Viral Pneumonia Screening on Chest X-Rays Using Confidence-Aware Anomaly Detection

Jianpeng Zhang, Yutong Xie, *Graduate Student Member, IEEE*, Guansong Pang, Zhibin Liao,
Johan Verjans, Wenxing Li, Zongji Sun, Jian He, Yi Li, Chunhua Shen, *Member, IEEE*,
and Yong Xia, *Member, IEEE*

Abstract— Clusters of viral pneumonia occurrences over a short period may be a harbinger of an outbreak or pandemic. Rapid and accurate detection of viral pneumonia using chest X-rays can be of significant value for large-scale screening and epidemic prevention, particularly when other more sophisticated imaging modalities are not readily accessible. However, the emergence of novel mutated viruses causes a substantial dataset shift, which can greatly limit the performance of classification-based approaches. In this paper, we formulate the task of differentiating viral pneumonia from non-viral pneumonia and healthy controls into a one-class classification-based anomaly detection problem. We therefore propose the confidence-aware anomaly detection (CAAD) model, which consists of a shared feature extractor, an anomaly detection module, and a confidence prediction module. If the anomaly score produced by the anomaly detection module is large enough, or the confidence score estimated by the confidence prediction module is small enough, the input will be accepted as an anomaly case (*i.e.*, viral pneumonia). The major advantage of our approach over binary classification is that we avoid modeling individual viral pneumonia classes explicitly and treat all known viral pneumonia cases as anomalies to improve the one-class model. The proposed model outperforms binary classification models on the clinical X-VIRAL dataset that contains 5,977 viral pneumonia (no COVID-19) cases, 37,393 non-viral pneumonia or healthy cases. Moreover, when directly testing on the X-COVID dataset that contains 106 COVID-19 cases and 107 normal controls without any fine-tuning, our model achieves an AUC of 83.61% and sensitivity of 71.70%, which is comparable to the performance of radiologists reported in the literature.

Index Terms— Viral pneumonia screening, deep anomaly detection, confidence prediction, chest X-ray.

I. INTRODUCTION

VIRAL pneumonia is a type of lung infection caused by viruses. It causes approximately 30 per cent of pneumonia cases, and is typically mild and healing without treatment. However, COVID-19 has demonstrated to be a much more dangerous virus causing severe pneumonia in a subset of patients, and has rapidly spread throughout the globe within a few months, threatening the health of billions of human beings with significant mortality [1], [2]. The clustering aspect of viral pneumonia occurrences, like SARS [3], MERS [4], and COVID-19, can often suggest a potential outbreak. Therefore, it is desirable to develop an accurate, fast, and cost-effective tool for viral pneumonia detection, which provides the prerequisite for rigorous detection, contact tracing, and isolation of infected subjects in a large district.

Take the recent COVID-19 outbreak for example. Viral nucleic acid detection using real-time polymerase chain reaction (RT-PCR) is the accepted standard diagnostic method [5], [6]. However, the test has suboptimal sensitivity and specificity and many hyperendemic regions and countries are not able to provide sufficient RT-PCR testing for tens of thousands of suspected subjects in a short period of time. Moreover, it would fail to detect the newly evolved coronavirus before extracting the DNA sequence of the new virus from the swab, which may delay the control of the epidemic [7].

Accordingly, medical imaging is considered as a vital technique to assist doctors to evaluate the disease and to optimise prevention and control measures. Clinically, chest X-ray is the most commonly used imaging modality in the diagnostic workup of patients with thoracic abnormalities, due to its fast imaging speed, low radiation and low cost [8].

Jianpeng Zhang and Yutong Xie are with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: james.zhang@mail.nwpu.edu.cn; xuyongxie@mail.nwpu.edu.cn).

Guansong Pang, Zhibin Liao, Johan Verjans, and Chunhua Shen are with the School of Computer Science, The University of Adelaide, SA 5005, Australia (e-mail: chunhua.shen@adelaide.edu.au).

Wenxing Li and Zongji Sun are with JF Healthcare Inc., Nanjing 210000, China.

Jian He is with the Department of Radiology, Nanjing Drum Tower Hospital-Affiliated Hospital, Medical School, Nanjing University, Nanjing 210029, China.

Yi Li is with GreyBird Ventures, LLC, Concord, MA 01742 USA.

Yong Xia is with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Research and Development Institute, Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China (e-mail: yxia@nwpu.edu.cn).

Digital Object Identifier 10.1109/TMI.2020.3040950

In comparison to computed tomography (CT), chest X-ray cannot provide 3D anatomy but is generally considered to be able to differentiate pneumonia although it is probably considered the most challenging plain film to interpret correctly [9]. Accurate interpretation is vital to patient management in the acute setting, and to help identify clustering occurrence of COVID-19. Therefore, we aim to develop an automated and accurate viral pneumonia screening technique using chest X-ray as a stopgap for early warning of clusters of an outbreak caused by COVID-19 or a novel coronavirus.

Recent advances in deep learning have led to breakthroughs in many long-standing medical image analysis tasks, such as the detection, staging, and delineation of pathological abnormalities. On the task of chest X-ray interpretation, deep convolutional neural networks (DCNNs) have been constructed to diagnose the most common and important thoracic diseases [10]–[13] and to differentiate between viral and bacterial pneumonia [14], [15]. In contrast, we attempt to distinguish viral pneumonia from all non-viral pneumonia ones (not merely bacterial pneumonia), aiming to rapidly detect the clusters of viral pneumonia (*e.g.*, COVID-19) caused by a novel virus before an outbreak. This task, nevertheless, remains challenging due to two intrinsic complexities. First, many different types of viruses can cause pneumonia, including the influenza A/B viruses, respiratory syncytial virus, coronaviruses, herpes simplex, measles, chickenpox, and more seriously, some novel viruses. The complex pathological cues of viral pneumonia cause considerable visual differences on X-ray images (see Fig. 1), leading to substantial intra-class variance and dataset shift (*e.g.*, newly emerged viral pneumonia cases have highly different lesions from the known viral pneumonia in the training data). Second, it is hard to collect a large number of positive (*i.e.*, viral pneumonia) samples in the early stage of an outbreak. Hence, the viral versus non-viral pneumonia classification is an extreme class-imbalance problem. These two complexities pose significant challenges to the commonly-used binary classification approaches since 1) they normally assume that the training and test data share an identical distribution (*i.e.*, no dataset shift) and there exists small class variance within each class; and 2) they often ignore the class-imbalance problem. Consequently, the classifiers yield poor sensitivity performance. The sensitivity, however, is clinically significant, since it can be much more disastrous to discharge a patient with viral pneumonia than to misdiagnose a healthy control. To address both challenges, we advocate the replacement of a classifier by an anomaly detector for this chest X-ray interpretation problem. As an one-class classification approach [16], anomaly detection is not only able to detect dissimilar or even previously unseen anomalies, but also less dependent on labeled anomaly data than binary classification [17]–[19].

In this paper, we propose a confidence-aware anomaly detection (CAAD) model to distinguish viral pneumonia cases from non-viral pneumonia cases and healthy controls using chest X-rays. We reformulate the viral pneumonia screening into a one-class classification-based anomaly detection task, instead of a binary classification one. Specifically, we design an anomaly detection module to assign each X-ray image an anomaly score and employ the contrastive loss function
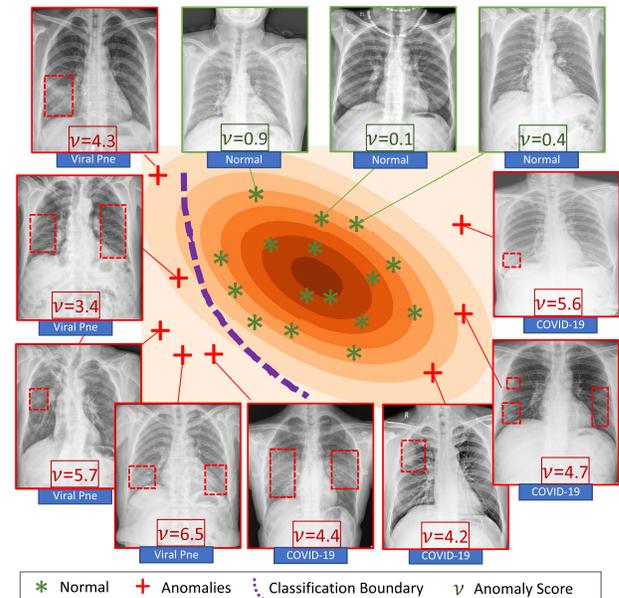

