Discovering Unknown Diseases with Explainable Automated Medical Imaging

Claire Tang(✉)

Lynbrook High School, San Jose, CA, USA
claire.tang2021@gmail.com

Abstract. Deep neural network (DNN) classifiers have attained remarkable performance in diagnosing known diseases when the models are trained on a large amount of data from known diseases. However, DNN classifiers trained on known diseases usually fail when they confront new diseases such as COVID-19. In this paper, we propose a new deep learning framework and pipeline for explainable medical imaging that can classify known diseases as well as detect new/unknown diseases when the models are only trained on known disease images. We first provide in-depth mathematical analysis to explain the overconfidence phenomena and present the calibrated confidence that can mitigate the overconfidence. Using calibrated confidence, we design a decision engine to determine if a medical image belongs to some known diseases or a new disease. At last, we introduce a new visual explanation to further reveal the suspected region inside each image. Using both Skin Lesion and Chest X-Ray datasets, we validate that our framework significantly improves the accuracy of new disease discovery, i.e., distinguish COVID-19 from pneumonia without seeing any COVID-19 data during training. We also qualitatively show that our visual explanations are highly consistent with doctors' ground truth. While our work was not designed to target COVID-19, our experimental validation using the real world COVID-19 cases/data demonstrates the general applicability of our pipeline for different diseases based on medical imaging.

Keywords: New/unknown disease discovery · DNN confidence calibration · Visual explanation · COVID-19

1 Introduction

Extensive AI-based research and attempts have been made on automated medical imaging. Recent researches have witnessed remarkable progress in diagnosing known diseases when DNN classifier models are trained on a large number of images on known diseases [7]. However, in real world, unknown/new diseases continuously emerge, i.e., COVID-19. Unfortunately, since no training data for the new/unknown diseases are available at training time, existing DNN classifiers trained only on the *known disease (in-domain data)* oftentimes fail on the

B. W. Papież et al. (Eds.): MIUA 2020, CCIS 1248, pp. 346–358, 2020.
https://doi.org/10.1007/978-3-030-52791-4_27

new/unknown disease (out-of-domain data) in open-world practice. This problem is challenging even for a human. When doctors see a new disease, they could wrongly diagnose such a new disease as some other known diseases. In fact, at the beginning of the COVID-19 outbreak, doctors mistook the new COVID-19 disease as Pneumonia/SARS/MERS which are known diseases from the past.

The detection of out-of-domain unknown diseases is currently a challenging open research problem. Unknown diseases are theoretically unlimited. For each unknown disease, again there are theoretically infinite variations. To make it even harder, none of these data is available at model training and learning time. Recent work [8] has shown that DNN classifiers oftentimes suffer from overfitting and overconfidence issues, i.e., prediction accuracy is much lower than average confidence scores for predictions. As a consequence, DNN classifiers mistake unknown out-of-domain diseases as one of the known in-domain diseases.

On the other hand, deep learning models are black boxes. It is not clear why it works when it works, and why it does not work when it fails. Blindly accepting the decision from computer-aided diagnosis based on DNN classifiers can have serious consequences on patients in practice. Thus, it is highly desirable for models to provide explanations that can assist doctors to think and make the right decisions. To explain deep networks, several methods have been proposed based on internal states of the network [15–17]. Recently, Selvaraju [14] proposed Grad-CAM to compute neuron importance as part of a visual explanation. However, these approaches are only designed to explain the decision for existing diseases and cannot be applied to explain the decision when an unknown/new disease is detected.

In this paper, we aim to develop a high-quality explainable automatic medical imaging system that can accurately detect new/unknown diseases as well as provide reliable visual explanations to doctors. Our Contributions can be summarized as follows:

- We provide in-depth mathematical analysis to explain the overconfidence phenomena that leads to misdiagnosis of new/unknown diseases and present the calibrated confidence that can mitigate the overconfidence; We develop an automatic unknown disease discovery capability via confidence calibration for DNN classifiers trained only on known diseases data.
- We develop an automatic visual explanation into deep learning models to reveal suspected evidence in medical images for potential unknown diseases.
- We propose a novel explainable deep learning framework and pipeline that incorporates the above two automatic modules.
- Based on our proposed new pipeline, we conduct comprehensive experimental evaluations showing that our system achieves significant performance improvement on both quantitatively (unknown disease detection) and qualitatively (visual explanation) on Skin Lesion and Chest X-Ray datasets.

