Visualizing and explaining deep learning predictions for pneumonia detection in pediatric chest radiographs

Sivaramakrishnan Rajaraman, Sema Candemir, George Thoma, Sameer Antani

Sivaramakrishnan Rajaraman, Sema Candemir, George Thoma, Sameer Antani, "Visualizing and explaining deep learning predictions for pneumonia detection in pediatric chest radiographs," Proc. SPIE 10950, Medical Imaging 2019: Computer-Aided Diagnosis, 109500S (13 March 2019); doi: 10.1117/12.2512752

SPIE.

Event: SPIE Medical Imaging, 2019, San Diego, California, United States

Visualizing and explaining deep learning predictions for pneumonia detection in pediatric chest radiographs

Sivaramakrishnan Rajaraman[*a], Sema Candemir[b], George Thoma[a], Sameer Antani[a]
[a]Lister Hill National Center for Biomedical Communications, National Library of Medicine,
National Institutes of Health, 8600 Rockville Pike, Bethesda, MD, 20894-0001, USA; [b]The Ohio
State University Wexner Medical Center, 410 West 10th Avenue, Columbus, OH, 43210, USA

ABSTRACT

Pneumonia is a severe inflammatory condition of the lungs that leads to the formation of pus and other liquids in the air sacs. The disease is reported to affect approximately 450 million people across the world, resulting in 2 million pediatric deaths every year. Chest X-ray (CXR) analysis is the most frequently performed radiographic examination for diagnosing the disease. Unlike pneumonia in adults, pediatric pneumonia is poorly studied. Computer-aided diagnostic (CADx) tools aim to improve disease diagnosis and supplement decision making while simultaneously bridging the gap in effective radiological interpretations during mobile field screening. These tools make use of handcrafted and/or convolutional neural networks (CNN) extracted image features for visual recognition. However, CNNs are perceived as black boxes since their performance lack explanations and poorly understood. The lack of transparency in the learned behavior of CNNs is a serious bottleneck in medical screening/diagnosis since poorly interpreted model behavior could unfavorably impact decision-making. Visualization tools are proposed to interpret and explain model predictions. In this study, we highlight the advantages of visualizing and explaining the activations and predictions of CNNs applied to the challenge of pneumonia detection in pediatric chest radiographs. We evaluate and statistically validate the models' performance to reduce bias, overfitting, and generalization errors.

Keywords: Deep learning, computer-aided diagnosis, pneumonia, convolutional neural networks, visualization, explanation, chest radiographs, decision-making

1. INTRODUCTION

Pneumonia is a significant cause of mortality in pediatrics across the world. According to the World Health Organization (WHO), around 2 million pneumonia-related deaths are reported every year in children under 5 years of age, making it the most significant cause of pediatric death[1]. Nearly 95% of the community-acquired pediatric pneumonia occurs in Africa and Southeast Asia. The disease is caused by bacterial and viral pathogens[2] that require different forms of management. Bacterial pneumonia is treated immediately with antibiotics while that caused by the viral pathogens, with supportive care. Timely diagnosis and treatment are highly indispensable for short- and long-term health outcomes. Chest X-ray (CXR) analysis is the most frequently performed radiographic examination for diagnosing and differentiating the disease[3,4]. Fig. 1 shows the instances of normal, bacterial and viral pneumonia infected CXRs. Rapid radiographic diagnoses and treatment are adversely impacted by the lack of expert radiologists in resource-constrained regions where pediatric pneumonia is highly endemic with alarming mortality rates. There are occasions when the radiologists fail to appreciate normal variations and influence of technical factors on the appearance of CXRs. This leads to inter/intra-observer variability and poses a serious threat to reliable interpretation.

Computer-aided diagnostic (CADx) tools aim to supplement medical decision making and reduce expert intervention in screening/diagnosis, particularly in disease-endemic regions with resource-constrained settings[5]. A vast majority of these tools make use of machine learning (ML) algorithms that employ handcrafted features for clinical decision-making[6]. However, these features are extracted with rule-based feature descriptors, the performance depends on the human expertise in developing algorithms to account for the variability in the morphology and position of the region of interest (ROI) and is computationally intensive. The process is adversely impacted by the restricted visibility to the degree of variability in the data during large-scale diagnoses[7]. On the other hand, data-driven approaches including deep learning (DL) overcome these limitations by self-discovering/learning hierarchical feature representation from the raw input pixels. DL models have revolutionized screening and diagnosis in medical visual recognition tasks. In convolutional neural network (CNN)

Medical Imaging 2019: Computer-Aided Diagnosis, edited by Kensaku Mori,
Horst K. Hahn, Proc. of SPIE Vol. 10950, 109500S · © 2019 SPIE
CCC code: 1605-7422/19/$18 · doi: 10.1117/12.2512752

based DL models, the lower-level features are abstracted to form higher-level features toward the process of learning complex, non-linear functions. This results in "end-to-end" feature extraction and classification that avoids the intricacies of using traditional feature extraction/classification methods[8]. Unlike kernel algorithms, the performance of CNNs scale with data and computational resources. CNNs are shown to be remarkable in object localization and detection tasks to signify the contribution of individual pixels toward decision-making[9].

