health care system (17,002 images from 5,853 patients total). DeepCOVID-XR consists of an ensemble of six CNN architectures (DenseNet-121, ResNet-50, InceptionV3, Inception-ResNetV2, Xception, and EfficientNet-B2). The CNNs in this ensemble were pretrained on a large publicly available data set of chest radiographs from the National Institutes of Health [26] and were then fine-tuned on a private clinical training set of patient's chest radiographs from over 20 sites across the Northwestern Memorial Health Care System who were tested for COVID-19 from February 2020 to April 2020. Note that the external testing was done on data from an institution that is affiliated with the same health care system from which the training data was obtained. Therefore, there is significant likelihood that the testing data have a very similar distribution and cover a very similar population as the training data. Thus, their good testing results (sensitivity of 75%) may not guarantee generalization. Their pretrained models were made available on Github [40]. We tested their system on some external sources of COVID-19 positive to assess the robustness and generalization ability of their solution. Table. 3 shows the testing results with some significant drop-offs in accuracy. Overall the model had a sensitivity of 58% (156 of 271).

This suggests that having large amount of training data from multiple sources and an ensemble of several advanced CNN architectures is not a complete solution to the problem of generalization to unseen sources. It may be impractical to have one set of data that covers the space of possible X-ray data acquisition parameters and protocols.

Additionally, we compared the modified Resnet-50 presented here against other CNN architectures. Instead of

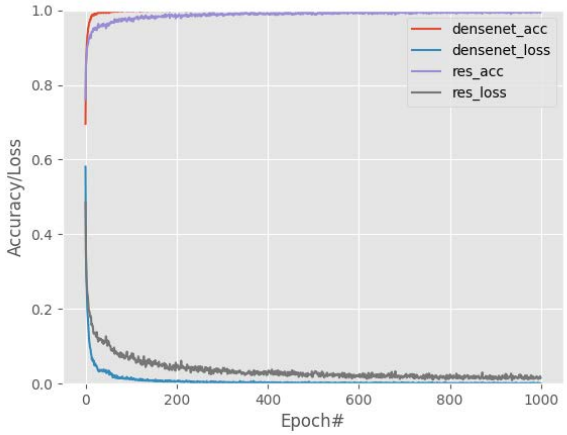


FIGURE 14. Comparison of training accuracy/loss per epoch for Resnet-50 vs Densenet-121 as base model.

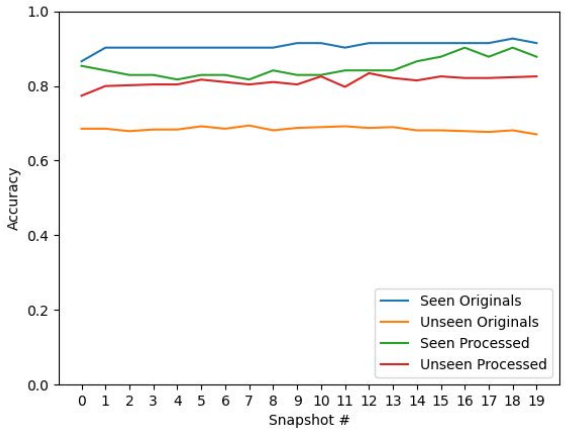


FIGURE 15. Test accuracy of the saved densenet-121 snapshots on seen vs unseen data sources after application of the proposed approach.

Resnet-50 as a pretrained base model, we used DenseNet-121. Fig. 14 shows a comparison of the training accuracy/loss per epoch of the compared models. For model training, as in Section IV-B, BIMCV-COVID-19+ data was used as the source for COVID-19 positive class and Padchest for the COVID-19 negative class. It can be seen that both models have similar training graph. Also, Densenet-121 was faster to reach 100% training accuracy than Resnet-50. The internal vs external testing results of Densenet-121 are presented in Table 4. The results (drop-off = 0.37) suggest that Densenet-121 suffers from shortcut learning as well. We can also notice that when a model has higher training, it achieves higher seen testing accuracy but has a larger drop-off when tested externally.