2 Our New Framework

In this section, we propose a novel framework and pipeline for explainable automated medical imaging. Figure 1 shows the whole framework including both

in-domain known disease diagnosis and out-of-domain unknown disease discovery. Next, we will present both training and testing processes with the focus of out-of-domain unknown disease discovery.

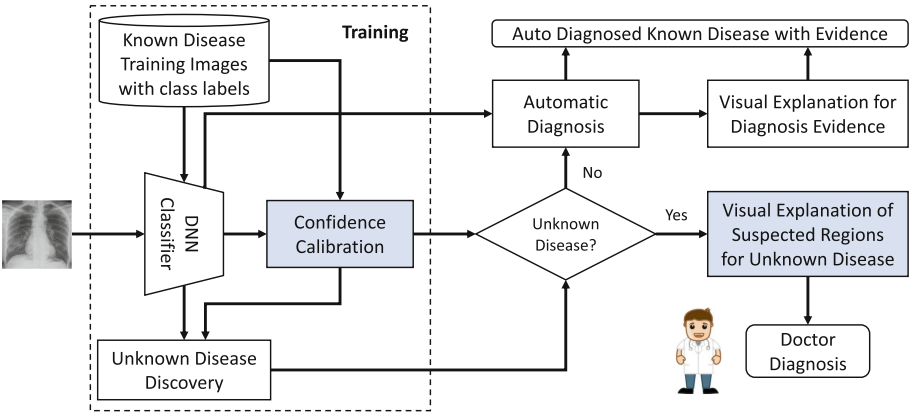

Fig. 1. Explainable automated diagnosis framework

Training Process: The components inside the dotted box in Fig. 1 indicate the training process. That said, DNN Classifier and Confidence Calibration for Unknown disease discovery will learn their parameters during training and later be used during testing. In the training, a DNN classifier is first trained *only on known disease training images with class labels*. Then, our confidence calibration component is to further adjust the confidence scores from DNN classifier output. This will largely mitigate the DNN overconfidence and avoid misdiagnosing a new disease as some known diseases.

To make our setting practical, our training process only takes the images of known diseases as inputs. We assume that new/unknown disease images are not available during model training time. In addition, our visual explanation component can automatically generate visual explanations only using the trained DNN classifier without needing to train a separate image segmentation model.

Testing Process (Diagnose Known and Unknown Diseases): The trained components in the dotted box are used in the testing process. Given an input image, it first goes through DNN classifier and confidence calibration components to generate the calibrated confidences. Next, we compare the calibrated confidence of the input image with a given threshold. If it is smaller than the threshold, we decide that this is a new/unknown disease; and we use our new visual explanation to show the potential suspected regions that have led to our detection of "new/unknown". Otherwise, we directly use the trained DNN classifier model to automatically diagnose to be one of the known diseases and provide its visual explanation [17] for doctors to review and confirm.

In the rest of our paper, we will focus on introducing our novel designs for the two blue components in Fig. 1.

3 Confidence Calibration for New Disease Discovery

Overconfidence phenomenon has been observed empirically in literature [8]. In this section, we propose a mathematical explanation of why overconfidence happens in deep neural networks (DNN) classifiers that lead to misdiagnosis of new/unknown diseases. Motivated by our mathematical logic, we shall present calibration solutions.

3.1 Mathematical Explanation of Overconfidence Observation

DNN classifiers implicitly assume all data are in-domain. Thus, they model:

$$\widehat{p}(\mathrm{y}_{in}|\mathrm{d}=1,\boldsymbol{x}) \text{ for a random variable d} = \begin{cases} 1 & \text{if } \boldsymbol{x} \text{ is an in-domain sample} \\ 0 & \text{if } \boldsymbol{x} \text{ is an out-of-domain sample} \end{cases}$$

In open world settings, one needs to learn:

$$\widehat{p}(\mathrm{y}_{in} \cup \mathrm{y}_{out}|\boldsymbol{x}) = \sum_{\mathrm{d}\in\{0,1\}} \widehat{p}(\mathrm{y}_{in} \cup \mathrm{y}_{out}|\mathrm{d},\boldsymbol{x})\widehat{p}(\mathrm{d}|\boldsymbol{x}) \tag{1}$$

Since unknown data is not available during training, we can only model the following based on training data:

$$\widehat{p}(\mathrm{y}_{in}|\boldsymbol{x}) = \sum_{\mathrm{d}\in\{0,1\}} \widehat{p}(\mathrm{y}_{in}|\mathrm{d},\boldsymbol{x})\widehat{p}(\mathrm{d}|\boldsymbol{x}) \tag{2}$$

Then, we hope to indirectly model out-of-domain probability:

$$\widehat{p}(\mathrm{y}_{out}|\boldsymbol{x}) = g(\widehat{p}(\mathrm{y}_{in}|\boldsymbol{x})) \tag{3}$$

Thus, the combination of $\widehat{p}(\mathrm{y}_{in}|\boldsymbol{x})$ and $\widehat{p}(\mathrm{y}_{out}|\boldsymbol{x})$ forms a probability distribution.