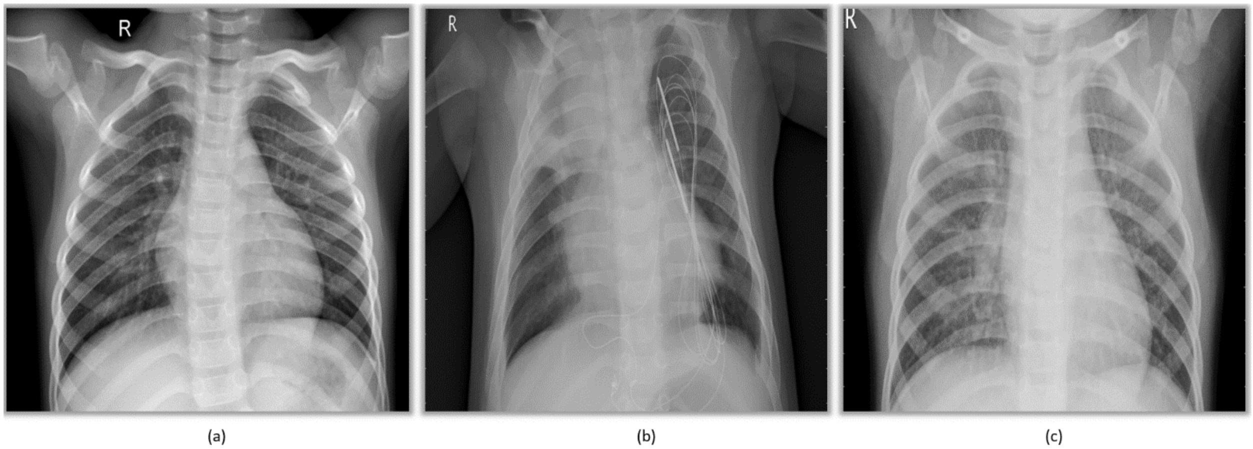

Figure 1. Pediatric CXRs. (a) Normal CXR showing clear lungs with no abnormal opacification, (b) Bacterial pneumonia exhibiting focal lobar consolidation in the right upper lobe and (c) Viral pneumonia manifesting with diffuse interstitial patterns in both lungs.

CNNs are shown to deliver superior results under the availability of a huge amount of annotated data. Under circumstances of sparse annotated data availability as with medical images, researchers use transfer learning methods that employ pre-trained CNNs to transfer the knowledge learned in the form of generic image features from large-scale datasets like ImageNet to the current task[10]. The transfer of knowledge is rather general than being task-specific. The pre-trained weights are either fine-tuned on the current task or the models are used as fixed feature extractors toward visual recognition[11].

Study of the literature reveals several works pertaining to the usage of CADx tools that employ handcrafted and/or CNN extracted features toward the challenge of pneumonia detection in CXRs. The authors of[12] developed a CAD system named PneumoCAD that uses handcrafted features to diagnose pediatric pneumonia. An AUC of 0.97 and 0.94 and a specificity of 80% and 90% is reported through different methods of weighting image classification. The authors of[5] released the National Institutes of Health (NIH) CXR dataset that contains 112,120 frontal CXRs, individually labeled to include up to 14 different diseases. The disease labels are text-mined from the radiological reports using natural language processing tools. The authors performed a multi-label image classification together with ROI localization to detect and spatially locate the disease. An AUC of 0.6333 is reported for pneumonia detection. In another study, the authors[13] proposed a 121-layer densely connected neural network to output the probability of pneumonia and localize the ROI with heat maps. The model is trained on the NIH CXR dataset and its performance is compared with that of four practicing radiologists, on a test set of 420 CXRs. It is observed that the model exceeded the average performance of the radiologists in the F1-metric. An AUC of 0.7680 is reported in detecting pneumonia. The authors of[14] used a pre-trained Inception-V3 network as a fixed feature extractor toward classifying normal and pneumonia and further distinguishing bacterial and viral pneumonia in pediatric chest radiographs. An accuracy of 92.8%, recall of 93.2%, specificity of 90.1%, and AUC of 0.968 is reported in classifying normal and pneumonia. An accuracy of 90.7%, recall of 88.6%, specificity of 90.9%, and AUC of 0.940 is reported in distinguishing bacterial and viral pneumonia in CXRs.

Despite the promising performance of CNNs in visual recognition, there is a lack of clarity on their learned behavior and predictions. While the literature discusses several methods for classification, very few researchers provide an explanation of the model predictions and/or validate how the performance is achieved[15,16]. Exploratory studies need to be performed in qualitatively understanding and interpreting the model performance and suggesting possible improvements. The lack of transparency in the learned representations of CNNs is a serious issue in medical applications since poorly understood behavior could adversely impact decision-making[17]. A comprehensive understanding and interpretation of the architecture, the internal operations and the learned behavior assists in trusting and explaining the predictions that indicate disease/abnormality. It is sensible to mention that the current literature leaves much room for progress in these aspects.

The unresolved issue of visualizing, understanding and explaining the predictions of CNNs applied to the challenge of pneumonia detection in pediatric CXRs is principally relevant and is the rationale behind this study.

In this study, we visualize the activations and explain the predictions of CNNs applied to the challenge of classifying normal and pneumonia infected CXRs and further differentiate bacterial and viral pneumonia to facilitate rapid referrals that require urgent medical intervention. We demonstrate that in the process of using the optimal model architecture and parameters for the underlying task, a trustworthy model can be built and the predictions are explained toward discriminating the classes. We evaluate and statistically validate the performance of an optimized, custom CNN and a pretrained VGG16[18] model to provide an accurate and timely diagnosis of the pathology. The study is organized as follows: Section 2 elaborates on the materials and methods, Section 3 discusses the results and Section 4 concludes the study.

2. MATERIALS AND METHODS

2.1 Data Collection and Preprocessing

The pediatric CXRs used in this study are made publicly available by the authors of[14]. The dataset includes anteroposterior chest radiographs collected from pediatrics of 1 to 5 years of age from Guangzhou women and children's medical center, Guangzhou. The characteristics of the dataset are mentioned in Table 1. Radiological imaging is performed as part of patients' clinical care. The radiographs are interpreted to confirm a diagnosis and referral decisions made thereafter. Bacterial pneumonia is referred for urgent antibiotic treatment while viral pneumonia is treated with supportive care. All CXRs are screened for quality control by removing unreadable scans. Training and testing data are graded by expert physicians to account for grading errors.

Table 1. Dataset and their characteristics.

Category	# Samples	# Training samples	#Test samples	Type	Depth
Normal	1583	1349	234	JPG	8-bit
Bacterial Pneumonia	2780	2538	242	JPG	8-bit
viral pneumonia	1493	1345	148	JPG	8-bit

The CXRs contain regions other than the lungs that are irrelevant toward pneumonia detection. The lung boundaries are detected using our atlas-based lung boundary detection algorithm[19,20]. It uses a reference set of radiographs with expert delineated lung boundaries as models to register with the objective CXRs. Fig. 2 shows the detected boundaries for the sample CXRs from the dataset under study.

