Applied Soft Computing Journal 166 (2024) 112139

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

An explainable weakly supervised model for multi-disease detection and localization from thoracic X-rays

Anwesh Kabiraj [a,1], Tanushree Meena [a,1], Kalyan Tadepalli [b,c], Sudipta Roy [a,*]

[a] Artificial Intelligence & Data Science, Jio Institute, Mumbai 410206, India
[b] Sir HN Reliance Foundation Hospital, Girgaon, Mumbai 400004, India
[c] Reliance Jio - Artificial Intelligence Centre of Excellence (AICoE), Hyderabad, India

HIGHLIGHTS

- Localization of anomaly without having annotated localization in training.
- Generalized model to achieve state of the art accuracy in terms of IoBB as well as dice scores.
- Novel class activation map pooling to achieve better results than SOTA methods.
- The CAN and ADN helps the model to utilize all components of features.
- The explainability of the overall process in each layer by showing deconvolutions and saliency maps to unboxing the black box.

ARTICLE INFO

Keywords:
Weakly supervised
Deep learning
Thoracic diseases
Medical imaging
Anomaly localization
Computer vision

ABSTRACT

Thoracic diseases are a major source of mortality, often requiring diagnosis from plain chest X-rays. However, differentiating between complex conditions based on subtle radiographic patterns poses challenges even for experts. Recently, deep learning methods have shown promise in automating thoracic disease detection from chest radiographs. Many existing approaches focus on the diseased organs in the radiographs by utilizing spatial regions that contribute significantly to the model's prediction. Expert radiologists, on the other hand, first identify the prominent region before determining whether those regions are abnormal or not. Therefore, incorporating localization information through deep learning models could result in significant improvements in automatic disease classification. Motivated by this, we have proposed a generalized weakly supervised Confidence-Aware Probabilistic Class Activation Map (CAPCAM) classification model that localizes anomalies for thoracic disease. The CAPCAM used CX-Ultranet as the backbone with the combination of Confidence Aware Network (CAN) and Anomaly Detection Network (ADN) without having any localization labeling. This learning from the backbone helps the model to utilize all components of the feature extracted and, therefore eliminating the need to train them individually reducing the time taken. We have experimentally shown that the proposed CAPCAM method sets a new state-of-the-art benchmark by achieving accuracy in terms of Intersection of bounding box (IoBB) in the range of 85 % - 94 %, and Dice scores in the range of 88 %-90 % for all thirteen diseases on two publicly available large-scale CXR datasets–NIH, Stanford and CheXpert. Testing across different noise levels and different levels of blurred level assessed real-world viability. We have also added a layer of explainability to show how the image is processed. This study demonstrates deep learning's potential to augment radiologists' decision-making by providing fast, accurate automated aids for thoracic disease diagnosis. The proposed CAPCAM model could be readily translatable to improve clinical workflows.

* Corresponding author.
E-mail addresses: anweshkabiraj17@gmail.com (A. Kabiraj), tanushree.meena@jioinstitute.edu.in (T. Meena), kalyan.tadepalli@rfhospital.org (K. Tadepalli), sudipta1.roy@jioinstitute.edu.in (S. Roy).
[1] Authors contributed equally

https://doi.org/10.1016/j.asoc.2024.112139
Received 19 April 2023; Received in revised form 9 August 2024; Accepted 13 August 2024
Available online 24 August 2024

1. Introduction

Respiratory and thoracic diseases are among the leading causes of morbidity and mortality globally [1]. While the armamentarium of tools to diagnose these diseases has significantly expanded, chest X- Rays are often the starting point of the diagnosis-to-cure pathway. Usage of chest X-rays is on the rise, a trend that is attributable both to the ease of availability as well as the low amount of radiation exposure involved. Plain chest radiographs (CXRs) frequently display multiple thoracic abnormalities including atelectasis, cardiomegaly, effusion, pleural thickening, and pneumonia. However, extracting accurate diagnostic information from these complex CXRs remains challenging. Radiographic interpretation relies extensively on the diagnostic acumen and expertise of the individual radiologist. Additionally, real-world clinical CXRs often contain noise and blurriness from various sources, which can potentially lead to misdiagnoses if the noise obscures or mimics findings. Recent computational applications in clinical studies show that the performance of radiologists improves with the use of deep learning (DL) models. Furthermore, deep learning models can also detect missed findings beyond a human's scope, which can minimize the reader-dependent issue among radiologists [2]. Hence, a DL model without using annotated data to detect thoracic diseases is clinically beneficial as data annotation is extremely labor-intensive and thus error prone.