Fig. 1. An illustration of binary classification vs. anomaly detection in distinguishing viral pneumonia cases (*i.e.*, 'Anomalies') from non-viral cases and healthy controls (*i.e.*, 'Normal'). Image samples are shown on a two-dimensional contour plot, where inner contour lines indicate large density. The red dotted boxes in viral pneumonia cases are the suspected lesion areas annotated by radiologists. Novel viral pneumonia (*i.e.*, COVID-19) can be either similar to or different from the cases caused by known viruses (denoted as 'Viral Pne'). The decision boundary given by a binary classification approach can well separate 'Normal' cases and 'Viral Pne' cases, but may not able to distinguish 'Normal' cases from COVID-19 cases. In contrast, our anomaly detection approach can distinguish both known and novel viral pneumonia from 'Normal' cases by assigning the former large anomaly scores and the latter small anomaly scores.

to ensure that the scores generated for anomalies (*i.e.*, viral pneumonia) are significantly larger than those for non-viral pneumonia cases and normal controls. We further introduce an additional confidence prediction module to describe the confidence of the anomaly detection module. According to the confidence level, we re-assign the samples with low confidence as suspected viral pneumonia for further medical tests, which helps reduce false-negative cases and thus improves the sensitivity. Both the anomaly detection module and confidence prediction module can be jointly optimized in an end-to-end manner. We have evaluated our CAAD model on the X-VIRAL dataset which contains 5,977 positive viral pneumonia subjects (anomalies) and 37,393 negative subjects (non-viral pneumonia and health cases, known as normal controls). Our proposal achieves the state-of-the-art performance, *i.e.*, 87.57% AUC, for viral pneumonia screening. Even with no exposure to COVID-19 cases during training, our CAAD model shows superior performance for the COVID-19 screening purpose, achieving an AUC of 83.61% and a sensitivity of 71.70% on our additional unseen X-COVID dataset with 106 confirmed and 107 normal subjects. This is comparable to the performance of radiologists reported in the literature [20].

The main contributions of this paper are summarized as follows.

- We formulate viral pneumonia screening into an anomaly detection problem and propose the CAAD model to

solve it, which is able to detect viral pneumonia caused by novel viruses and is less dependent on labeled viral pneumonia data than classification models.

- We propose to predict failures of anomaly detection by modeling its confidence level so as to further improve the screening sensitivity.
- Our experimental results demonstrate the effectiveness and strong generalizability of our model in viral pneumonia screening and the potential on epidemic prevention and control.

II. RELATED WORK

A. Chest X-Ray for Pulmonary Disease Screening

Chest X-ray is one of the most commonly used imaging modalities for visualizing and quantifying the structural and functional consequences of thoracic diseases, providing high-resolution pictures of disease progression and therapy response. Magree *et al.* [21] documented the incidence of pneumonia confirmed with X-ray imaging and demonstrated a high incidence, which guided the later prevention and treatment of vaccine. Jacobi *et al.* [22] described the most common manifestations and patterns of lung abnormality on chest X-ray in COVID-19 and suggested that the medical community can frequently rely on portable chest X-ray due to its widespread availability. Wong *et al.* [20] demonstrated that the common CT findings of bilateral involvement, peripheral distribution, and lower zone dominance can also be appreciated on chest X-ray, which shows the potential of using chest X-ray as a tool for COVID-19 identification. Borghesi *et al.* [23] presented an experimental chest X-ray scoring system and applied it to hospitalized patients with COVID-19 to quantify and monitor the severity and progression of COVID-19. Different from these studies, we focus on the viral pneumonia screening and aim to develop a fast and accurate algorithm to differentiate viral pneumonia from non-viral pneumonia and normal controls for the prevention and control of a possible outbreak.

B. Deep Learning for Chest X-Ray Interpretation

To improve efficiency and ease the burden of radiologists, researchers gradually adapt the recent advances of deep learning to interpret chest X-ray images. For computer-aided diagnosis of 14 common thoracic diseases, Wang *et al.* [11] proposed a weakly-supervised classification and localization framework, Rajpurkar *et al.* [12] constructed a 121-layer dense convolutional neural network that can perform the task at a level exceeding practicing radiologists, and Wang *et al.* [13] introduced an attention mechanism to help the model focus on the lesion area and thus further improved the diagnosis performance. Besides, many attempts [14], [15] have been made to investigate DCNN-based classification models for pneumonia detection and the differentiation between viral and bacterial pneumonia, aiming to facilitate rapid referrals for children who need urgent intervention. In these studies, these diagnostic tasks are formulated as classification problems, which are usually solved based on the intra-class similarity and inter-class dissimilarity of pathological patterns. Such classification models may fail to distinguish viral and non-viral

pneumonia since the category of viral pneumonia contains cases with highly variable visual appearances.

C. Failure Prediction

Despite their success, deep learning models still make mistakes, particularly when applied to real-world applications. To avoid the decision risk caused by the inherent defects of deep learning models, failure prediction is of great necessity. Hendrycks *et al.* [24] proposed to detect failures and out-of-distribution examples in neural networks via the prediction/maximum class probability method. However, it is hard to distinguish the failures if they are misclassified with a high probability. To address this issue, Corbiere *et al.* [25] considered the true class probability as a suitable confidence criterion for failure estimation. Xie *et al.* [26], [27] proposed a deep segmentation-emendation model for gland instance segmentation, in which an emendation network is designed to predict the inconsistency between the ground truth masks and pixel-wise predictions of segmentation network, and the failure predictions made by an emendation network can be utilized to refine the segmentation result. Inspired by these works, we attempt to predict failures of anomaly detection, which improves not only the detection performance but also the diagnosis credibility of our model.

D. Deep Anomaly Detection

Anomaly detection is the task of identifying unusual samples from the majority of the data [28]. Traditional anomaly detection mainly focuses on the kernel-based one-class classification. Typical methods include the One-Class SVM (OC-SVM) [29] and Support Vector Data Description (SVDD) [29], which attempt to separate the anomalies from normal controls using a hyperplane. These methods, however, suffer from a poor computational scalability and the curse of dimensionality [30]. By contrast, deep learning specializes in automatically learning feature representations from large-scale data. Many research efforts have been devoted to transfer the advantages of deep learning to anomaly detection. Ruff *et al.* [30] presented a deep SVDD model, which uses a deep neural network to minimize the volume of a hypersphere. Schlegl *et al.* [31] proposed a fast unsupervised anomaly detection framework with generative adversarial networks (f-AnoGAN), which is able to detect the unseen anomalies from health subjects after being trained on healthy OCT images. Although superior to traditional methods, these deep anomaly detection models, also known as novelty detection, employ unsupervised deep learning techniques, such as autoencoders and generative adversarial networks, to characterize the normal class, without using the information provided by anomalies [18], [19], [32]. Differently, we leverage the anomaly data, *i.e.*, clinically available viral pneumonia, in the training procedure, using them as the prior knowledge to reinforce the one-class model.

III. METHODS

The proposed CCAD model is composed of an anomaly detection network and a confidence prediction network
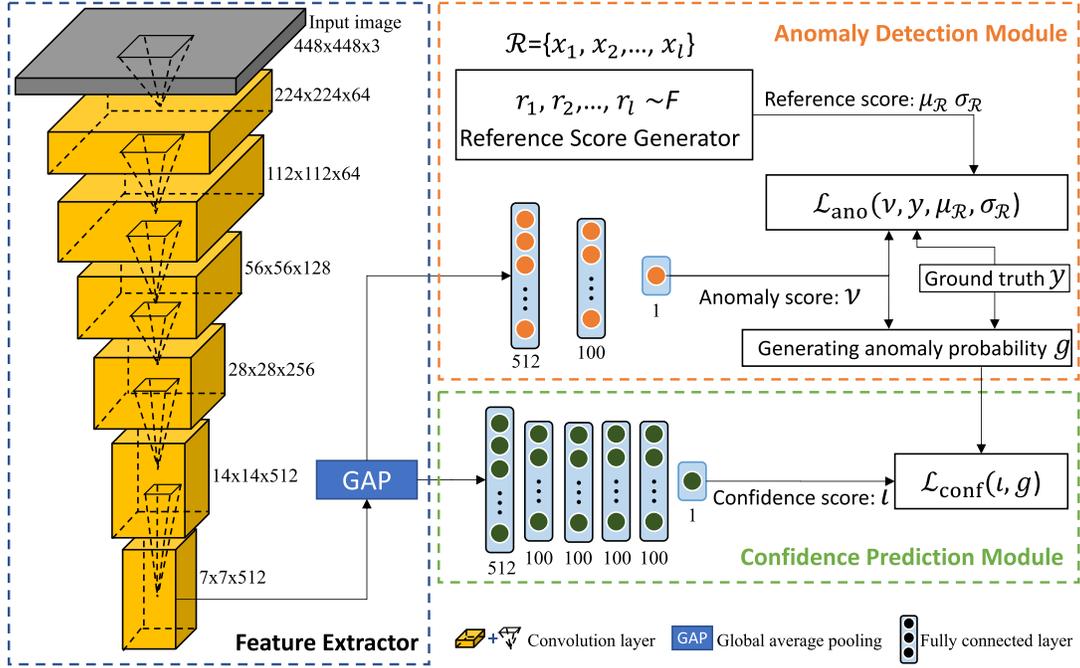
Fig. 2. Diagram of the proposed CAAD model. This model is composed of an anomaly detection module and a confidence prediction module, which are designed to predict the anomaly score and confidence score of each input, respectively.