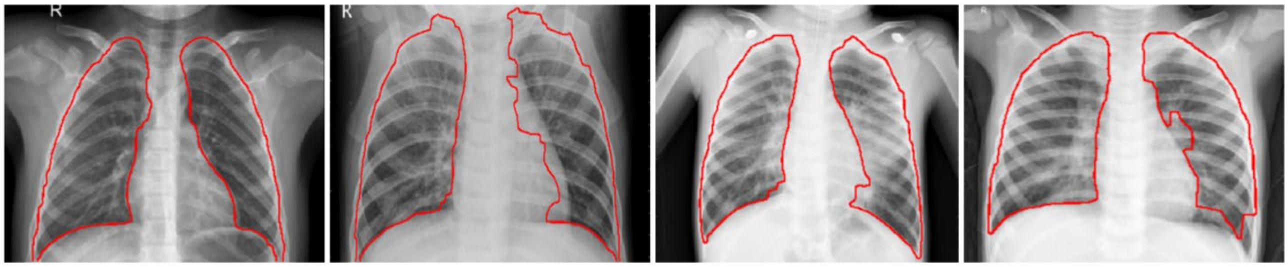

Figure 2. Detected boundaries in example pediatric CXRs.

As models, we used pediatric chest radiographs and their corresponding lung masks that were collected and studied in[20]. When a patient CXR is presented, the algorithm finds the most similar CXRs in the reference model set to the objective CXR. The similarity between the CXRs is measured by comparing their horizontal and vertical intensity histograms, which are rough shape models of the lung blobs. As mentioned in[19,20], we used the Bhattacharyya distance as the similarity measure. The main purpose of measuring the similarity is to increase the correspondence performance and decrease the computational expense during registration. After model selection, the algorithm computes a correspondence map between the model CXRs and the objective chest radiograph using local image features and finds the most similar locations with the SIFT-flow algorithm[21]. This map is the transformation mapping applied to the model masks to morph them into the approximate lung model for the objective CXR. The segmented images are cropped to the size of a bounding box that includes the lung pixels. The cropped lung ROI is resampled to 1024×1024 pixel dimensions and mean normalized to assist the models in faster convergence. We used the NIH Biowulf Linux cluster and the NVIDIA DGX-1 facility available at the National Library of Medicine (NLM), Keras® with Tensorflow® backend, Matlab R2017b® and CUDA 8.0/cuDNN 5.1 dependencies for GPU acceleration.

2.2. Model Configuration

We evaluated the performance of a custom CNN and pre-trained VGG16 network to classify normal and pneumonia and further differentiate between bacterial and viral pneumonia in CXRs. Selecting and tuning the custom CNN model parameters is extremely complex and computationally expensive. We performed Bayesian optimization[22] to search for the optimal model parameters and training options for the underlying task. The process helps in optimizing non-differentiable, discontinuous and computationally expensive functions by internally using a Gaussian process model of an objective function and its evaluation in model training. The optimization variables used in this study include network depth, initial learning rate, momentum, and L2 regularization. The number of filters in the convolutional layer is increased by a factor of two when the spatial dimensions are down-sampled with max-pooling layers to ensure roughly the same number of computations. The convolutional layer is followed by a rectified linear unit (ReLU) layer to introduce non-linearity into the computations and prevent vanishing gradients[8]. The search ranges for the network depth, initial learning rate, momentum, and L2 regularization are set as [1 9], [1e-5 1e-1], [0.8 0.95] and [1e-10 1e-1] respectively. An objective function is defined to take the optimization variables as inputs, train, validate, and save the optimal network. The number of objective function evaluations is specified. Optimization is achieved by minimizing the classification error on the test set. We also evaluated the performance of a pre-trained VGG16 network for the underlying task. The architecture and pre-trained weights are downloaded from the GitHub repository[23]. The model is truncated in the deepest convolutional layer. A global average pooling (GAP) and a logistic layer are added on top of the truncated model as shown in Fig. 3. The model is optimized for hyper-parameters by a randomized grid search method[24]. The search ranges for the learning rate, SGD momentum and L2-regularization parameters are set to [1e-3 5e-2], [0.8 0.95] and [1e-10 1e-1] respectively. The pre-trained model is trained end-to-end with very small weight updates. Hold-out testing is performed after every step using the test set independent of the training set by passing each test image through the model without performing backpropagation and gradient descent. Callbacks are used to view the internal states and statistics during the training phase and the best performing model is kept for analysis. The performance of the customized and pre-trained CNNs is evaluated in terms of accuracy, area under the receiver operating characteristic (ROC) curve (AUC), precision, recall, specificity, F1-metric, and Matthews correlation coefficient (MCC).

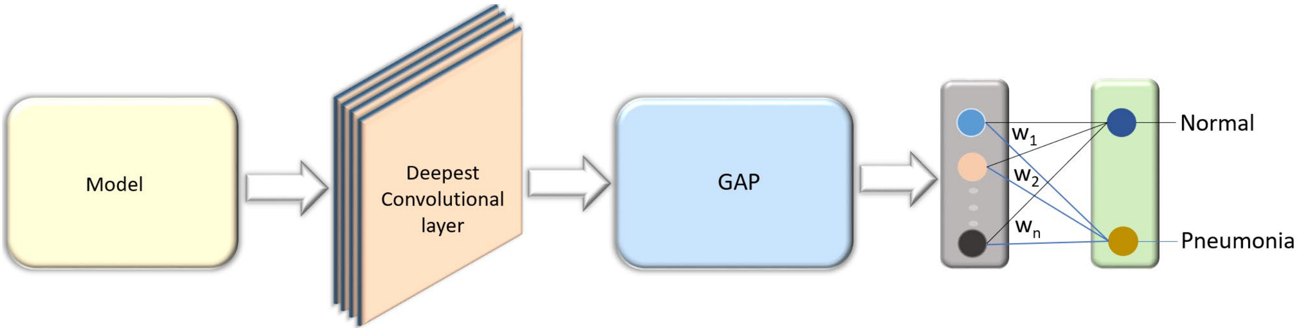

Figure 3. VGG16 network truncated at the deepest convolutional layer and added with a GAP and dense layer.