Many traditional thoracic DL systems are based on manual feature extraction and use machine learning (ML) techniques to analyse shape texture, contrast, intensity, and shape, among others [3]. The radiographic manifestations of many thoracic diseases can demonstrate substantial overlap and subtleties that pose challenges for accurate differential diagnosis. Owing to the complex, heterogeneous presentation of these diseases on imaging, developing a single standardized feature set for multi-class recognition on chest radiographs has remained an open problem. Therefore, enabling automated systems to reliably detect multiple concurrent thoracic diseases from chest radiographs is an important capability that would assist clinicians by providing data-driven diagnostic decision support. In the recent years, DL techniques, particularly convolutional neural networks (CNN) have developed the potential to automatically detect multiple diseases from a chest radiograph. However, in clinical care, these DL techniques continue to confront the issues like noisy data, robustness, explainability, data annotation, generalizability, and reliability challenges. Therefore, in this work, we propose a robust model for localization of thoracic diseases which can address all the aforementioned challenges. We have performed extensive experiments to justify our claim. The contributions of this paper are summarized as follows:

- Localization of anomaly without having annotated localization in training.
- An efficient, weakly supervised, and generalized model to achieve state-of-the-art accuracy in terms of IoBB as well as dice scores.
- Novel probabilistic class activation map (CAM) pooling to achieve better results than other state-of-the-art pooling methods.
- Different components in our network, like CAN and ADN help the model to utilize all components of feature extracted therefore eliminating the necessity to train them individually and hence time taken.
- The explainability of the overall process in each layer by showing deconvolutions and saliency maps to find out area of interest ultimately helps in unboxing the black box.
- Noise sensitivity analysis to make the model robust for clinical usage.

The rest of the paper is organized as follows. Section 2 discusses the related work followed by methodology in Section 3. Experimental details, results and discussions are in Sections 4 and 5 respectively. In the last section, we summarize our contributions.

2. Related work

2.1. Weakly supervised learning

In recent years, class activation mapping (CAM) [4] and gradient-weighted CAM (Grad-CAM) [5] have emerged as weakly supervised techniques that provide explainability into model classification decisions by highlighting discriminative input image regions most influential on the output, without requiring extensive localization labelling data. The concept of CAM [4] was introduced for the visualization of features which a model uses to make the final decision. In [5] Grad-CAM was introduced for the purpose of visualization of important features generated during back propagation algorithm using the gradient information. These are adaptable to a wide range of models including the ones used for image-capturing [6,7] and classification. Both CAMs and Grad CAMs depict only the most discriminative region instead of the entire regions for the purpose of classification. To solve the above issue, several weakly supervised segmentation and recognition algorithms were developed [8] that recognizes the location of the object from the image-level information. The grid drop method [9] was proposed where an input image is segmented into a m-by-m matrix and each matrix is dropped randomly with a specific probability. The transformed image is then utilized again and again for training, allowing a network to locate the entire region of the object. But this method requires a significant amount of time for training. Though the methods are generally designed utilizing attention-based approaches, but they do not take into account the feature of medical images in which there are indistinguishable plain radiograph with multiple diseases and infection.

2.2. Localization of thoracic diseases

A weakly supervised deep learning system which consists of squeeze-and-excitation blocks, multi-map transfer, and max-min pooling was proposed in [10] for the classification and localization of the diseased region. ChestX-ray14, [11] a large-scale hospital dataset consisting of 108,948 anterior plain radiographs of 32,717 unique patients, was published along with labels for 13 different thoracic pathologies and one normal class. As a result, each radiograph may contain multiple annotations. While ChestXray14 is a massive and high-quality dataset, the annotations are noisy. This poses a challenge in disease classification. Furthermore, [11] it has been demonstrated experimentally that these common thoracic diseases could be accurately diagnosed or even spatially localized using a unified weakly supervised multi-label learning framework trained by generated noisy annotations. The Resnet surpassed the other SOTA CNN which includes Alex Net [12], Google Net [13] and, VGGNet-16 [14] by achieving a class-wise ROC-AUC score of 0.8141 in the case of cardiomegaly. Several observations, such as "Mass" and "Pneumonia" were substantially lowered to 0.5609 and 0.6333, respectively. In real-life conditions, this can be extremely problematic since these are fairly common clinical issues. This study also highlighted a long-held prejudice regarding the inability of conventional CNN to develop meaningful representations under poor annotations. The main challenge in applying deep learning models to medical problems is the lack of high-quality annotated data.