(see Fig. 2). Both networks share a feature extractor. Given an input chest X-ray image $\mathbf{x}$, the anomaly detection network aims to learn an anomaly scoring function $\varphi : \mathbf{x} \to \mathbb{R}$. For any two inputs $\mathbf{x}_i$ and $\mathbf{x}_j$, we have $\varphi(\mathbf{x}_i) > \varphi(\mathbf{x}_j)$ if $\mathbf{x}_i$ is abnormal and $\mathbf{x}_j$ is normal. The confidence prediction network targets at approximating a confidence scoring function $\zeta : \mathbf{x} \to [0, 1]$, where 1 indicates highest model confidence and 0 indicates the opposite. In the inference mode, if the anomaly score is larger than T_{ano} or the confidence score is less than T_{conf}, we accept the input as an anomaly case (*i.e.*, viral pneumonia). We now delve into each part of our model.

A. Feature Extractor

Although a DCNN with any architecture can be embedded in our CAAD model as the feature extractor, we choose the state-of-the-art EfficientNet [33] with the B0 architecture pretrained on ImageNet [34], due to the trade-off between the performance and complexity. This network is mainly composed of mobile inverted bottleneck (MBConv) blocks [35], [36] with squeeze-and-excitation module [37]. The six stages of layer-by-layer convolution operations are represented by yellow rectangles in Figure 2, and the architecture details are listed in Table I. For each input chest X-ray image $\mathbf{x}$, it is first processed by several MBConv blocks, and then transformed into a d-dimensional (d equals the number of channels in the last convolution layer) feature vector by a global average pooling layer. We denote the parameters of feature extractor as $\boldsymbol{\theta}$.

B. Anomaly Detection Network

The anomaly detection network is composed of the feature extractor and anomaly detection module that is a multi-layer

TABLE I
ARCHITECTURE OF OUR FEATURE EXTRACTOR. S: STAGE, L: NUMBER OF STACKED LAYERS, *W*/*H*/*C*: WEIGHT/HEIGHT/CHANNEL, GAP: GLOBAL AVERAGE POOLING

S	Operator	Input $\rightarrow$ Output ($W \times H \times C$)	L
1	Conv3x3	$448 \times 448 \times 3 \rightarrow 224 \times 224 \times 32$	1
	MBConv1, k3x3	$224 \times 224 \times 32 \rightarrow 224 \times 224 \times 16$	1
2	MBConv6, k3x3	$224 \times 224 \times 16 \rightarrow 112 \times 112 \times 24$	2
3	MBConv6, k5x5	$112 \times 112 \times 24 \rightarrow 56 \times 56 \times 40$	2
4	MBConv6, k3x3	$56 \times 56 \times 40 \rightarrow 28 \times 28 \times 80$	3
	MBConv6, k5x5	$28 \times 28 \times 80 \rightarrow 28 \times 28 \times 112$	3
5	MBConv6, k5x5	$28 \times 28 \times 112 \rightarrow 14 \times 14 \times 192$	4
	MBConv6, k3x3	$14 \times 14 \times 192 \rightarrow 14 \times 14 \times 320$	1
6	Conv3x3	$14 \times 14 \times 320 \rightarrow 7 \times 7 \times 320$	1
	GAP	$7 \times 7 \times 320 \rightarrow 1 \times 1 \times 320$	1

perceptron with three 100-neuron hidden layers and an one-neuron output layer. It aims to generate an anomaly score for each input image $\mathbf{x}$, formulated as

$$\nu = \varphi(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\alpha}) \qquad (1)$$

where $\boldsymbol{\alpha}$ is the trainable parameter of the anomaly detection module.

Extensive results in [38] show that Gaussian distribution fits the anomaly scores very well in a range of datasets. To guide the learning procedure of the anomaly detection module, we compute another scalar score as a reference. We randomly sample l scalar values from an univariate Gaussian distribution, *i.e.*, $r_1, r_2, \ldots, r_l \sim \mathcal{N}(\mu, \sigma^2)$, and define a reference score as $\mu_R = \frac{1}{l} \sum_{i=1}^{l} r_i$ and $\sigma_R^2 = \frac{1}{l} \sum_{i=1}^{l} (r_i - \mu_R)^2$. Following [19], we set $\mu = 0$, $\sigma = 1$, and $l = 5,000$. With the obtained anomaly score and reference score, we employ the following contrastive loss [39], [40] to optimize the anomaly

detection module

$$\mathcal{L}_{ano}(v, y, \mu_R, \sigma_R) = (1 - y) \left| \frac{v - \mu_R}{\sigma_R} \right|$$
$$+ y \max \left(0, \mathtt{margin} - \frac{v - \mu_R}{\sigma_R}\right) \quad (2)$$

where σ_R is the standard deviation of the anomaly scores of randomly selected l normal data, y is the ground truth label, *i.e.*, $y = 0$ indicates that the input is a negative case and $y = 1$ indicates that the input is a positive case. Besides, $\mathtt{margin}$ represents the Z-score confidence interval parameter, which is empirically set to 5 for this study.

Different from binary classification where performance can be largely degraded when there is imbalanced class distribution, our one-class classification-based anomaly detection network is inherently resilient to the class imbalance. The reason is that the anomaly detection network aims to learn a one-class description model from the large-scale negative data (*i.e.,* non-viral pneumonia). By doing so, it avoids modeling the positive class with the limited amount of labeled data; the limited positive data is used instead to reinforce the one-class modeling to achieve tighter one-class description.

C. Confidence Prediction Network

The current approach of anomaly detection does not have an error correction mechanism. However, we observe that the model does produce false predictions. To alleviate this issue, we follow the work of failure prediction in image classification and segmentation [25], [26], and make the shift in thinking that we can predict the failures of anomaly detection. Hence, we propose a confidence prediction network to learn a confidence score for each input, which reflects the confidence of the anomaly score estimated by our model.

1) Confidence Criterion for Anomaly Detection: The predicted anomaly score $v \in \mathbb{R}$ explicitly describes the abnormality degree of a given image, varying from very confirmed positive cases, *i.e.*, viral pneumonia ($v >= \mathtt{margin}$), to confirmed negative cases, *i.e.*, non-viral pneumonia or healthy patients ($v \approx 0$). However, it is difficult to describe the abnormality degree in the form of probability, which is important for confidence prediction. Hence, we propose to employ the probability density function (PDF) to estimate the prediction probability. In order to normalize v, we introduce the Gaussian PDF as

$$\mathrm{PDF}(v) = \frac{1}{\sigma \sqrt{2\pi}} \exp(-\frac{(v - \mu)^2}{2\sigma}), \quad (3)$$

We then approximate the prediction probability of anomaly detection by the normalized PDF, where PDF is commonly regarded as relative probability and the normalization scales the values into [0, 1], similarly to the use of sigmoid, hyperbolic tangent, or softmax functions in deep learning. The approximated prediction probability is expressed as:

$$\mathrm{prob} = \frac{\mathrm{PDF}(v)}{\max(\mathrm{PDF})}$$
$$= \exp(-\frac{(v - \mu)^2}{2\sigma}). \quad (4)$$

Even if the scores estimated by the model is not Gaussian, it will not affect the computation of the prediction probability. However, it is hard to use such a prediction probability to distinguish failure predictions from successful ones (discussed in Section V-B). To further address this issue, we propose the anomaly probability, formulated as:

$$g = \begin{cases} \mathrm{prob} & \text{if } y = 0 \\ 1 - \mathrm{prob} & \text{if } y = 1 \end{cases} \quad (5)$$

where $g \in [0, 1]$ is the confidence criterion for distinguishing successful and erroneous predictions. Intuitively, a robust model should successfully predict the true labels of input cases with a high confidence, whereas should have low confidence when making erroneous predictions. Therefore, in Eq. (5), g is close to 0 when the anomaly detection module fails to predict the true labels of given image and close to 1 when the true labels are successfully predicted.