2.3. Visualizing the network layer activations

The learned behavior of the trained model is understood by visualizing its activations to gain a comprehensive knowledge of the input patterns that activate the filters in a given convolutional layer[9]. The learned weights have high interpretability in the earlier layers than in the deeper layers. The network layer activations of an optimally trained model are compared with the input image to interpret the learned features toward classifying normal and pneumonia and to further distinguish bacterial and viral pneumonia in chest radiographs.

2.4. Visual explanations through gradient weighted class activation mapping

It is imperative to understand the parts of the image the model looked in to arrive at the predictions. Gradient weighted class activation maps (grad-CAM)[25] helps to visualize and debug the predictions of the CNN particularly in the case of a classification error when the network arrives at a decision based on the surrounding context. This helps in answering questions pertaining to the model's ability to categorize and localize the ROI in an image specific to the expected class. CXRs are fed to the trained model and the output feature map of the deepest convolutional layer is recorded. The gradient of the expected class with respect to the output feature map of this convolutional layer is computed and pooled to compute the mean intensity of the gradient over the feature map channels. Each channel in the feature map array is multiplied by the weights with respect to the expected class. The heat map of the class activation is obtained by the channel-wise mean of the resultant feature map. The generated heat map is superimposed on the input CXR to localize the ROI with respect to the expected class. The localized regions explain how the model answers the difference between the classes.

2.5. Local Interpretable Model-Agnostic Explanations (LIME)

Training metrics can be misleading since the data may accidentally leak into the held-out data and/or the model predicts based on the surrounding context and not the ROI. In ML applications, particularly in medical screening/diagnosis, a measure of trust is often necessitated prior to deployment. Despite the fact that DL models are perceived as black-boxes, interpreting the rationale behind the predictions helps to decide their trustworthiness. An "explainer" is needed to explain the predictions to highlight the features that are most relevant to decision-making and crucial for effective human-computer interactions. Local Interpretable Model-Agnostic Explanations (LIME) is an effective visualization tool that explains the predictions and evaluates the usefulness of the models in decision-making[26].

Being model-agnostic, LIME provides a qualitative understanding of the relationship between the interpretable components and predictions. The image is divided into contiguous superpixels, a dataset of perturbed instances is generated by turning on/off these interpretable components. The perturbed images are weighted by their similarity to the explained instance. An explanation is generated by approximating the complex, non-linear model by a linear one weighted in the neighborhood of the prediction to be explained. The superpixels with the highest positive weights are finally presented as an explanation. The process helps to select the model with reduced generalization errors, improve model performance by optimizing its parameters and gain crucial insights into its behavior.

Let $y \in \mathbb{R}^d$ be the original instance to be explained. Let $y' \in \{0, 1\}^d$ be the binary vector for its interpretable representation that denotes the presence/absence of a superpixel. Let $h \in H$ denote the model explanation where H denotes a class of potentially interpretable linear models. Let $\Omega(h)$ denote the measure of the complexity of the explanation $h \in H$. For a linear model, $\Omega(h)$ denotes the number of non-zero weights. Let $g: \mathbb{R}^d \rightarrow \mathbb{R}$ denote the model to be explained and g(y), the probability that y belongs to a certain class. Let $\Pi_y(z)$ denote the proximity measure between the instance z to y and $L(g, h, \Pi_y)$, a measure of low fidelity of h in approximating g in the locality defined by Π_y. The value $L(g, h, \Pi_y)$ is minimized to ensure interpretability. The number of non-zero coefficients $\Omega(h)$ remains low enough to be interpretable. The explanations produced are given by,

$$\mathbb{y}(y) = \underset{\mathrm{h} \in \mathrm{H}}{\operatorname{argmin}} L\big(g, h, \Pi_y\big) + \Omega(h) \tag{1}$$

The value $L(g, h, \Pi_y)$ is approximated by drawing samples weighted by Π_y. Instances are sampled around y by drawing non-zero elements of y' uniformly at random. Given a perturbed sample $a' \in \{0, 1\}^{d'}$ that contain a fraction of non-zero elements of y', the sample is recovered in the original representation $a \in \mathbb{R}^d$ to obtain g(a) that is used as a label for the explanation model given by,

$$\Pi_y^a = \exp(-D\ \frac{(y,a)^2}{\sigma^2}) \tag{2}$$

An exponential kernel defined on the L2-distance function (D) with width σ is given by,

$$L\big(g, h, \Pi_y\big) = \sum_{a,a' \in A} \Pi_y^a\ (g(a) - h(a'))^2 \tag{3}$$

LIME is model-agnostic, leverages simple concepts, helps to decide the trustworthiness of predictions, improve the model performance with parameter optimization and identify why the model should not be trusted.

3. RESULTS AND DISCUSSIONS

3.1. Bayesian optimization applied to custom CNN

The objective function takes as input, the optimization variables including network depth, initial learning rate, momentum, and L2 regularization, trains, validates, and saves the optimized model. The number of objective function evaluations is set to 100. Optimization is achieved by minimizing the classification error on the test set. The optimal values for the network depth, initial learning rate, momentum, and L2 regularization are found to be 6, 1e-3, 0.9 and 1e-6 respectively. The configuration of the optimized custom CNN is shown in Fig. 4.

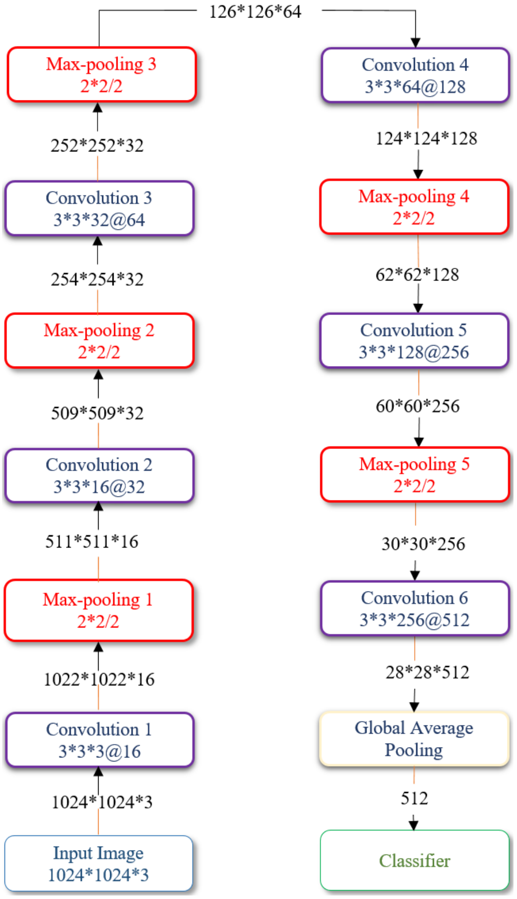

Figure 4. The optimized architecture of custom CNN.