Immediately after the publication of ChestX-ray14, [15] introduced a cutting-edge CNN model called CheXNet, which consisted of 121 layers and DenseNet as the backbone. The model takes plain radiographs as input and returns the probability of disease as well as a heatmap that localizes the most likely disease regions on the input images. However, the network does not have any significance for learning under weak supervision. Although [15] achieved significant results, still there is a space for improvement. A combination of attention-mining and activation maps was proposed in [16]. In this approach, two activation maps were generated. The second activation map was created from the first after removing the active area from the first activation map. The L2 distance was then utilized to analyse the links between multiple diseases

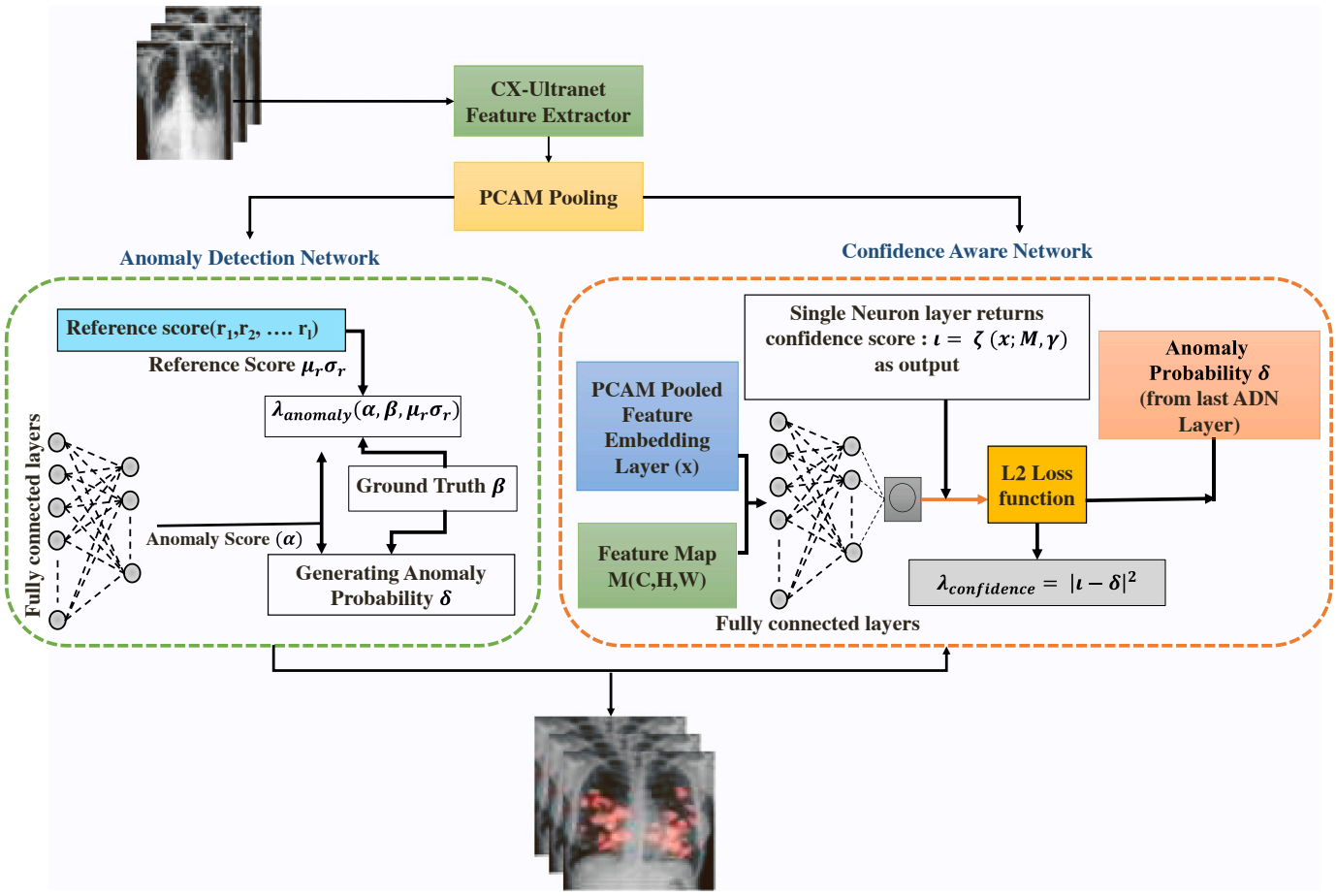

Fig. 1. The overall flow of the CAPCAM architecture. Starting with CX-Ultranet as feature extractor, PCAM pooling the feature maps and embeddings and then passing them through the ADN and CAN visualize the precise localizations.

inside a radiograph. In [17], the authors addressed the challenge of collecting instance-level labelling for localization of the multi-thorax diseases. They developed a model and a conditional loss function which can be trained simultaneously with a smaller number of labelled and non-labelled data. The authors in [26] emphasized the lack of high-quality radiographs in the ChestX-ray14 dataset. They created an alignment module as well as a contrast-induced attention network, which enhanced attention to the problem area by contrasting negative images with equivalent positive images.