Proposition 1. *For an unknown image* $\boldsymbol{x}$*, we have* $\widehat{p}(\mathrm{y}_{in}|\mathrm{d}=1,\boldsymbol{x}) \approx \frac{\widehat{p}(\mathrm{y}_{in}|\boldsymbol{x})}{\widehat{p}(\mathrm{d}=1|\boldsymbol{x})}$.

Proof: In this case, $\widehat{p}(\mathrm{y}_{in}|\mathrm{d}=0,\boldsymbol{x})\widehat{p}(\mathrm{d}=0|\boldsymbol{x})$ is small given the small value of $\widehat{p}(\mathrm{y}_{in}|\mathrm{d}=0,\boldsymbol{x})$ in open world (since y_{in} is a label for in-domain samples). Thus, Eq. 2 can be rewritten as follows:

$$\begin{aligned} \widehat{p}(\mathrm{y}_{in}|\boldsymbol{x}) &= \sum_{\mathrm{d}\in\{0,1\}} \widehat{p}(\mathrm{y}_{in}|\mathrm{d},\boldsymbol{x})\widehat{p}(\mathrm{d}|\boldsymbol{x}) \\ &= \widehat{p}(\mathrm{y}_{in}|\mathrm{d}=1,\boldsymbol{x})\widehat{p}(\mathrm{d}=1|\boldsymbol{x}) + \cancelto{0}{\widehat{p}(\mathrm{y}_{in}|\mathrm{d}=0,\boldsymbol{x})}\ \widehat{p}(\mathrm{d}=0|\boldsymbol{x}) \\ &\approx \widehat{p}(\mathrm{y}_{in}|\mathrm{d}=1,\boldsymbol{x})\widehat{p}(\mathrm{d}=1|\boldsymbol{x}) \end{aligned}$$

Thus, we reorganize the formula:

$$\widehat{p}(\mathrm{y}_{in}|\mathrm{d}=1,\boldsymbol{x}) \approx \frac{\widehat{p}(\mathrm{y}_{in}|\boldsymbol{x})}{\widehat{p}(\mathrm{d}=1|\boldsymbol{x})} \tag{4}$$

Hypothesis 1. *Let* f^c *be the unnormalized probability* $\widehat{p}(\mathrm{y}_{in}|\boldsymbol{x})$ *and* f^d *be the unnormalized probability* $\widehat{p}(\mathrm{d}=1|\boldsymbol{x})$*, i.e.,* $\widehat{p}(\mathrm{y}_{in}|\boldsymbol{x}) = norm(f^c(\boldsymbol{x}))$*,* $\widehat{p}(\mathrm{d}=1|\boldsymbol{x}) = norm(f^d(\boldsymbol{x}))$*. We call the unnormalized probability "logits". We hypothesize the following:*

$$norm\Big(\frac{f^c(\boldsymbol{x})}{f^d(\boldsymbol{x})}\Big) \approx \frac{norm(f^c(\boldsymbol{x}))}{norm(f^d(\boldsymbol{x}))}$$

Mathematical Explanation of Overconfidence Observation [8]**:** Assuming that Hypothesis 1 holds, we show the explanation via Eq. 4. Given Hypothesis 1, we can rewrite Eq. 4 as follows:

$$\widehat{p}(\mathrm{y}_{in}|\mathrm{d}=1,\boldsymbol{x}) \approx \frac{\widehat{p}(\mathrm{y}_{in}|\boldsymbol{x})}{\widehat{p}(\mathrm{d}=1|\boldsymbol{x})} \approx \frac{norm(f^c(\boldsymbol{x}))}{norm(f^d(\boldsymbol{x}))} \approx norm\Big(\frac{f^c(\boldsymbol{x})}{f^d(\boldsymbol{x})}\Big)$$

Then, we use the following "softmax function" [7] to normalize the logits to be a probability distribution:

$$\widehat{p}(\mathrm{y}_{in}|\mathrm{d}=1,\boldsymbol{x}) \approx softmax\Big(\frac{f^c(\boldsymbol{x})}{f^d(\boldsymbol{x})}\Big) = \frac{exp(\frac{f_i^c(\boldsymbol{x})}{f_i^d(\boldsymbol{x})})}{\sum_{j=1}^{C} exp(\frac{f_j^c(\boldsymbol{x})}{f_j^d(\boldsymbol{x})})}$$