3.2. Performance Evaluation

The performance of optimized custom CNN and pretrained VGG16 network are evaluated as two separate tasks: (i) classifying normal and pneumonia; and, (ii) distinguishing bacterial and viral pneumonia in pediatric CXRs. We evaluated the models with the baseline data and cropped lung ROI from the CXRs. In classifying normal and pneumonia, we found that the performance of custom CNN is better with the cropped ROI in terms of AUC, precision, and specificity. In differentiating bacterial and viral pneumonia infected CXRs, no significant difference in performance is observed between the baseline data and cropped ROI, except for AUC where cropped ROI gave better results. Table 2 shows the performance of the optimized custom CNN with respect to the baseline and cropped lung ROI. Similar trends are observed in evaluating the performance of a pretrained VGG16 network that is trained end-to-end for the current task. No significant difference in performance is observed with the baseline data and cropped ROI in classifying bacterial and viral pneumonia. However, cropped ROI gives better results than the baseline in terms of all performance metrics except for recall, in classifying normal and pneumonia infected CXRs. Table 3 shows the performance of the pretrained VGG16 network with the baseline and cropped lung ROI. The pretrained VGG16 network outperforms the optimized custom CNN in all performance metrics except for recall, in both the classification tasks. This may be due to the fact that the availability of a fewer number of training samples didn't give enough opportunity for the custom CNN to learn discriminative features across the classes. The pretrained VGG16 network learned generic image features from the large-scale ImageNet dataset to transfer to the current task. Compared to random weight initialization, the pretrained ImageNet weights served as a good initialization that is fine-tuned end-to-end on the current task to assist in earlier convergence, reduced overfitting, bias, and generalization errors.

Table 2. Performance of custom CNN with baseline and cropped lung ROI.

Data	Task	Accuracy	AUC	Precision	Recall	F1	Specificity	MCC
Baseline	Normal v. Pneumonia	0.943	0.983	0.920	0.995	0.957	0.855	0.878
	Bacterial v. Viral Pneumonia	0.928	0.954	0.909	0.984	0.946	0.838	0.848
Cropped Lung ROI	Normal v. Pneumonia	0.941	0.984	0.930	0.980	0.955	0.877	0.873
	Bacterial v. Viral Pneumonia	0.928	0.956	0.909	0.984	0.946	0.838	0.848

Table 3. Performance of pretrained VGG16 network with baseline and cropped lung ROI.

Data	Task	Accuracy	AUC	Precision	Recall	F1	Specificity	MCC
Baseline	Normal v. Pneumonia	0.957	0.990	0.951	0.983	0.967	0.915	0.908
	Bacterial v. Viral Pneumonia	0.936	0.962	0.920	0.984	0.951	0.860	0.862
Cropped Lung ROI	Normal v. Pneumonia	0.962	0.993	0.977	0.962	0.970	0.962	0.918
	Bacterial v. Viral Pneumonia	0.936	0.954	0.920	0.984	0.951	0.860	0.862

Fig. 5 shows the confusion matrices for the performance of the pretrained VGG16 network with the cropped lung ROI and Fig. 6 shows the AUC achieved for the corresponding tasks. The lack of significant difference in performance of the models with the baseline and cropped ROI may be attributed to the fact that the baseline data already appeared as cropped, except for a few training and testing instances. The lung segmentation algorithm resulted in under-segmentation in a few instances near the costophrenic angle. We observed that the test accuracy and loss are better than the training metrics since noisy, low-quality images are included in the training set to reduce bias, overfitting and generalization errors. The CXRs are fed to the VGG16 network and the filter activations are compared with the corresponding input pixel locations. Strong positive activations are visible as white pixels and negative activations as black pixels. Fig. 7 shows a random selection of activations observed in the filters of different convolutional layers in the VGG16 network when presented with an input CXR. We observed that the earlier network layers act as a collection of various edge detectors and the activations retain almost all information present in the input image. The activations begin to encode higher level features, become less interpretable and more abstract with increasing depth. These higher representations carry increasingly less representation about the visual contents of the image and more information pertaining to the expected class. The sparsity of activations is also found to increase with the layer depth. In the earlier layers, all filters are activated by the input image wherein the deeper layers, more filters are blank that infer that the patterns encoded by these filters are not found in the input image. The CNN acts as an information distillation pipeline in which the input pixels are transformed to filter out irrelevant information and useful characteristics pertaining to the class of the image is magnified and retained. CXRs are fed to the VGG16 network and its predictions are decoded. When presented with a CXR infected with pneumonia, grad-CAM generates a heat map for the expected pneumonia class that indicates the visual differences in the "pneumonia-like" parts of the image. The heat map is generated as a two-dimensional grid of scores, computed for each pixel location in the input image to indicate the importance of the location with respect to the expected class. The generated heat map is superimposed on the input CXR to localize the ROI with respect to the pneumonia class. The process helps in locating the spread of the disease and also explains why the model considered the CXR to belong to the pneumonia class.