TABLE 4. Comparison of internal vs external performance of Resnet-50 vs Densenet-121 as base model.

Model	Num of params	Training Acc	Seen Acc	Unseen Acc	Drop-off
ResNet-50	184,449	0.9956	0.82	0.53	0.29
DenseNet-121	149,441	1.0	0.93	0.56	0.37

In order to determine if our method helps with DenseNet-121, we apply the proposed solution and compare seen vs unseen testing performance before and after application of the proposed method. Fig. 15 shows results. It can be seen that the drop-off between internal and external testing has been reduced with the application of our proposed pipeline. Note for faster testing results, the unseen testing set used in this experiment is small subset (232 positive and 232 negative cases) randomly chosen from the unseen testing set used in Table 4 which consisted of 1200 positive and 1200 negative cases.

On the other hand, to our knowledge there are no similar studies to compare with our approach. However, in a related paper, a study on congestive heart failure [13] observed in CXRs addressed shortcuts by pre-training a model originally trained on Imagenet on a related problem without shortcuts (pneumonia and other lung diseases which show up in CXR). They found this approach reduced the likelihood of the final tuning capturing shortcuts. However, the approach requires choosing one particular known bias to precisely target and it needs careful data curation and extra steps in training. In our case, we are thinking of unknown, "concealed" shortcuts.

VI. CONCLUSION

Deep learning based medical diagnostic systems cannot be considered clinically mature until they are also shown to be generalizable across source populations. The issues raised here apply to other image modalities and other diseases for which X-ray imaging is used. We point out biases to look-out for while training models and suggest steps for training data curation and preprocessing to avoid shortcut learning. For lung diseases like variants of COVID-19, the use of lung segmentation, histogram equalization, and added noise to images significantly reduced the gap in performance between cross validation results and unseen source predictions. While overcoming shortcut learning entirely may require new neural network formulations, in this work we show important steps towards mitigating it. Thus, leading to more fair, reliable, robust and transparent deep learning based clinical decision support systems. Our proposed methods enabled the models to perform statistically significantly higher on unseen sources. Thus, lowering the performance drop to only 9% instead of 20%.

VII. FUTURE WORK

Our work here is concerned with mitigating shortcuts on the data level. In the future, we are planning to explore the possibility of solving shortcut learning at the model level. The black-box nature of CNNs continues to be further explored. More complex loss functions may be needed to avoid learning simple functions which capture shortcuts.

In the context of clinical applications, an important step for employing a DL approach for CXR support systems, will be fine-tuning or calibration of the system to avoid the unique artifacts and confounders at a source. Especially revealing is the sensitivity of confounding factors, and each source will have to be carefully analyzed to avoid issues like those found in this research. When working with medical images, it is very important to visually examine a significant sample from each source.

In the future, new approaches may alleviate the problems encountered in cross-source generalization. Numerous alternate state-of-the-art techniques may provide significant improvements for medical imaging classification tasks on unseen sources and deserve study such as: multi-task learning [41], few-shot learning [42], semi-supervised learning [43], attention models [44], data synthesis (especially of confounders such as synthetic tubes/devices) [45], and federated learning [46]. Each of these approaches improve generalization by ensuring that feature distillation is appropriately abstract, while also ensuring that the system is focused on classifying with the appropriate features. While federated learning has become a hot topic for the medical industry, it requires a high degree of coordination between facilities and one may not be able to examine potential shortcuts at non-local facilities. It must be noted that even if there is a broad variety of data there can still be shortcuts learned. Such as a high likelihood of a person with a device attached having the disease to be found. While a device indicates a suspected illness, no human doctor would diagnose based on its presence.

The medical industry, as well as any mission critical application, requires trust in an algorithm, but this trust need not be implicit. Explainable AI (XAI) is another key factor for engineers and developers to successfully transition technology from the lab to the clinic. Decision support systems for the medical community can be hugely beneficial to the radiomics and pathology communities, but recommendations will need to be evidence based. Technologies such as GradCAM are helping researchers, but other more sophisticated XAI solutions will be needed in the clinical space.