We illustrate our overconfidence explanation in Fig. 2 using an example: Assuming there are two in-domain classes in our classifier. For an out-of-domain $\boldsymbol{x}$, it is expected that $f_c(\boldsymbol{x})$ (the unnormalized $\widehat{p}(\mathrm{y}_{in}|\boldsymbol{x})$) (blue points in Fig. 2) for both classes are small, e.g., 0.5 and 0.8. The normalization function maps $f_c(\boldsymbol{x})$ to probabilities 43% and 57%. However, for an out-of-domain $\boldsymbol{x}$, $f_d(\boldsymbol{x})$ ((the unnormalized $\widehat{p}(\mathrm{d}=1|\boldsymbol{x})$) is a very small number, e.g., 0.1. After $f_c(\boldsymbol{x})$ is divided by the small $f_d(\boldsymbol{x})$, the final model logits (red points in Fig. 2) for both classes become 5 and 8. The softmax normalization maps them to probabilities 5% and 95%. With that, the model will conclude that $\boldsymbol{x}$ is classified into class #2 with a confidence level of 95%. This shows how a wrong decision can be made with overconfidence for out-of-domain images.

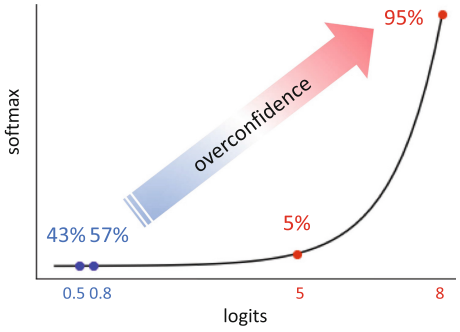

Fig. 2. Overconfidence explanation for an unknown image (Color figure online)

3.2 Confidence Calibration Without Overconfidence

Based on the above mathematical explanation of overconfidence, an intuitive solution to mitigate overconfidence is temperature scaling [11], i.e., scale $f^d(\boldsymbol{x})$

with a large temperature T to compute the calibrated confidence score.

$$S_c(\boldsymbol{x}) = softmax\left(f_c(\boldsymbol{x})\right) \approx softmax\left(\frac{f_c(\boldsymbol{x})}{f_d(\boldsymbol{x})T}\right) = \frac{exp(\frac{f_i^c(\boldsymbol{x})}{f_i^d(\boldsymbol{x})T})}{\sum_{j=1}^{C} exp(\frac{f_j^c(\boldsymbol{x})}{f_j^d(\boldsymbol{x})T})}$$

where S_c is calibrated confidence score for each class c. Unfortunately, since this temperature T is not trainable, it is hard to determine the right temperature for any case. In our experiments, T is simply set as a large number.

Thus, we present another confidence calibration approach using *Mahalanobis distance* based on a *generative classifier layer* to replace with the softmax layer [10]. According to a simple theoretical connection, the pretrained softmax classifier is likely to follow class-conditional Gaussian distribution. That said, we can parameterize the class-conditional Gaussian distribution with class mean μ_i and covariance matrix Σ as follows:

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} f(\boldsymbol{x}_i), \quad \hat{\Sigma} = \frac{1}{N} \sum_{c} \sum_{i:y_i=c} (f(\boldsymbol{x}_i) - \hat{\mu}_c)(f(\boldsymbol{x}_i) - \hat{\mu}_c)^T$$

During testing, given an input image, we can compute its confidence score based on Mahalanobis distance (distance between a point and a probability distribution) as follows:

$$S_c(\boldsymbol{x}) = -(f(\boldsymbol{x}) - \mu_c)^T \Sigma^{-1} (f(\boldsymbol{x}) - \mu_c) \tag{5}$$

where S_c is the same for each class c. $f(\boldsymbol{x})$ represents the output features at the penultimate layer of DNN classifier models. Since all S_c does not have to form a probability distribution, we will introduce how these scores are matched to the final decision engine in the next section.

3.3 Decision Engine

In this section, our goal is to use the confidence scores to derive the final probability distribution, i.e., both $\widehat{p}(\mathrm{y}_{in}|\boldsymbol{x})$ and $\widehat{p}(\mathrm{y}_{out}|\boldsymbol{x})$.