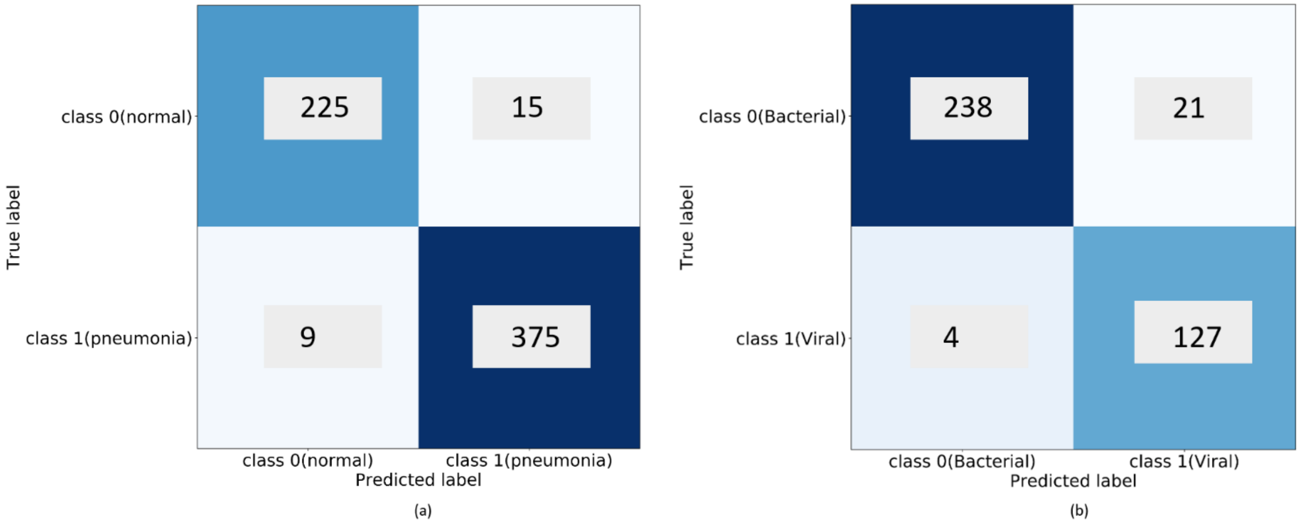

Figure 5. Confusion matrices for the performance of pretrained VGG16 network on the classification tasks. (a) Normal/pneumonia and (b) Bacterial/viral pneumonia.

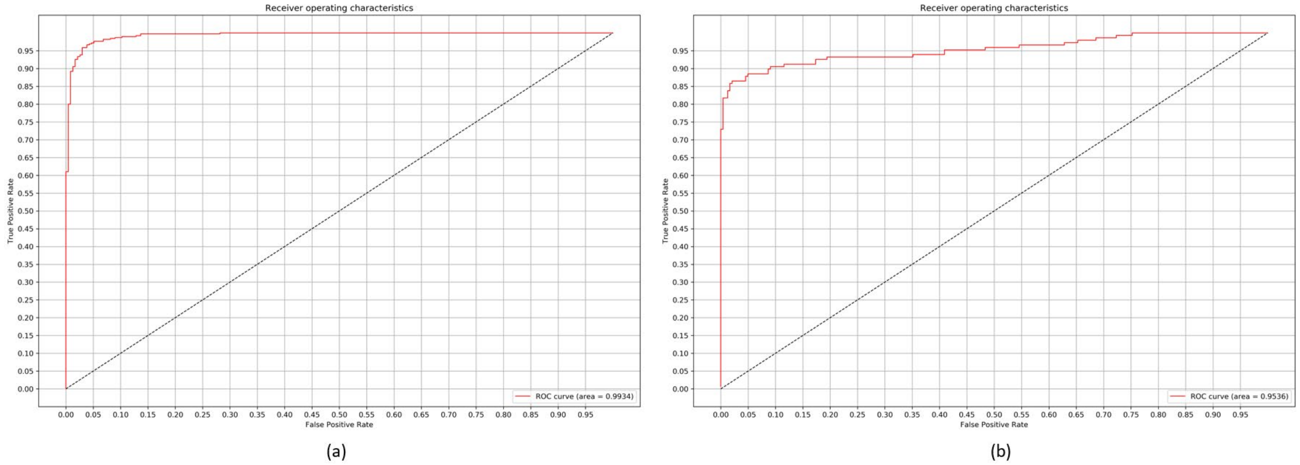

Figure 6. AUC for the performance of pretrained VGG16 network on the classification tasks. (a) Normal/pneumonia and b) Bacterial/viral pneumonia.

Fig. 8 shows several instances of pneumonia infected CXRs, their corresponding grad-CAM, and LIME outputs. The lung masks are applied to the original images to extract the lung regions. It is interesting to note that the regions of high opacity are strongly activated which probably made the network to classify the image to belong to the pneumonia class. The explanations generated by the LIME explainer are shown as the superpixels with the highest positive weights, superimposed on the original input. The explainer shows that the model is forcing on the regions of high opacity. There are also a few false positive superpixels reported.

The current implementation uses linear models to approximate local behavior. The assumption holds well when looking into the neighborhood of prediction but not powerful enough to explain the behavior of the original, non-linear model. The explanation may not be faithful if the underlying model exhibits a high degree of non-linearity in the locality of predictions. It may be for this reason that we could observe a number of false positive superpixels been reported in the explanations. These explanations result from a random sampling process, the exact same explanation could not be expected. In these circumstances, the number of samples is increased to improve the confidence in the explanation.