REFERENCES

[1] B. Zhao, "Understanding sources of variation to improve the reproducibility of radiomics," *Frontiers Oncol.*, vol. 11, p. 826, Mar. 2021, doi: 10.3389/fonc.2021.633176.

[2] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002683.

[3] E. H. P. Pooch, P. Ballester, and R. C. Barros, "Can we trust deep learning based diagnosis? The impact of domain shift in chest radiograph classfication," in *Thoracic Image Analysis*, J. Petersen, R. S. J. Estépar, A. Schmidt-Richberg, S. Gerard, B. Lassen-Schmidt, C. Jacobs, R. Beichel, and K. Mori, Eds. Lima, Peru: Springer, 2020, pp. 74–83.

[4] J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand, "On the limits of cross-domain generalization in automated X-ray prediction," in *Proc. 3rd Conf. Med. Imag. With Deep Learn.*, in Proceedings of Machine Learning Research, vol. 121, T. Arbel, I. Ben Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, and C. Pal, Eds. Montreal, QC, Canada: PMLR, Jul. 2020, pp. 136–155.

[5] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Mach. Intell.*, vol. 3, no. 7, pp. 610–619, May 2021.

[6] K. B. Ahmed, G. M. Goldgof, R. Paul, D. B. Goldgof, and L. O. Hall, "Discovery of a generalization gap of convolutional neural networks on COVID-19 X-rays classification," *IEEE Access*, vol. 9, pp. 72970–72979, 2021.

[7] J. D. López-Cabrera, R. Orozco-Morales, J. A. Portal-Diaz, O. Lovelle-Enríquez, and M. Pérez-Díaz, "Current limitations to identify COVID-19 using artificial intelligence with chest X-ray imaging," *Health Technol.*, vol. 11, no. 2, pp. 411–424, Mar. 2021.

[8] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[9] S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 456–473.

[10] L. Rieger, C. Singh, W. Murdoch, and B. Yu, "Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 8116–8126.

[11] M. Nauta, R. Walsh, A. Dubowski, and C. Seifert, "Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis," *Diagnostics*, vol. 12, no. 1, p. 40, Dec. 2021.

[12] D. Cherezov, R. Paul, N. Fetisov, R. J. Gillies, M. B. Schabath, D. B. Goldgof, and L. O. Hall, "Lung nodule sizes are encoded when scaling CT image for CNN's," *Tomography*, vol. 6, no. 2, pp. 209–215, Jun. 2020, doi: 10.18383/j.tom.2019.00024.

[13] S. Jabbour, D. Fouhey, E. Kazerooni, M. W. Sjoding, and J. Wiens, "Deep learning applied to chest X-rays: Exploiting and preventing shortcuts," in *Proc. 5th Mach. Learn. for Healthcare Conf.*, F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., vol. 126, Aug. 2020, pp. 750–782. [Online]. Available: https://proceedings.mlr.press/v126/jabbour20a.html

[14] Z. Lin, Z. He, S. Xie, X. Wang, J. Tan, J. Lu, and B. Tan, "AANet: Adaptive attention network for COVID-19 detection from chest X-ray images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4781–4792, Nov. 2021.

[15] J. Civit-Masot, F. Luna-Perejón, M. D. Morales, and A. Civit, "Deep learning system for COVID-19 diagnosis aid using X-ray pulmonary images," *Appl. Sci.*, vol. 10, no. 13, p. 4640, Jul. 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/13/4640

[16] S. H. Yoo, H. Geng, T. L. Chiu, S. K. Yu, D. C. Cho, J. Heo, M. S. Choi, I. H. Choi, C. Cung Van, N. V. Nhung, B. J. Min, and H. Lee, "Deep learning-based decision-tree classifier for COVID-19 diagnosis from chest X-ray imaging," *Frontiers Med.*, vol. 7, Jul. 2020, doi: 10.3389/fmed.2020.00427.