Consider the calibrated confidences S_c for any class c. Accordingly, since y_{out} has only one unknown class, i.e., y_{out} equivalent to $\mathrm{d} = 0$. We consider the following threshold-based function to derive the probability of out-of-domain probability:

$$\widehat{p}(\mathrm{y}_{out}|\boldsymbol{x}) = \widehat{p}(\mathrm{d} = 0|\boldsymbol{x}) = \begin{cases} 1 & \text{if } s \leq \delta \\ 0 & \text{otherwise} \end{cases}$$

where $s = \max_c S_c$ meaning the largest confidence scores among all in-domain classes. δ is the threshold based on the true positive rate requirement. Note that the in-domain images ($s > \delta$) will be further diagnosed as one of the unknown diseases using the conventional softmax layer in DNN classifiers. When an image $\boldsymbol{x}$ is detected as in-domain, we directly compute its classification probability as:

$$\widehat{p}(\mathrm{y}_{in}|\boldsymbol{x}) = \widehat{p}(\mathrm{y}_{in}|\mathrm{d} = 1, \boldsymbol{x})$$

4 Visual Explanation of Suspected Regions (DisCAM)

Existing work tried to provide visual explanations for the in-domain classification decision [17]. It produces heat maps to visualize the most indicative regions in the image regarding the diagnosed disease using class activation mappings (CAM). To generate CAM M_c for an input image with diagnosed in-domain class c, the DNN model extracts the all k feature maps f_k that are output by the final convolutional layer.

$$M_c = \sum_k w_{c,k} f_k$$

where $w_{c,k}$ the weight in the final classification layer for feature map k for in-domain class c.

However, CAM cannot be directly used to discover unknown regions since none of the in-domain disease classes is diagnosed. Thus, we devise the *Discovery CAM (DisCAM)* based on the original CAM. We use the calibrated confidence to combine the weights in the final classification layer as follows:

$$M = \sum_k \sum_i \frac{S_i}{\sum_{j=1}^{C} S_j} w_{i,k} f_k$$

At last, we follow CAM to generate a heat map based on the neuron importance weights M by upscaling M to the dimensions of the image and overlaying the image for each pixel.

5 Experimental Evaluation

5.1 Datasets

We have conducted experimental evaluations based on our proposed new deep learning pipeline. We use two medical datasets in our experiment. For each dataset, we discuss its in-domain and out-of-domain data respectively.

Skin Lesion Dataset
For in-domain images, we use the latest ISIC2019 Skin Lesion Challenge Dataset [1]. It contains 25,331 training images and each image is labeled as one of 8 categories/classes, including 7 different diseases and 1 benign. The task is to classify an image into one of these eight classes. Since the class ground truth of testing images are not available, we evaluate our approaches via 10-fold cross-validation on training data and report the average results. For out-of-domain images, we download the images in the "unknown" category from Gallery in ISIC archive website [2]. Dermatologists have determined these images do not belong to any of the above 8 categories. In addition, each image is provided with segmentation ground truth by dermatologists.

Chest X-Ray Dataset
For in-domain images, we use the Chest X-Ray dataset [3] from Kaggle. It contains 5,863 training images and 624 validation images. Each image is labeled

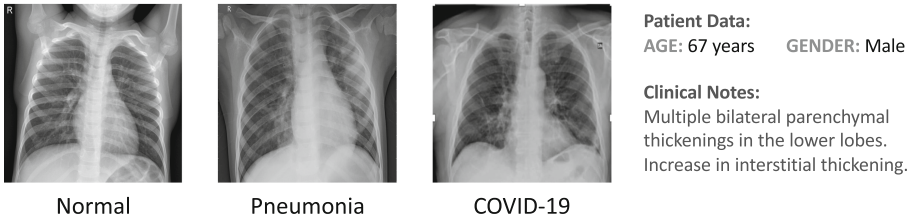

Fig. 3. COVID-19 data example

as either Pneumonia or Normal. For out-of-domain images, we collect the chest X-ray images of on-going spreading new disease COVID-19 from three online resources [4,5], and [6]. Since the in-domain data only have frontal chest X-ray, we only keep frontal X-ray out-of-domain COVID-19 images for our testing purpose. Each COVID-19 data sample consists of a chest X-ray image, a patient's basic information, and clinical notes from doctors. Figure 3 shows a sample chest X-ray images including in-domain normal and pneumonia as well as a new out-of-domain COVID-19 image.

COVID-19 started in late 2019 and is caused by a new virus, a.k.a. severe acute respiratory syndrome coronavirus 2, or Sars-CoV-2. The infection may result in severe pneumonia with clusters of illness onsets. Its impacts on public health make it paramount to clarify the clinical features with other pneumonia. Thus, the computer-aided discovery of COVID-19 is challenging but in the meanwhile practically very useful.