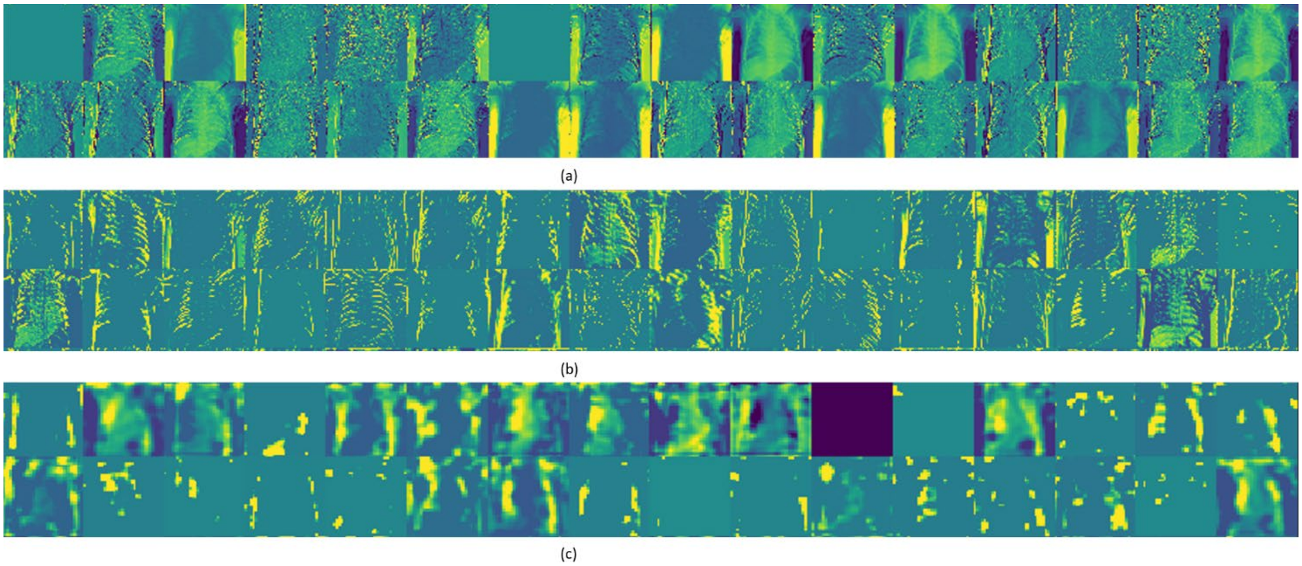

Figure 7. Visualizing a random selection of filter activations in the VGG16 model. (a) first convolutional layer, (b) eighth convolutional layer, and (c) deepest convolutional layer.

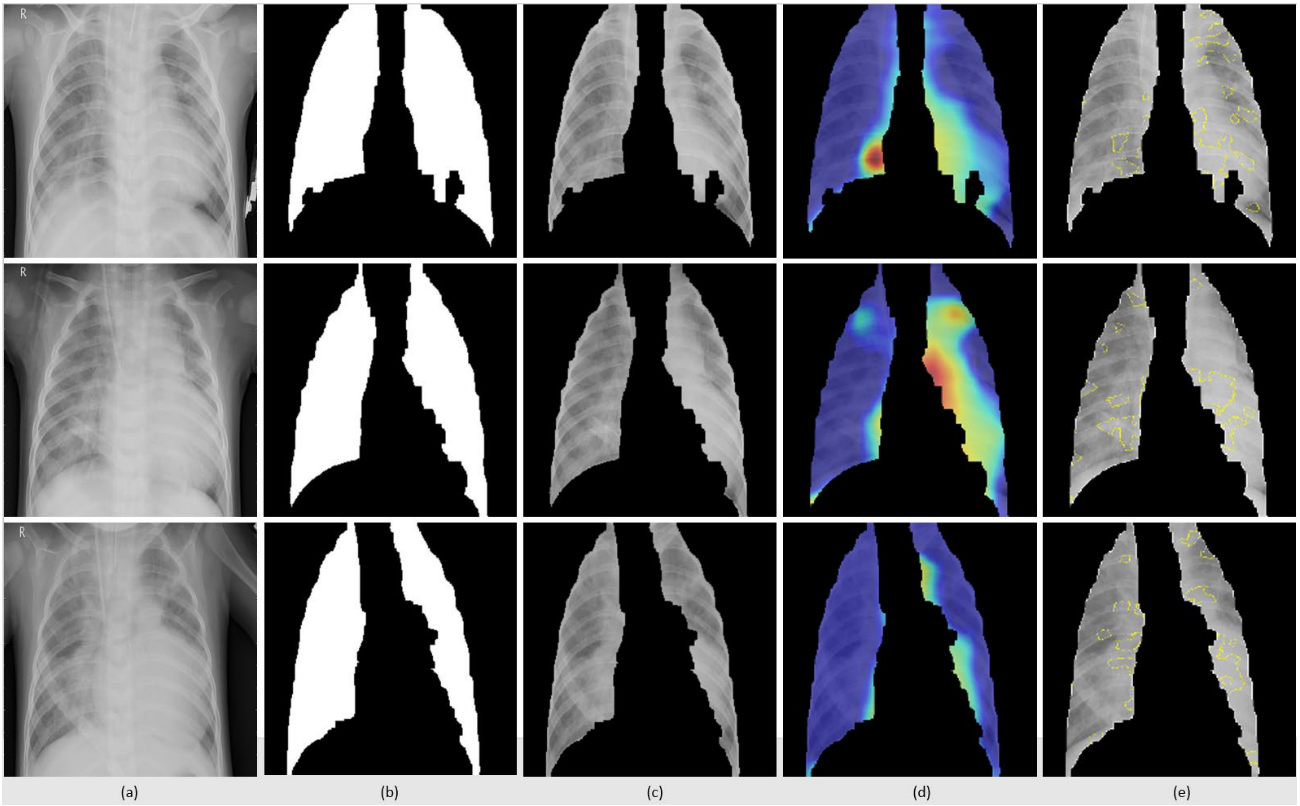

Figure 8. Visual explanations through class activation maps and LIME. (a) original image, (b) lung mask, (c) cropped lung regions, (d) grad-CAM visualization and e) LIME visualization.

The results obtained with the cropped lung ROI are compared to the state-of-the-art as shown in Table 4. We observed that the pretrained VGG16 network trained end-to-end on the cropped lung ROI outperformed the state-of-the-art in all performance metrics in classifying normal and pneumonia and further differentiating bacterial and viral pneumonia. However, the optimized custom CNN demonstrated better results for recall. If a model has to be selected considering the balance between precision and recall as demonstrated by the F1 metric and MCC, the pretrained VGG16 outperformed the custom CNN and the state-of-the-art in the classification tasks.

Table 4. Comparing the classification performance with the state-of-the-art.

Task	Model	Accuracy	AUC	Precision	Recall	specificity	F1	MCC
Normal v. Pneumonia	Pre-trained VGG16	**0.962**	**0.993**	**0.977**	0.962	**0.962**	**0.970**	**0.918**
	Custom CNN	0.941	0.984	0.930	**0.980**	0.877	0.955	0.873
	Kermany et al.[14]	0.928	0.968	-	0.932	0.901	-	-
Bacterial v. Viral Pneumonia	Pre-trained VGG16	**0.936**	0.954	**0.919**	**0.984**	**0.859**	**0.951**	**0.862**
	Custom CNN	0.928	**0.956**	0.909	**0.984**	0.838	0.946	0.848
	Kermany et al.[14]	0.907	0.940	-	0.886	0.909	-	-

*Bold numbers indicate superior performance.