In medical imaging endeavours like disease localization and classification, achieving precise and robust results is crucial for practical application. Many existing methods in this field rely on traditional approaches or deep learning algorithms to process images. However, these methods often encounter difficulties when dealing with images affected by various forms of degradation. Traditional image processing techniques and certain deep learning models can be highly affected by noise, which can significantly reduce their performance. This limitation hinders their effectiveness in handling real-world situations where images are often subjected to noise from various sources like sensors, compression, or transmission. Additionally, many existing methods struggle to maintain accuracy when dealing with varying signal-to-noise ratio (SNR) levels. As the SNR decreases and noise becomes more prominent, these methods struggle to extract relevant features and patterns, resulting in inaccuracies in tasks like anomaly localization or object detection.

Furthermore, some methods that are trained on clean images may not perform well in noisy environments. Deploying such methods in practical settings with noisy image data can lead to decreased performance due to their inability to effectively handle noise-related challenges. While deep learning methods have shown promise in image processing, they often require substantial computational resources and large datasets for training. This can be impractical and resource-intensive, particularly for real-time applications or devices with limited processing capabilities.

The NIH and the CheXpert chest x-ray image datasets have continued to remain significant drivers for innovative approaches in the domain of computer vision. One of the innovative methods involved end-to-end architectures that incorporate attention mechanisms to optimize the selection of the thoracic region of interest [30]. To accomplish the multiclass classification problem, an attention-driven, and spatially unsupervised Stern network was utilized. Another innovation has been integrating connections between discriminative image features and corresponding disease labels using approaches like dual-path decoders. This approach involved the concatenation of spatial reduction attention, dual path attention, and a feature enhancement module. In another case, multi-modal learning was utilized to incorporate inter-relationships between different pathological conditions, consisting of a representational, a multi-modal bridge, and a graph learning module [31]. Further, global and local attention supervision to boost lesion detection has also been explored [32]. In this model, they utilized a Dual attention optimization (DualAttNet) to refine feature representations which performed better on the ChestX-ray8, among others, than earlier methods,