5.2 Implementation and Model Training

We implement our code using the PyTorch 1.1.0 framework. The experiment is run on 8 NVIDIA GPUs (Tesla V100 16 GB GPU).

Our first step is to train state-of-the-art based CNN models for both datasets. We first normalized both datasets using the mean and standard deviation calculated on the statistics of all training images. The skin lesion dataset has mean (0.679, 0.526, 0.519) and standard deviation (0.181, 0.185, 0.198), and chest X-ray dataset has mean (0.480, 0.480, 0.480) and standard deviation (0.232, 0.232, 0.232). Note that the gray-scale chest X-ray images have the same values for all RGB channels. For each image, we first resize it to be 256×256. We performed dynamic in-memory augmentation by randomly cropping to 224×224, horizontal & vertical flips, and zooming by appropriate transformations in the PyTorch data loader. Following the previous work [12], we conduct transfer learning with ResNet-50, ResNet-101 and ResNet-152 pre-trained on the ImageNet [13]. We also use batch size 64 and use the same approach in [12] to choose the optimal learning rate. Using this learning rate, we continue following the two-step model training in [12].

To validate our model training, we first evaluate the performance of in-domain classification on all trained models. We use Top-1 accuracy and AUC

Table 1. In-domain classification evaluation results

Dataset	Skin Lesion		Chest X-Ray	
Metrics	ACC	AUC	ACC	AUC
ResNet-50	85.44	86.78	92.95	91.03
ResNet-101	85.81	86.79	93.30	92.34
ResNet-152	86.41	86.88	94.12	93.53

metrics for image classification. AUC stands for Area under the Receiver Operating Characteristic (ROC) Curve, which the ROC curve is a graph plotting TPR against the FPR = FP/(FP + TN) by varying a threshold. Table 1 shows the in-domain performance. Since our out-of-domain detection will not retrain the model, the in-domain classification performance will not be impacted in our new deep learning pipeline.

5.3 New/Unknown Disease (Out-of-Domain) Detection Evaluation

We follow the evaluation metrics in the literature [9]. Let TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. We use the following out-of-domain evaluation metrics:

- **TNRkTPR (high)** (True Negative Rate (TNR) at k% True Positive Rate (TPR)): can be interpreted as the probability that an out-of-domain image is classified correctly when the true positive rate (TPR) for in-domain data is as high as k%, where TPR = TP/(TP + FN). In our experiment, we choose k to be 85.
- **Detection Error (low)**: measures the misclassification probability when TPR is k%. Detection error is defined as follows:

$$\min_{\delta}\{P_{IND}(s \leq \delta)p(\boldsymbol{x} \in P_{IND}) + P_{OOD}(s > \delta)p(\boldsymbol{x} \in P_{OOD})\}$$

  where s is a confidence score. We follow the same assumption that both IND and OOD examples have an equal probability of appearing in the test set.
- **AUROC (high)** (Area under the Receiver Operating Characteristic Curve): The ROC curve is a graph plotting TPR against the FPR = FP/(FP + TN) by varying a threshold.
- **AUPR (high)** (Area under the Precision-Recall Curve): The PR curve is a graph plotting the precision (TP/(TP + FP)) against recall (TP/(TP + FN)) by varying a threshold.

Table 2 and Table 3 show the unknown disease detection results on Skin Lesion and Chest X-Ray diseases respectively. As one can see, the baseline suffers from failure due to overconfidence. Temperature Scaling (TS) improves the performance but still not satisfactory due to the untrainable temperature T. Generative Classifier (GC), after removing the source of overconfidence by replacing

Table 2. OOD detection results for Skin Lesion dataset

Model	Method	TNR85TPR	Det. error	AUROC	AUPR
ResNet50	Baseline	25.16	41.91	64.34	96.84
	TS	45.24	29.29	75.90	97.59
	GC	61.35	23.40	83.72	98.47
ResNet101	Baseline	26.55	39.44	65.60	96.89
	TS	48.99	28.48	76.37	97.35
	GC	64.59	23.31	84.32	98.53
ResNet152	Baseline	28.18	37.74	66.34	97.01
	TS	51.39	28.04	77.34	97.34
	GC	65.25	22.88	84.45	98.99