2) Confidence Prediction Network: Our confidence prediction network is also built upon the shared feature extractor, and particularly contains the confidence prediction module with four 100-neuron hidden layers (see Fig. 2) for a strong prediction ability, as done in [25]. The forward computation of the confidence prediction network can be formally expressed as

$$\iota = \zeta(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\beta}) \quad (6)$$

where ι is the confidence score of the corresponding anomaly detection, and $\boldsymbol{\beta}$ represents the ensemble of parameters of this module.

Since the confidence score ι takes a value from the range [0, 1], we formulate confidence prediction as a regression task and employ the standard $l2$ loss to optimize the confidence prediction module.

$$\mathcal{L}_{conf}(\iota, g) = |\iota - g|^2 \quad (7)$$

D. Training and Inference

We resize each training image to a fixed size of 512×512 pixels and applied several data augmentation strategies, including random cropping patches of size 448×448 and zooming (90%~110%) and horizontally flipping cropped patches, to alleviate overfitting on the it. Then, the proposed CAAD model is trained in three steps, Following [25], the proposed CAAD model was trained in three steps. First, we trained the anomaly detection network, which is the combination of the anomaly detection module and shared feature extractor, via minimizing the loss $\mathcal{L}_{ano}$ using the vanilla stochastic gradient descent (SGD) algorithm. Second, we fixed the feature extractor and trained the confidence prediction network, which is the combination of the confidence prediction module and shared feature extractor, via minimizing the loss $\mathcal{L}_{conf}$ using the Adam algorithm. In both steps, we set the max epoch to 20, the batch size to 40, and the initial learning rate to 5×10^{-4}, which decays linearly to 10^{-6} during the entire training process. Third, we fine-tuned the entire model in an end-to-end manner via minimizing the combination of $\mathcal{L}_{ano}$ and $\mathcal{L}_{conf}$ with the Adam optimizer. Here, the weight

TABLE II
COMPARISON OF RESULTS OF BINARY CLASSIFICATION MODELS AND THE ANOMALY DETECTION MODEL ON THE X-VIRAL DATASET

Mode		Feature extractor	Performance %			
Binary classification	Anomaly detection		Accuracy	Sensitivity	Specificity	AUC
✓		ResNet	78.52	78.28	78.56	86.24
	✓	ResNet	80.04	84.44	79.34	87.18
✓		EfficientNet	78.71	79.09	78.65	86.30
	✓	EfficientNet	**80.65**	**85.51**	**79.87**	**87.42**

Algorithm 1: Training CAAD Model

Input: $\mathfrak{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ - training data and labeled ground truth. Initialize $\boldsymbol{\theta}$ of feature extractor with the pretrained weights in ImageNet. Randomly initialize $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Output: Anomaly detection network φ, and Confidence prediction network ζ.

1: $-$ **Step1**: *Training anomaly detection network*
2: **while** *not converge* **do**
3:    Randomly sample a batch with m samples with half positive cases and half negative cases
4:    Randomly sample l scalar values from Gaussian distribution to compute μ_R, σ_R as the reference score
5:    Compute the anomaly score ν_i via Eq. (1) for each sample $\mathbf{x}_i$
6:    Compute the anomaly detection loss

$$\frac{1}{m} \sum_{i=1}^m \mathcal{L}_{ano}(\nu_i, y_i, \mu_R, \sigma_R)$$

7:    Update parameters $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ by using back-propagation
8: **end while**
9: $-$ **Step2**: *Training confidence prediction network*
10: Fix the feature extractor $\boldsymbol{\theta}$ and anomaly detection module $\boldsymbol{\alpha}$
11: **while** *not converge* **do**
12:    Randomly sample a batch with m samples
13:    Compute the confidence score ι_i via Eq. (6) for each sample $\mathbf{x}_i$
14:    Generate the anomaly probability g_i for each sample
15:    Compute the confidence prediction loss $\frac{1}{m} \sum_{i=1}^m \mathcal{L}_{conf}(\iota_i, g_i)$
16:    Update parameters $\boldsymbol{\beta}$ by using back-propagation algorithm.
17: **end while**
18: $-$ **Step3**: *Joint training in an end-to-end manner*
19: **while** *not converge* **do**
20:    Randomly sample a batch with m samples
21:    Compute μ_R, σ_R
22:    Compute ν_i and ι_i, then generate g_i according to Eq. (5)
23:    Update parameters $\boldsymbol{\theta}, \boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ by using back-propagation in an end-to-end manner.
24: **end while**

ratio of $\mathcal{L}_{ano}$ and $\mathcal{L}_{conf}$ was set to 1 : 1. In this step, we set the max epoch to 10, the batch size to 40, and the initial learning rate to 10^{-7}, aiming to avoid deviating too much

from the original anomaly detection scores. Note that the shared feature extractor was pretrained on ImageNet, and the anomaly detection module and confidence prediction module were randomly initialized.

In the inference stage, we input a test image into the well-trained model and generate a scalar anomaly score ν and a confidence score ι via the forward propagation. For the anomaly detection, we assume $g = 0.5$ in Eq. (5) as the boundary point. According to Eq. (4) and Eq. (5), we can compute the corresponding boundary anomaly score $\nu \approx 1.18$. Therefore, we set $T_{ano} = 1.18$ as the threshold to distinguish the abnormal cases and normal cases, *i.e.*, detected as abnormal if $\nu \geq T_{ano}$ or otherwise. As for the confidence prediction, we empirically set $T_{conf} = 0.9$ as the threshold to correct the erroneous predictions with low confidence. In practice, we only re-label the predictions, recognized as normal or abnormal, with low confidence, *i.e.*, $\iota < T_{conf}$, as abnormal cases to achieve the high sensitivity, which is significant in clinical study. Therefore, the final diagnosis made by our CAAD model is formulated as:

$$diag = \begin{cases} 1 & if \ \nu \geq T_{ano} \ or \ \iota < T_{conf} \\ 0 & if \ \nu < T_{ano} \ and \ \iota \geq T_{conf} \end{cases} \tag{8}$$

Specifically, if either condition is met (*i.e.*, the anomaly score is larger T_{ano} or the confidence score is less than T_{conf}), our model gives a 1 : POSITIVE diagnosis and recommend to be further examined by radiologists; otherwise our model gives a 0 : NEGATIVE diagnosis.

E. Performance Metrics

For this study, the diagnostic performance of an algorithm is quantitatively assessed by the area under the receiver operator curve (AUC), sensitivity, specificity, and, accuracy. AUC reflects the probability that a recognition model ranks a randomly chosen positive instance higher than a randomly chosen negative case. It is the most commonly used metric to evaluate the overall classification performance. The sensitivity and specificity give the proportion of positives and negatives that were correctly identified, respectively. Accuracy gives the percentage of correctly classified cases, including both positive and negative ones.