[17] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam, "Can AI help in screening viral and COVID-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132665–132676, 2020.

[18] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q Duong, and M. Ghassemi, "COVID-19 image data collection: Prospective predictions are the future," 2020, *arXiv:2006.11988*.

[19] F. Li, X. Lu, and J. Yuan, "MHA-CoroCapsule: Multi-head attention routing-based capsule network for COVID-19 chest X-ray image classification," *IEEE Trans. Med. Imag.*, vol. 41, no. 5, pp. 1208–1218, May 2021.

[20] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[21] M. de la Iglesia Vayá, J. M. Saborit, J. A. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. García-García, M. Caparrós, G. González, and J. M. Salinas, "BIMCV COVID-19+: A large annotated dataset of RX and CT images from COVID-19 patients," 2020, *arXiv:2006.01174*.

[22] S. Desai, A. Baghal, T. Wongsurawat, P. Jenjaroenpun, T. Powell, S. Al-Shukri, K. Gates, P. Farmer, M. Rutherford, G. Blake, T. Nolan, K. Sexton, W. Bennett, K. Smith, S. Syed, and F. Prior, "Chest imaging representing a COVID-19 positive rural U.S. population," *Sci. Data*, vol. 7, no. 1, Nov. 2020.

[23] G. Wang *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 Biomed. Eng.*, vol. 5, no. 6, pp. 509–521, Apr. 2021.

[24] H. B. Winther, H. Laser, S. Gerbel, S. K. Maschke, J. B. Hinrichs, J. Vogel-Claussen, F. K. Wacker, M. M. Höper, and B. C. Meyer, "COVID-19 image repository," Inst. Diagnostic Intervent. Radiol., Hannover Med. School, Hanover, Germany, May 2020, doi: 10.6084/m9.figshare.12275009.v1.

[25] S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J. L. Suárez, J. Luengo, M. A. Valero-González, P. García-Villanova, E. Olmedo-Sánchez, and F. Herrera, "COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 12, pp. 3595–3605, 2020.

[26] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

[27] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, C. Chute, R. Ball, J. Seekins, S. S. Halabi, R. Jones, D. B. Larson, C, P. Langlotz, B. N. Patel, and M. P. Lungren, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 590–597.
[28] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest X-ray image dataset with multi-label annotated reports," *Med. Image Anal.*, vol. 66, Dec. 2020, Art. no. 101797, doi: 10.1016/j.media.2020.101797.
[29] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002683.
[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[31] M. Horry, S. Chakraborty, B. Pradhan, M. Paul, J. Zhu, H. W. Loh, P. D. Barua, and U. R. Arharya, "Debiasing pipeline improves deep learning model generalization for X-ray based lung nodule detection," 2022, *arXiv:2201.09563*.
[32] L. O. Teixeira, R. M. Pereira, D. Bertolini, L. S. Oliveira, L. Nanni, G. D. C. Cavalcanti, and Y. M. G. Costa, "Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images," *Sensors*, vol. 21, no. 21, p. 7116, Oct. 2021.
[33] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *J. Digit. Imag.*, vol. 32, no. 4, pp. 582–596, 2019.
[34] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio, and S. K. Antani, "Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays," *IEEE Access*, vol. 8, pp. 115041–115050, 2020.
[35] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, Jun. 1947, doi: 10.1007/bf02295996.
[36] (2022). A. Murphy. *Radiographic Contrast*. [Online]. Available: https://radiopaedia.org/articles/58718
[37] J. A. Onofrey, D. I. Casetti-Dinescu, A. D. Lauritzen, S. Sarkar, R. Venkataraman, R. E. Fan, G. A. Sonn, P. C. Sprenkle, L. H. Staib, and X. Papademetris, "Generalizable multi-site training and testing of deep neural networks using image normalization," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 348–351.
[38] S. A. Mali, A. Ibrahim, H. C. Woodruff, V. Andrearczyk, H. Müller, S. Primakov, Z. Salahuddin, A. Chatterjee, and P. Lambin, "Making radiomics more reproducible across scanner and imaging protocol variations: A review of harmonization methods," *J. Personalized Med.*, vol. 11, no. 9, p. 842, Aug. 2021. [Online]. Available: https://www.mdpi.com/2075-4426/11/9/842
[39] R. M. Wehbe, J. Sheng, S. Dutta, S. Chai, A. Dravid, S. Barutcu, Y. Wu, D. R. Cantrell, N. Xiao, B. D. Allen, G. A. MacNealy, H. Savas, R. Agrawal, N. Parekh, and A. K. Katsaggelos, "DeepCOVID-XR: An artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large U.S. clinical data set," *Radiology*, vol. 299, no. 1, pp. E167–E176, Apr. 2021.
[40] (2021). IVPLatNU. *Deepcovidxr*. [Online]. Available: https://github.com/IVPLatNU/deepcovidxr
[41] R. Caruana, "Multitask Learning," *Machine Learn.*, vol. 28, pp. 41–75, Jul. 1997.
[42] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1126–1135.
[43] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (Chapelle, O. et al., Eds.; 2006)[book reviews]," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, p. 542, 2009.
[44] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2401–2410.
[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
[46] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, B. McMahan, T. Van Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," in *Proc. Mach. Learn. Syst.*, vol. 1, A. Talwalkar, V. Smith, and M. Zaharia, Eds. Stanford, CA, USA, 2019, pp. 374–388.