Table 3. OOD detection results for Chest X-Ray dataset

Model	Method	TNR85TPR	Det. error	AUROC	AUPR
ResNet50	Baseline	9.65	49.54	38.84	84.19
	TS	37.24	47.83	50.77	87.64
	GC	90.49	12.96	94.02	98.21
ResNet101	Baseline	8.04	47.48	47.09	88.20
	TS	42.98	34.02	65.80	93.37
	GC	90.98	12.15	94.38	98.61
ResNet152	Baseline	14.94	46.71	52.35	88.40
	TS	48.04	29.73	71.55	94.84
	GC	91.92	11.22	94.63	99.11

the softmax layer, achieves significant performance improvement for all metrics. The GC performance on the Skin Lesion dataset is slightly lower since it contains colorful images and there are many varieties of noises on the images such as color, illumination, skin hair, etc. GC improved baseline performance by over 6 times on Chest X-Ray to detect new out-of-domain COVID-19 disease using the model trained on known pneumonia and normal images. In fact, we achieved almost perfect detection of COVID-19 when tested on a small dataset.

5.4 Case Study: Visual Explanation for Suspected Regions

Next, we conduct some case study of visual explanation on new/unknown diseases to (1) qualitatively validate our visual explanation method by comparing with ground truth doctor explanation, and (2) visually elaborate the underlying reason why our unknown disease detection method works well. Each Skin Lesion unknown image has a doctor provided segmentation ground truth (green lines). Each Chest X-Ray COVID-19 image has clinic notes which explain the suspected

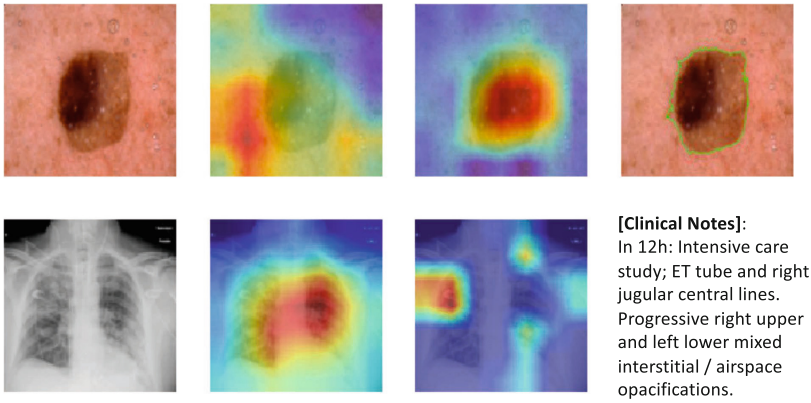

Fig. 4. Case 1: CAM looks at wrong regions

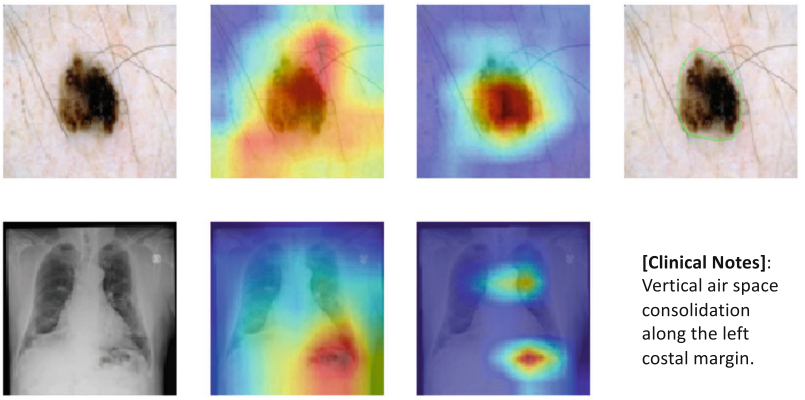

Fig. 5. Case 2: CAM looks at too broad regions

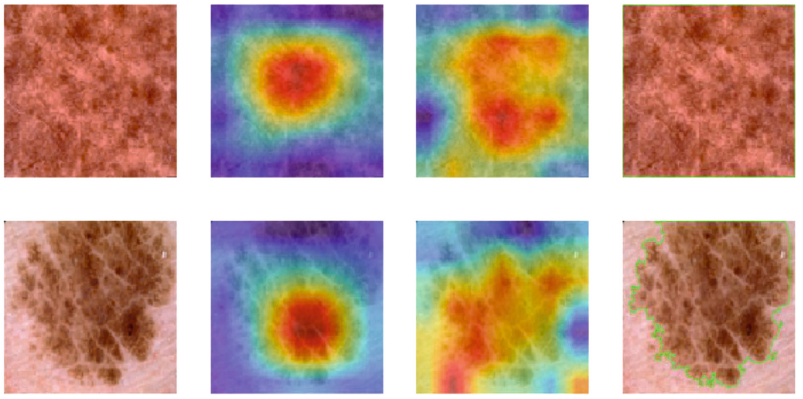

Fig. 6. Case 3: CAM looks at too narrow regions

regions in X-Ray that indicate COVID-19 diagnosis. It is important to note that the left and right sides are flipped over in conventional X-Ray images.