IV. DATASETS

Two in-house X-ray image datasets, X-VIRAL and X-COVID, were used for this study. The X-VIRAL dataset contains 5,977 viral pneumonia cases, 18,619 non-viral pneumonia cases, and 18,774 healthy controls (*i.e.*, 5977 positive and 37,393 negative cases) collected from 390 township
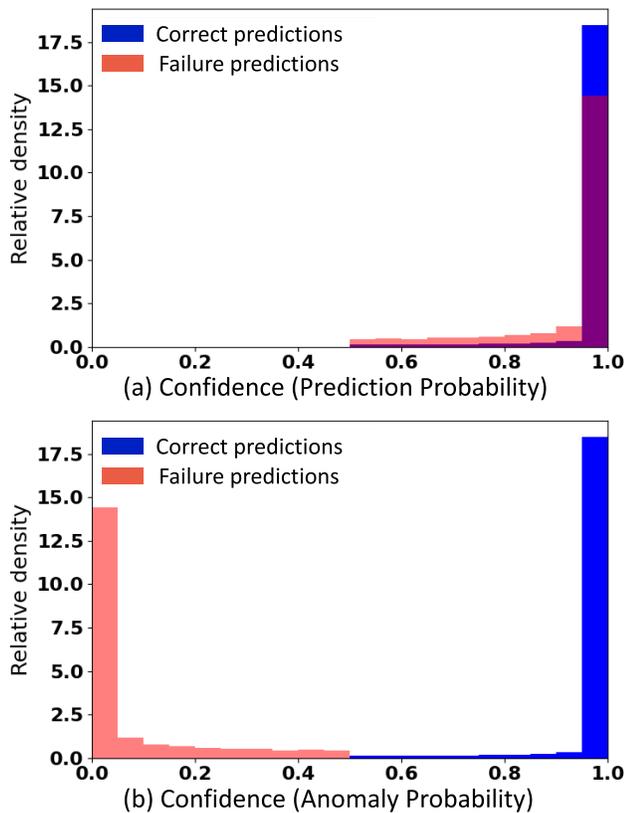
Fig. 3. Comparison of confidence learning based on (a) the prediction probability and (b) anomaly probability for failure prediction on the validation set of X-VIRAL.

hospitals through a telemedicine platform of JF Healthcare during 2019. Each X-ray image has a high resolution, varying from 1000 to 3000 for height and weight, and was annotated by one of three board-certified radiologists. Note that all viral pneumonia cases were collected before the COVID-19 outbreak, and hence do not contain any COVID-19 cases. The X-COVID dataset was collected from 6 institutions during March 2020. It consists of 106 confirmed COVID-19 cases and 107 normal controls. Besides, a public COVID-19 dataset[1] [41], called Open-COVID, was used for external validation. This dataset contains the X-ray images of 493 confirmed COVID-19 cases, 16 confirmed SARS [3] cases, and 10 confirmed MERS [4] cases.

V. Experiments and Results

A. Anomaly Detection Vs. Binary Classification

We first compared anomaly detection models to binary classification models on the X-VIRAL dataset using the five-fold cross-validation. The feature extractor used in all models is either the 18-layer ResNet [42] or EfficientNet-B0 [33], both being pre-trained on ImageNet. The binary classification models were optimized by minimizing the cross-entropy loss function. To indulge a fair comparison, we also employed the balanced sampling strategy to keep the balance of both positive and negative categories during the training. The obtained

[1]https://github.com/ieee8023/covid-chestxray-dataset

performance metrics were listed in Table II. It shows that the binary classification using ResNet achieves the baseline performance, *i.e.*, an accuracy of 78.52%, a sensitivity of 78.28%, a specificity of 78.56%, and an AUC of 86.24%, which can be improved a little bit by introducing a stronger feature extractor, *i.e.*, EfficientNet-B0. By contrast, an anomaly detection model always outperforms (particularly in terms of sensitivity) the corresponding binary classification model. In this experiment, the anomaly detection model using EfficientNet-B0 achieves the highest accuracy of 80.65%, highest specificity of 79.87%, and highest AUC of 87.42%, and also remarkably improves the sensitivity from 79.09% to 85.51%. The results suggest that anomaly detection has distinct advantages over binary classification in distinguishing viral pneumonia cases from non-viral pneumonia cases and healthy controls, especially with an extraordinary ability to detect positive cases as evidenced by a high sensitivity, which is particularly useful for viral pneumonia screening in clinical practice.

B. Confidence Learning for Failure Prediction

To analyze the effectiveness of confidence learning using either the anomaly probability or prediction probability, we compared the distribution of prediction probability and anomaly probability obtained on the validation set of X-VIRAL in Fig. 3. It is apparent that the prediction probability distributions of correct predictions and false predictions have a lot of overlap, which hinders the confidence prediction module from accurately distinguishing them. By contrast, the proposed anomaly probability can successfully separate false predictions from correct ones.

In Fig. 3(b) and Eq. (5), we observed that, if (1) the predicted anomaly score for negative cases is close to 0 or (2) the predicted anomaly score for positive cases is larger than T_{ano}, our CAAD model has high confidence; otherwise low confidence. Such conjecture was confirmed with the results given in Fig. 4, which shows 16 chest X-ray images from the X-VIRAL validation set, each being equipped with the ground truth label and the anomaly scores and confidence scores predicted by our CAAD model. For each negative case in the top row, it shows that if the predicted anomaly scores are very low, close to 0, the confidence score is close to 1. Similarly, for each positive case in the bottom row, it shows that if the predicted anomaly score is larger than T_{conf}, the confidence score is still close to 1. In contrast, the confidence becomes very low if a case is wrongly diagnosed, as those marked with red bounding boxes. These results demonstrate the effectiveness of using the proposed anomaly probability to learn the confidence for failure prediction.

C. Importance of Confidence Prediction

To evaluate the effectiveness of confidence prediction, we compared the anomaly detection network (denoted by AnoDet) with our CAAD model, in which the confidence threshold T_{conf} ranges from 0.5 to 0.95. Note that the CAAD model can be treated as a combination of AnoDet with a confidence prediction module. The results obtained on the X-VIRAL dataset are shown in Table III. It reveals that (1)

| True label: Negative Anomaly score: 0.01 Confidence: 0.95 | True label: Negative Anomaly score:0.03 Confidence: 0.92 | True label: Negative Anomaly score:0.11 Confidence: 0.90 | True label: Negative Anomaly score:0.23 Confidence: 0.88 | True label: Negative Anomaly score:0.33 Confidence: 0.85 | True label: Negative Anomaly score:0.56 Confidence: 0.76 | True label: Negative Anomaly score:0.79 Confidence: 0.65 | True label: Negative Anomaly score:1.80 Confidence: 0.44 |

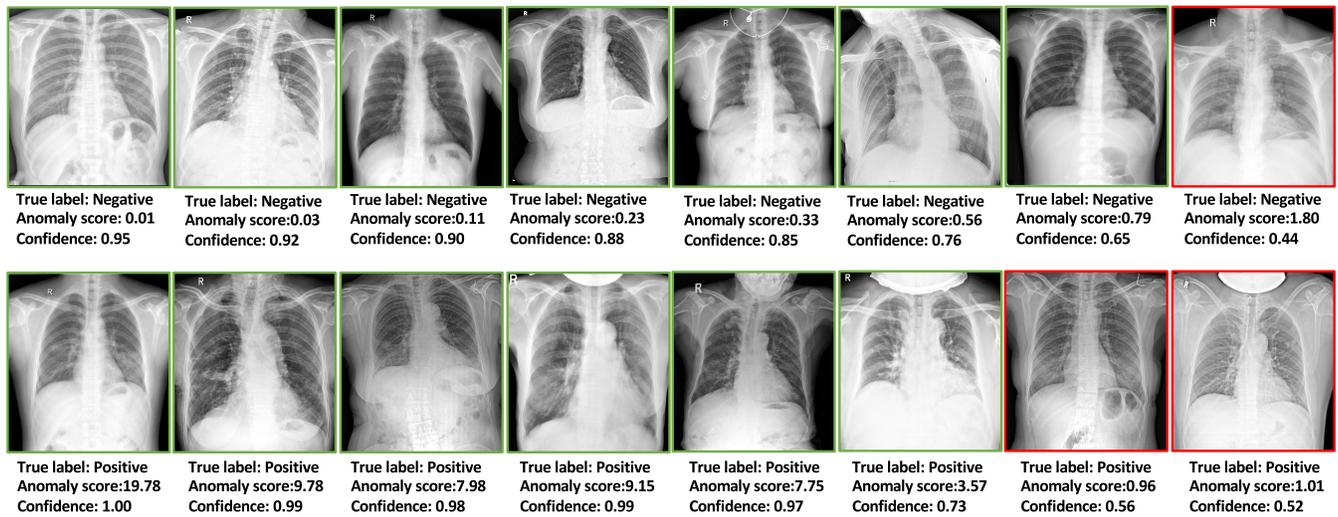| True label: Positive Anomaly score:19.78 Confidence: 1.00 | True label: Positive Anomaly score:9.78 Confidence: 0.99 | True label: Positive Anomaly score:7.98 Confidence: 0.98 | True label: Positive Anomaly score:9.15 Confidence: 0.99 | True label: Positive Anomaly score:7.75 Confidence: 0.97 | True label: Positive Anomaly score:3.57 Confidence: 0.73 | True label: Positive Anomaly score:0.96 Confidence: 0.56 | True label: Positive Anomaly score:1.01 Confidence: 0.52 |