KAOUTAR BEN AHMED is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of South Florida. Her research interests include artificial intelligence, machine learning, data mining, and deep learning.

LAWRENCE O. HALL (Fellow, IEEE) received the B.S. degree in applied mathematics from the Florida Institute of Technology, in 1980, and the Ph.D. degree in computer science from Florida State University (FSU), in 1986. He is currently a Distinguished University Professor in computer science and engineering with the University of South Florida (USF), Tampa, FL, USA, where he co-directs the Institute for Artificial Intelligence+X. He has authored over 300 publications in journals, conferences, and books. His research interests include distributed machine learning, extreme data mining, bioinformatics, pattern recognition, and integrating AI into image processing. He has received funding from the National Institutes of Health, NASA, DOE, DARPA, and National Science Foundation. He is a fellow of AAAS, AIMBE, IAPR, and AAIA. He received the Norbert Wiener Award, in 2012, from the IEEE SMC Society and the 2021 Fuzzy Pioneer award from the IEEE Computational Intelligence Society.

DMITRY B. GOLDGOF (Fellow, IEEE) is currently an Educator and a Scientist working in the area of medical image analysis, image and video processing, computer vision and pattern recognition, ethics and computing, bioinformatics, and bioengineering. He is also a Professor and the Vice Chair of the Department of Computer Science and Engineering, University of South Florida (USF), Tampa, FL, USA. He has graduated 44 M.S. and 28 Ph.D. students, published over 95 journals, 220 conference papers, 20 books chapters; and edited five books (GS citations impact: H-index 59, G-index 114). His research interests include the area of biomedical image analysis and machine learning with application in MR, CT, PET and microscopy images, radiomics and bioinformatics, area of video motion analysis with biometrics, face analysis, surveillance, and biomedical applications. He is a fellow of IAPR, a fellow of AAAS, and a fellow of AIMBE.

RYAN FOGARTY (Graduate Student Member, IEEE) received the B.S.E.E. degree from Florida State University, in 1996, and the M.S. degree in ECE from the University of Florida, in 1998. He is currently pursuing the Ph.D. degree with the University of South Florida. He was formerly a Principal Engineer working in array signal processing for sonar and radar embedded systems, in which he worked for over 20 year's. His current research is supported through an internship at the Moffitt Cancer Center, Tampa, FL, USA, investigating machine learning for radiomics and histopathology.