4. CONCLUSION

We proposed a DL based AI system to diagnose pneumonia in pediatric CXRs to expedite diagnosis, facilitate early treatment and improve clinical decision-making. We also explained the model predictions through ROI localization that is most relevant to decision-making and crucial for effective human-computer interactions. The study presents a generalized platform to apply to an extensive range of medical applications. Classifying CXRs is a difficult task due to the presence of a large number of variable objects that are extraneous to pneumonia diagnosis. The promising performance of the pretrained VGG16 network trained end-to-end on the pediatric pneumonia dataset suggests that the pretrained model effectively learns from progressively complicated data with reduced bias and generalization errors using a relatively small data collection. The transfer learning mechanism can further be explored and analyzed for other biomedical imaging applications including screening/diagnosis of common diseases.

ACKNOWLEDGMENTS

This work is supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

CONFLICT OF INTEREST

The authors have no conflict of interest to report.

REFERENCES

[1] Rudan, I., Boschi-Pinto, C., Biloglav, Z., Mulholland, K. and Campbell, H., "Epidemiology and etiology of childhood pneumonia," Bull. World Health Organ. 86(5), 408-416 (2008).

[2] Virkki, R., Rikalainen, H., Svedström, E., Juven, T., Mertsola, J. and Ruuskanen, O., "Differentiation of bacterial and viral pneumonia in children," Thorax 57(5), 438-441 (2002).

[3] Cherian, T., Mulholland, E. K., Carlin, J. B., Ostensen, H., Amin, R., De Campo, M., Greenberg, D., Lagos, R., Lucero, M., Madhi, S. A., O'Brien, K. L., Obaro, S. and Steinhoff, M. C., "Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies," Bull. World Health Organ. 83(5), 353-359 (2005).

[4] Sivaramakrishnan, R., Antani, S., Candemir, S., Xue, Z., Abuya, J., Kohli, M., Alderson, P. and Thoma, G., "Comparing deep learning models for population screening using chest radiography," Proc. SPIE 10575, 105751E (2018).

[5] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. and Summers, R. M., "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," Proc. IEEE CVPR, 3462-3471 (2017).

[6] Goldbaum, M., Moezzi, S., Taylor, A., Chatterjee, S., Boyd, J., Hunter, E. and Jain, R., "Automated Diagnosis and Image Understanding with Object Extraction, Object Classification, and Inferencing in Retinal Images," Proc. IEEE ICIP, 695-698 (1996).

[7] Neuman, M. I., Lee, E. Y., Bixby, S., Diperna, S., Hellinger, J., Markowitz, R., Servaes, S., Monuteaux, M. C. and Shah, S. S., "Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children," J. Hosp. Med. 7, 294-298 (2012).

[8] Krizhevsky, A., Sutskever, I. and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," Commun. ACM 60(6), 84-90 (2017).

[9] Zeiler, M. D. and Fergus, R., "Visualizing and understanding convolutional networks," Proc. ECCV, 818-833 (2014).

[10] Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. and Summers, R. M., "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," IEEE Trans. Med. Imaging 35(5), 1285-1298 (2016).

[11] Rajaraman, S., Candemir, S., Xue, Z., Alderson, P. O., Kohli, M., Abuya, J., Thoma, G. R., Antani, S. and Member, S., "A novel stacked generalization of models for improved TB detection in chest radiographs," Proc. IEEE EMBC, 718-721 (2018).

[12] Oliveira, L. L. G., Silva, S. A., Ribeiro, L. H. V., de Oliveira, R. M., Coelho, C. J. and Ana Lúcia, A. L. S., "Computer-aided diagnosis in chest radiography for detection of childhood pneumonia," Int. J. Med. Inform. 77, 555-564 (2008).

[13] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P. and Ng, A. Y., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning" CoRR abs/1711.05225 (2017).

[14] Kermany, D. S. et al., "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," Cell 172(5), 1122-1124.e9 (2018).

[15] Rajaraman, S., Antani, S. K., Poostchi, M., Silamut, K., Hossain, M. A., Maude, R. J., Jaeger, S. and Thoma, G. R., "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," PeerJ 6, e4568 (2018).

[16] Rajaraman, S., Silamut, K., Hossain, M. A., Ersoy, I., Maude, R. J., Jaeger, S., Thoma, G. R. and Antani, S. K., "Understanding the learned behavior of customized convolutional neural networks toward malaria parasite detection in thin blood smear images," J. Med. Imaging 5(3), 34501-34511 (2018).

[17] Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E. and Greenspan, H., "Chest pathology detection using deep learning with non-medical training," Proc. IEEE ISBI, 294-297 (2015).

[18] Simonyan, K. and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," CoRR abs/1409.1556 (2014).

[19] Candemir, S., Jaeger, S., Palaniappan, K., Musco, J. P., Singh, R. K., Xue, Z., Karargyris, A., Antani, S., Thoma, G. and McDonald, C. J., "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," IEEE Trans. Med. Imaging 33(2), 577–590 (2014).

[20] Candemir, S., Antani, S., Jaeger, S., Browning, R. and Thoma, G., "Lung Boundary Detection in Pediatric Chest X-rays," Proc. SPIE 9418, 94180Q (2015).

[21] Liu, C., Yuen, J. and Torralba, A., "Sift flow: Dense correspondence across scenes and its applications," IEEE TPAMI 33(5), 978-994 (2011).

[22] Snoek, J., Rippel, O. and Adams, R. P., "Scalable Bayesian Optimization Using Deep Neural Networks," Proc. ICML, 2171-2180 (2015).

[23] Chollet, F., "Building powerful image classification models using very little data," Keras Blog, 2016, <https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html> (4 January 2017).

[24] Bergstra, J. and Bengio, Y., "Random Search for Hyper-Parameter Optimization," J. Mach. Learn. Res. 13, 281-305 (2012).

[25] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Proc. IEEE ICCV, 618-626 (2017).

[26] Ribeiro, M. T., Singh, S. and Guestrin, C., "Why Should I Trust You? Explaining the Predictions of Any Classifier," Proc. ACM KDD, 1135-1144 (2016).