Fig. 4. A set of 16 chest x-ray images from the X-VIRAL validation set. The predicted anomaly score and confidence score are displayed beneath each image. Positive cases are shown in the top row, and negative cases are shown in the bottom row. Totally, 13 cases (marked with green boxes) were corrected diagnosed by our CAAD model, and three (marked with red boxes) were wrongly diagnosed.
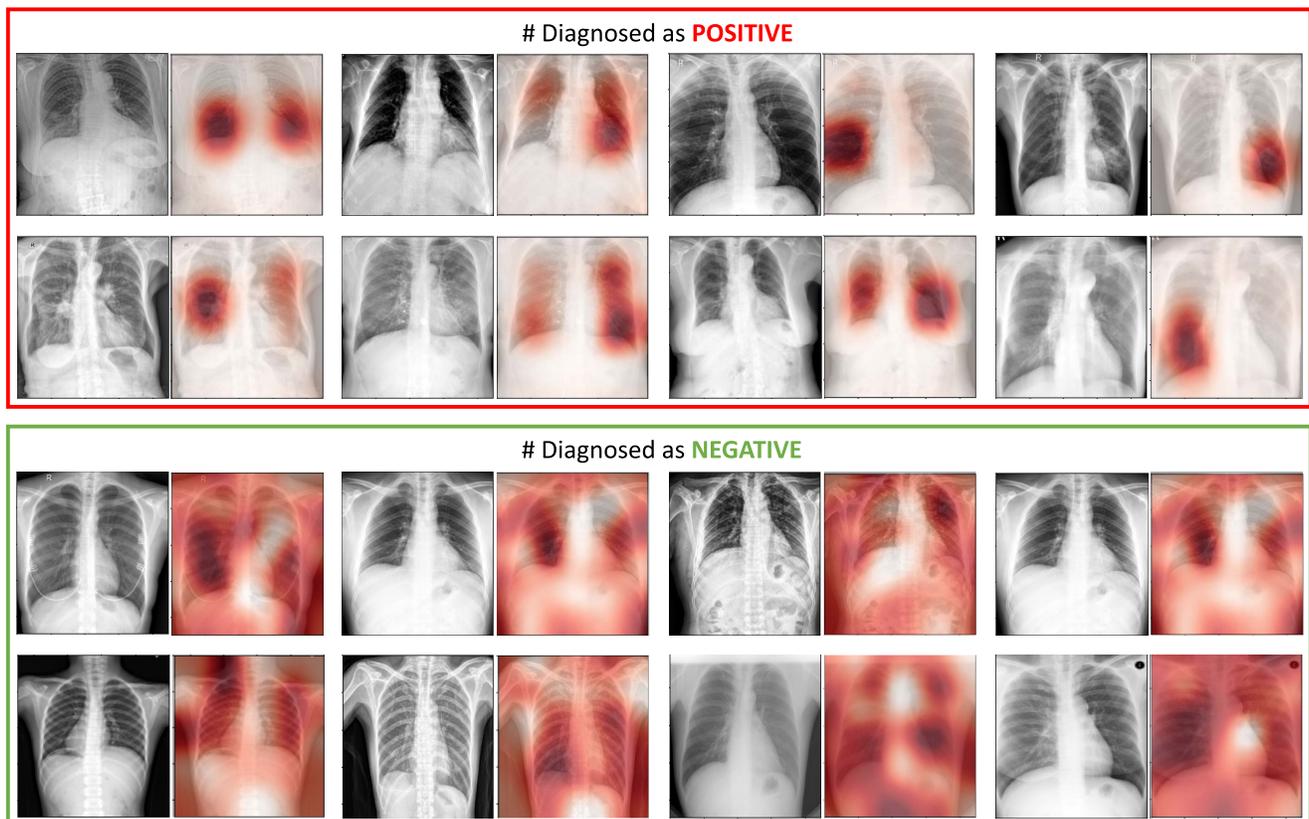

Fig. 5. Visualization of 16 chest X-ray images from the X-VIRAL validation set and their Grad-CAMs. The images in two top rows are diagnosed as positive by our CAAD model, while the images in two bottom rows are diagnosed as negative.

using confidence prediction leads to a slightly improved AUC of 87.57%, improved sensitivity, and deteriorated accuracy and specificity; (2) when setting the confidence threshold T_{conf} to 0.5, the performance of our model is comparable to that of AnoDet; and (3) with the increase of the confidence threshold T_{conf}, the deterioration of accuracy and specificity becomes severe and the improvement of sensitivity becomes substantial. Specifically, when setting the confidence threshold T_{conf} to 0.9, our CAAD model can boost the sensitivity from 85.51% to 93.01% while having a specificity of 67.72%. The improvement in the sensitivity suggests the advantage of using confidence prediction in our model for the clinical screening

TABLE III

PERFORMANCE OF THE ANODET MODEL (WITHOUT CONFIDENCE PREDICTION) AND OUR CAAD MODEL (WITH VARIABLE CONFIDENCE THRESHOLD T_{conf}) ON THE X-VIRAL DATASET. NOTE THAT WE SET THE SAME THRESHOLD $T_{ano} = 1.18$ FOR BOTH MODELS

Methods	T_{conf}	Accuracy	Sensitivity	Specificity	AUC
AnoDet	/	80.65	85.51	79.87	87.42
CAAD	0.5	80.33	85.88	79.44	
	0.6	79.47	87.02	78.27	
	0.7	78.48	88.05	76.95	87.57
	0.8	76.79	89.63	74.74	
	0.9	71.21	93.01	67.72	
	0.95	46.44	97.69	38.25	

TABLE IV

COMPARING PERFORMANCE OF FOUR MODELS (W/O FINE-TUNING) ON THE X-COVID DATASET

Model	Performance %			
	Accuracy	Sensitivity	Specificity	AUC
EfficientNet [33]	69.95	45.28	94.39	74.45
ConfiNet [25]	68.08	69.81	66.36	74.89
AnoDet [19]	73.24	55.66	90.65	82.97
CAAD	72.77	71.70	73.83	83.61

Fig. 6. Comparison of different models with and without using fine-tuning on the X-COVID dataset.

of viral pneumonia. It should noted that, as mentioned in Section III-C, we propose the confidence prediction module to distinguish the successful predictions with a high confidence from failed predictions with a low confidence, instead of further improving the performance of anomaly detection.

D. Visualizing Region of Diagnosis

For the visual explanation of the decision reasoning of our CAAD model, we adopted the gradient-weighted class activation mapping (Grad-CAM) [43] to "see" which regions play an important role during the inference. Fig. 5 shows 16 chest X-ray images from the X-VIRAL validation set, each being accompanied with the Grad-CAM maps overlaid on it. Eight cases in two top rows were diagnosed as positive by our CAAD model, while the other cases in two bottom rows were diagnosed as negative. It reveals that our CAAD model is able to focus on the suspected lesions and, accordingly, diagnose the input image as POSITIVE. However, if no highly suspected lesion is found, our model assigns the homogeneous activation values to almost the entire image. The visualization demonstrates the good interpretability of our CAAD model.

E. Generalize to Unseen X-COVID Dataset

To demonstrate its generalization ability, the well-trained CAAD model was directly tested on the unseen X-COVID dataset without fine-tuning. It was also compared to a binary classifier using EfficientNet [33], a binary classifier with confidence prediction (*i.e.*, ConfidNet) [25], and a anomaly detection model (*i.e.*, AnoDet) [19]. Note that all of these models were never trained on the COVID-19 cases. TABLE IV gives the performance of these models on the X-COVID dataset. It reveals that (1) anomaly detection models are superior to
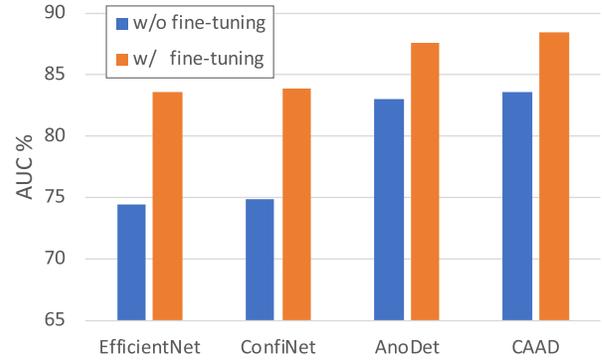
both binary classifiers, especially in terms of AUC, which demonstrates the effectiveness of anomaly detection in viral pneumonia screening; (2) confidence prediction is an effective strategy to predict failures in both a classifier and an anomaly detector, contributing to a big improvement in sensitivity; and (3) the performance of these models, however, drops when comparing to their performance in viral pneumonia screening. Compared to binary classification methods, anomaly detection based AnoDet and CAAD, as one-class classification approaches, have a better ability to detect dissimilar or even previously unseen COVID-19 cases. In summary, we expect that our model trained on the X-VIRAL dataset would have the ability to detect unseen COVID-19 cases as anomaly. Our results suggest that, despite the significant drop of sensitivity, our CAAD model still achieves an AUC of 83.61% and a sensitivity of 71.70% for COVID-19 screening, the highest performance obtained in our experiment.