Figure 4, Fig. 5 and Fig. 6 show three different types of wrong regions baseline CAM method looks at, which leads to all wrong decisions. In Fig. 4, CAM looks at completely wrong regions which no doubt leads to wrong predictions. Figure 5 and Fig. 6 are more interesting. In Fig. 5, although the regions CAM looks at include the correct region, it also looks at other distracting regions. For example, the hair on the skin and white abdomen area in Chest X-Ray possibly confused the decision engine. On the other hand, Fig. 6 shows that cases where CAM looks at too narrow regions and missed the holistic view of the disease. Meanwhile, our DisCAM looks at correct regions in all these three cases and also correctly detects all these unknown disease images.

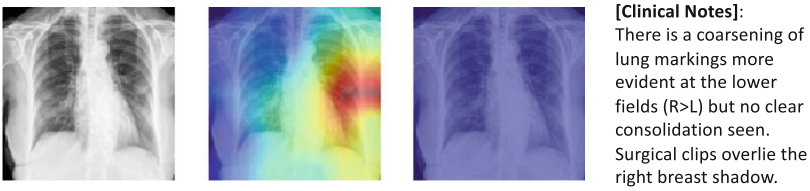

Fig. 7. Case 4: DisCAM does not find any evidence

Figure 7 shows another interesting visual explanation in which our DisCAM method shows no particular suspected region in the image. That said, our trained model does not discover any suspected regions based on the learned knowledge of known diseases and therefore also concludes this is a new/unknown disease. In contrast, CAM identifies completely wrong regions and mistakes the unknown disease as a known disease.

6 Conclusion and Future Work

We proposed a framework for explainable automatic medical imaging that can discover unknown diseases and provide a visual explanation for that decision. We first mathematically analyzed and explained why existing models oftentimes fail to classify new/unknown data correctly. We then showed calibration methods that can mitigate the overconfidence. We validated the new calibration method with multiple datasets and demonstrated its effectiveness for unknown data detection via quantitative evaluations. We successfully detected COVID-19 with our new deep learning pipeline trained with only known Pneumonia data. We provided visual explanations of our new/unknown detection decisions based on the calibrated confidence methods. Our explanations are consistent with doctors' ground truth and clinical notes. For future work, we will continue to validate our work by evaluating more and larger datasets. As a natural next step, we also plan to continue working on few-shot learning using a small amount of new disease data to efficiently learn the new diseases for future predictions/classifications.

Acknowledgement. I would like to express my sincere gratitude to Professor Jiang Du and his staff from the Department of Radiology at the University of California at San Diego who introduced me to the wonderful medical imaging space, especially on MRIs and image segmentation. I would also like to express my deep appreciation to Dr. Jeremy Shen from Samsung for educating me on related work and the pitfalls. I am very grateful to AI4ALL for awarding me a travel grant to attend the 2019 NeurIPS workshop in Vancouver, Canada for "Machine Learning 4 Health". Without their support, this work would have been impossible.

References

1. https://challenge2019.isic-archive.com
2. https://www.isic-archive.com/!/topWithHeader/onlyHeaderTop/gallery
3. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
4. https://radiopaedia.org/
5. https://www.kaggle.com/bachrr/covid-chest-xray
6. https://www.kaggle.com/andrewmvd/convid19-x-rays
7. Goodfellow, I.J., Bengio, Y., Courville, A.C.: Deep Learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge (2016)
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML, pp. 1321–1330 (2017)
9. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
10. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS, pp. 7167–7177 (2018)
11. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)
12. Mishra, S., Imaizumi, H., Yamasaki, T.: Interpreting fine-grained dermatological classification by deep learning. In: CVPR Workshops, pp. 2729–2737 (2019)
13. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
14. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. **128**(2), 336–359 (2020)
15. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV, pp. 818–833 (2014)
16. Zhou, B., Khosla, A., Lapedriza, À, Oliva, A., Torralba, A.: Object detectors emerge in deep scene CNNs. CoRR, abs/1412.6856 (2014)
17. Zhou, B., Khosla, A., Lapedriza, À, Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In CVPR, pp. 2921–2929 (2016)