To further improve their performance on the X-COVID dataset, we fine-tuned EfficientNet, ConfiNet, AnoDet, and our CAAD model, which had been well-trained on the X-VIRAL dataset, on the X-COVID dataset. We conducted the two-fold cross-validation and compared the AUC values obtained by these models in Fig. 6. It shows that fine-tuning is an effective strategy to improve the performance of all four models. Although fine-tuning narrows the gaps between the classification-based and anomaly detection-based models, it reveals that classification-based models suffer from a poorer generalization ability than anomaly detection-based models.

F. Generalize to Unseen Open-COVID Dataset

The EfficientNet, ConfiNet, AnoDet, and our CAAD model trained on the X-VIRAL dataset were further tested on the Open-COVID dataset without any fine-tuning. The performance of these models was displayed in TABLE V. It shows that our CAAD model achieves an accuracy of 94.93% for COVID-19 detection and an accuracy of 100% for SARS and MERS detection, which are superior to those of three competing methods. Such results indicate that our CAAD model has a strong generalization ability on the unseen Open-COVID dataset.

Besides, we randomly split 2,000 negative cases (1,000 non-viral pneumonia and 1,000 healthy controls) from the
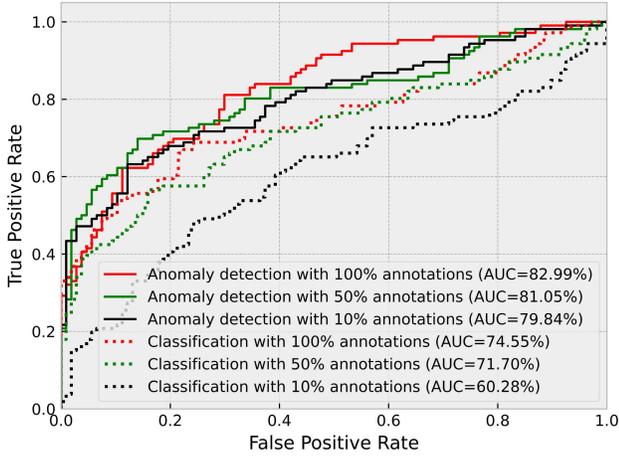
Fig. 7. AUC curves of anomaly detection and binary classification models obtained on X-COVID when both models were trained on X-VIRAL using all negative samples and 100%, 50%, and 10% positive samples, respectively.

TABLE V

COMPARING PERFORMANCE OF FOUR MODELS ON THE OPEN-COVID DATASET

Model	Detection accuracy (sensitivity) %			
	COVID (493 images)	SARS (16 images)	MERS (10 images)	All (519 images)
EfficientNet [33]	63.08	93.75	80.00	64.35
ConfiNet [25]	90.26	**100.00**	**100.00**	90.75
AnoDet [19]	83.37	**100.00**	**100.00**	84.20
CAAD	**94.93**	**100.00**	**100.00**	**95.18**

X-VIRAL dataset, combined them with the healthy controls from the X-COVID dataset, and also combined the COVID-19 cases from the X-COVID and Open-COVID datasets. Thus, we have a new dataset that contains 599 positive COVID-19 cases and 2107 negative cases. In TABLE VI, we compared the performance of two binary classification based methods (*i.e.*, EfficientNet and ConfiNet) and two anomaly detection based methods (*i.e.*, AnoDet and CAAD) on this dataset. Note that all these models never saw any COVID-19 X-ray images in the X-COVID and Open-COVID datasets. It shows that anomaly detection based models, especially our CAAD model, perform distinctly better than binary classification based models. Meanwhile, we also compared the performance of ConfiNet and CAAD when setting the confidence threshold T_{conf} to different values. It reveals that, when we set T_{conf} to 0.5, both models can achieve not only the highest accuracy, but also high and balanced sensitivity and specificity. Moreover, comparing to the results in TABLE VI and TABLE IV, it shows that all four models, especially both binary classification based ones, achieve better performance on this dataset than on the X-COVID dataset. The performance gain can be largely attributed to the high accuracy of COVID-19 detection on the Open-COVID dataset, as shown in TABLE V.

G. Learning With Less Positive Samples

In the real-word scenario, it is much more difficult to collect positive (*i.e.*, viral pneumonia) samples than to collect negative samples. Hence automated viral pneumonia screening has

TABLE VI

COMPARING PERFORMANCE OF FOUR MODELS ON THE X-COVID AND OPENCOVID DATASET (599 POSITIVES, 2107 NEGATIVES)

Method	T_{conf}	Acc	Sen	Spe	AUC
EfficientNet [33]	/	75.87	62.6	79.64	78.92
ConfiNet [25]	0.95	34.85	96.33	17.37	79.10
	0.9	46.45	92.65	33.32	
	0.8	57.5	86.14	49.36	
	0.7	65.56	82.47	60.75	
	0.6	70.18	76.63	68.34	
	0.5	76.39	63.27	80.11	
AnoDet [19]	/	79.79	69.78	82.63	83.34
CAAD	0.95	42.46	97.33	26.86	84.43
	0.9	69.73	90.65	63.79	
	0.8	75.09	84.14	72.52	
	0.7	76.72	80.8	75.56	
	0.6	77.72	78.63	77.46	
	0.5	78.57	77.13	78.97	

a class-imbalance nature. Anomaly detection methods avoid modeling the limited positive classes explicitly and hence better handle the class-imbalance issue than classification methods. To verify this, we trained the anomaly detection model and binary classification model, both using EfficientNet as the feature extractor, under much more class imbalanced conditions, *i.e.*, using less positive samples, and then tested both without further fine-tuning on the X-COVID dataset. Specifically, in the training stage, we kept the number of negative samples unchanged and reduced the positive samples from 100% to 50% and 10%, respectively. The obtained AUC curves are shown in Fig. 7. It reveals that training with less positive data leads to poor performance for both anomaly detection and binary classification models. Nevertheless, the anomaly detection model shows less performance degradation compared to the binary classification model, which indicates the advantage of using anomaly detection in terms of addressing the imbalanced problem.

VI. DISCUSSION

In the escalation of the COVID-19 epidemic, many attempts have been made to develop fast and accurate COVID-19 screening by means of chest medical imaging [44]–[47]. Kang *et al.* [45] leveraged different types of features extracted from chest CT images, and introduced a multi-view representation learning method to distinguish positive COVID-19 from negative cases. Shi *et al.* [47] presented an infection size aware random forest model to classify 1658 positive subjects confirmed COVID-19 and 1027 negative subjects without COVID-19 infection using chest CT, and achieved the sensitivity of 90.70% and specificity of 83.30% on this large-scale CT dataset. However, CT imaging takes considerably more time than X-ray imaging, and needs more complex sanitization procedures between switching patients. Besides, sufficient high-quality CT scanners may not be available in many under-developed regions, making it difficult for a timely viral pneumonia screening. In contrast, X-ray imaging is the most common and widely available chest imaging technique, playing a crucial role in clinical care and epidemiological studies [48], [49]. Most ambulatory care facilities, even in rural

regions, has X-ray imaging capability. Besides, X-ray imaging is real-time which can significantly speed up the screening of a mass population in a relatively short time and at a significantly reduced cost. Wang *et al.* [50] and Apostolopoulos *et al.* [51] introduced DCNN-based binary classification models for the detection of COVID-19 cases using chest X-Ray imaging.

Different from these COVID-19 screening works, we view COVID-19 as a novel type of viral pneumonia and attempt to distinguish it, together with other types of viral pneumonia, from non-viral pneumonia and healthy controls. To this end, we reformulate the binary classification problem in an anomaly detection fashion. Besides, we introduce a confidence prediction module to estimate the reliability of model diagnosis by learning an anomaly probability as the model confidence. The proposed CAAD model achieves an AUC of 83.61% on COVID-19 screening, which outperforms other AI-based methods [52]. Although achieving a sensitivity of only 71.70%, our CAAD model shows a screening ability that is comparable to that of radiologists, as a sensitivity of 69% was reported in [20]. The reason of such a low sensitivity may attribute to the observation that some subjects have not developed radiographic visible pathology in their lungs at the early stage of viral pneumonia when the X-ray was taken [53].

VII. Conclusion

In this paper, we have proposed the CAAD model for viral pneumonia screening. Our results on two chest X-ray datasets indicate that (1) anomaly detection works well in term of viral pneumonia screening on chest X-ray images and is superior to binary classification methods, and (2) learning model confidence is useful to predict failures, greatly reducing the false negatives, and (3) our CAAD model, never seeing any COVID-19 cases, achieves an AUC of 83.61% and sensitivity of 71.70% on the unseen X-COVID dataset, which is comparable to the performance of medical professionals. Our future work will focus on further reducing the false negative rate and, if possible, decreasing the false positive rate as well. We will also investigate how to differentiate the viral pneumonia severity using chest X-ray and then detect the potentially severe cases for early interventions, which requires more clinical diagnostic information.

References

[1] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, and M. Wang, "Presumed asymptomatic carrier transmission of COVID-19," *Jama*, vol. 323, no. 14, pp. 1406–1407, 2020.

[2] H. Chen *et al.*, "Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: A retrospective review of medical records," *Lancet*, vol. 395, no. 10226, pp. 809–815, Mar. 2020.

[3] W. Li *et al.*, "Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus," *Nature*, vol. 426, no. 6965, pp. 450–454, 2003.

[4] E. I. Azhar *et al.*, "Evidence for Camel-to-Human transmission of MERS coronavirus," *New England J. Med.*, vol. 370, no. 26, pp. 2499–2505, Jun. 2014.

[5] V. M. Corman *et al.*, "Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR," *Eurosurveillance*, vol. 25, no. 3, Jan. 2020, Art. no. 2000045.

[6] L. Lan *et al.*, "Positive RT-PCR test results in patients recovered from COVID-19," *J. Amer. Med. Assoc.*, vol. 323, no. 15, pp. 1502–1503, 2020.

[7] J. Watson, P. F. Whiting, and J. E. Brush, "Interpreting a COVID-19 test result," *Brit. Med. J.*, vol. 369, pp. 1–7, 2020.

[8] W. H. Self, D. M. Courtney, C. D. McNaughton, R. G. Wunderink, and J. A. Kline, "High discordance of chest X-ray and computed tomography for detection of pulmonary opacities in ED patients: Implications for diagnosing pneumonia," *Amer. J. Emergency Med.*, vol. 31, no. 2, pp. 401–405, Feb. 2013.

[9] R. Joarder and N. Crundwell, *Chest X-Ray in Clinical Practice*. Springer, 2009.

[10] L. Luo *et al.*, "Deep mining external imperfect data for chest X-ray disease screening," 2020, *arXiv:2006.03796*. [Online]. Available: http://arxiv.org/abs/2006.03796

[11] 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.

[12] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-Rays with deep learning," 2017, *arXiv:1711.05225*. [Online]. Available: http://arxiv.org/abs/1711.05225

[13] H. Wang, H. Jia, L. Lu, and Y. Xia, "Thorax-net: An attention regularized deep neural network for classification of thoracic diseases on chest radiography," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 475–485, Feb. 2020.

[14] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[15] S. Rajaraman, S. Candemir, I. Kim, G. Thoma, and S. Antani, "Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs," *Appl. Sci.*, vol. 8, no. 10, p. 1715, Sep. 2018.

[16] C. Bellinger, S. Sharma, and N. Japkowicz, "One-class versus binary classification: Which and when?" in *Proc. 11th Int. Conf. Mach. Learn. Appl. (ICMLA)*, vol. 2, Dec. 2012, pp. 102–106.

[17] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 665–674.

[18] G. Pang, L. Cao, L. Chen, and H. Liu, "Learning representations of ultrahigh-dimensional data for random distance-based outlier detection," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2041–2050.

[19] G. Pang, C. Shen, and A. van den Hengel, "Deep anomaly detection with deviation networks," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 353–362.

[20] H. Y. F. Wong *et al.*, "Frequency and distribution of chest radiographic findings in COVID-19 positive patients," *Radiology*, vol. 296, no. 2, pp. E72–E78, 2020.

[21] H. Magree *et al.*, "Chest X-ray-confirmed pneumonia in children in Fiji," *Bull. World Health Org.*, vol. 83, pp. 427–433, Mar. 2005.

[22] A. Jacobi, M. Chung, A. Bernheim, and C. Eber, "Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review," *Clin. Imag.*, vol. 64, pp. 35–42, Aug. 2020.

[23] A. Borghesi and R. Maroldi, "COVID-19 outbreak in italy: Experimental chest X-ray scoring system for quantifying and monitoring disease progression," *La Radiologia Medica*, vol. 125, no. 5, pp. 509–513, May 2020.

[24] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–12.

[25] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez, "Addressing failure prediction by learning model confidence," in *Adv. Neural Inf. Process. Syst.*, 2019, pp. 2898–2909.

[26] Y. Xie, H. Lu, J. Zhang, C. Shen, and Y. Xia, "Deep segmentation-emendation model for gland instance segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 469–477.

[27] Y. Xie, J. Zhang, H. Lu, C. Shen, and Y. Xia, "SESV: Accurate medical image segmentation by predicting and correcting errors," *IEEE Trans. Med. Imag.*, early access, Sep. 21, 2020, doi: 10.1109/TMI.2020.3025308.

[28] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.

[29] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[30] L. Ruff *et al.*, "Deep one-class classification," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4393–4402.

[31] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.*, vol. 54, pp. 30–44, May 2019.

[32] M. A. Siddiqui, A. Fern, T. G. Dietterich, R. Wright, A. Theriault, and D. W. Archer, "Feedback-guided anomaly discovery via online optimization," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2200–2209.

[33] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[36] M. Tan *et al.*, "MnasNet: platform-aware neural architecture search for mobile," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2820–2828.

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[38] H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek, "Interpreting and unifying outlier scores," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, 2011, pp. 13–24.

[39] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.

[40] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020, *arXiv:2002.05709*. [Online]. Available: http://arxiv.org/abs/2002.05709

[41] J. Paul 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*. [Online]. Available: http://arxiv.org/abs/2006.11988 and https://github.com/ieee8023/covid-chestxray-dataset

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[43] 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.

[44] D. Ting, L. Carin, V. Dzau, and T. Y. Wong, "Digital technology and COVID-19," *Nature Med.*, vol. 26, no. 4, pp. 459–461, 2020.

[45] H. Kang *et al.*, "Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2606–2614, Aug. 2020.

[46] P. Zhang, Y. Zhong, X. Tang, Y. Deng, and X. Li, "Learning diagnosis of COVID-19 from a single radiological image," 2020, *arXiv:2006.12220*. [Online]. Available: http://arxiv.org/abs/2006.12220

[47] F. Shi *et al.*, "Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification," 2020, *arXiv:2003.09860*. [Online]. Available: http://arxiv.org/abs/2003.09860

[48] T. Cherian *et al.*, "Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies," *Bull. World Health Org.*, vol. 83, pp. 353–359, Sep. 2005.

[49] T. Franquet, "Imaging of pneumonia: Trends and algorithms," *Eur. Respiratory J.*, vol. 18, no. 1, pp. 196–208, Jul. 2001.

[50] L. Wang and A. Wong, "COVID-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," 2020, *arXiv:2003.09871*. [Online]. Available: http://arxiv.org/abs/2003.09871

[51] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, Jun. 2020.

[52] K. Murphy *et al.*, "COVID-19 on the chest radiograph: A multi-reader evaluation of an AI system," *Radiology*, vol. 296, no. 3, pp. E166–E172, 2020.

[53] J. Russell, A. Echenique, S. Daugherty, and M. Weinstock, "Chest X-ray findings in 636 ambulatory patients with COVID-19 presenting to an urgent care center: A normal chest X-ray is no guarantee," *J. Urgent Care Med.*, vol. 14, no. 7, pp. 8–13, 2020.