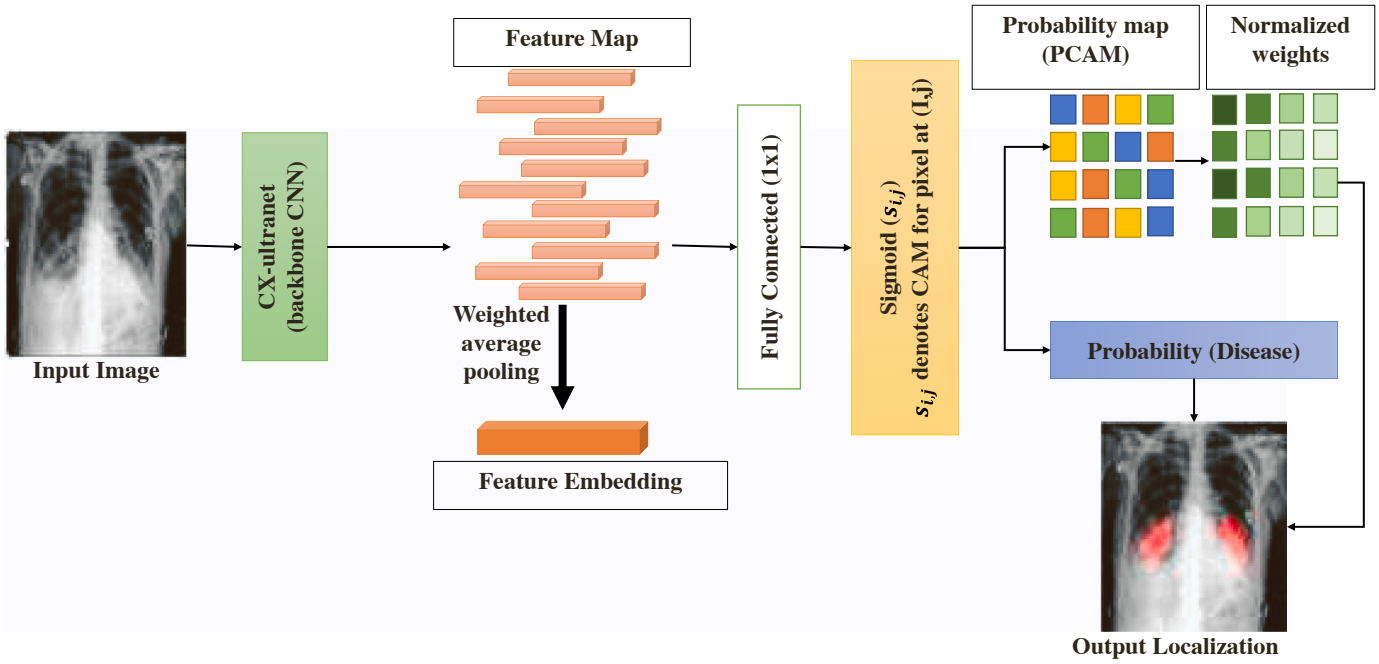

Fig. 2. Probabilistic CAM generation from the input image.

which relied only on single-scale attention. In contradistinction to the said approaches, our proposed approach utilizes probabilistic activation mapping and anomaly detection techniques for weakly supervised localization without additional spatial annotation needs. The interpretable model architecture also provides clinical insights.

To address these shortcomings and enhance image processing performance in noisy environments, this study proposes a novel approach. The proposed method involves deliberately introducing Gaussian noise to images and varying the SNR levels to mimic real-world conditions. This systematic evaluation aims to gauge the model's ability to handle noise interference. Specifically, the proposed method focuses on fine-tuning and optimizing existing deep learning algorithms to process images affected by different levels of Gaussian noise. By understanding how the model behaves under various SNR conditions, potential improvements can be identified to enhance its performance and ability to generalize to noisy image data. Additionally, the proposed method seeks to strike a balance between accuracy and computational efficiency, making it more suitable for real-time applications and devices with limited resources. This approach is motivated by the need to overcome the limitations of current methods, including sensitivity to noise, lack of robustness, limited generalization, and resource-intensive requirements. By addressing these issues, the proposed method aims to advance the field of medical imaging, enabling more reliable and effective performance in challenging real-world scenarios. The classification and localization of medical imaging is a very active area of research in the field of computer vision. However, existing models are plagued by the dual issues of class imbalance and the lack of explainability [50–52]. In this work, we proposed a pooling-based CAM model for the localization of multiple diseases. For explainability [34] [35], we used saliency maps and deconvolution.

3. Methodology

3.1. CX-Ultranet as a highly efficient feature extractor

The CNN is used in combination with the *Confidence-Aware Probabilistic Class Activation Map* (CAPCAM) model. We used CX-Ultranet [25], which was pre-trained on the NIH Chest X-ray dataset, and the CheXpert dataset. Initially, the image is processed using layer-by-layer mobile inverted CNN. For the separation of the optimal feature maps and training the CNN, the multiclass cross-entropy loss function is used. The retrieved feature maps and then fed into the innovative CAPCAM pooling network. In order to prevent data leakage, duplicate is removed. Image data generator class from Keras Framework was used for data augmentation. We perform sample wise centre, standard normalization, along with shear range = 0.1, zoom range = 0.15, rotation range = 5, width shift range = 0.1, and height shift range of 0.05. The horizontal flipping method was used for centreline extraction and disease related to symmetrical positions. This image data generator class also converts single-channel X-ray images (Grayscale) to three-channel format by repeating the values in the image across all channels. The image size is kept at 320×320 to lower computational complexities. The mean and standard deviation of the dataset is normalised, and the input image is shuffled after each epoch. Another advantage of using a section of the CX-Ultranet [36] as a feature extractor is its high precision and efficiency. We applied the same finding to support our proposed model because CX-Ultranet [25] predicts various diseases with high accuracy. The network can learn different anomalies on its own, but its ability of classification is substantially lower than that of CX-Ultranet. For more precise anomaly localization, it has been used as a post-processing tool.

3.2. Probabilistic class activation map (PCAM) pooling method

The proposed confidence-aware probabilistic CAM pooling [33] for anomaly localization consists of a feature extractor coupled with PCAM

pooling, anomaly prediction network and confidence prediction network is shown in Fig. 1.

Attention pooling is based on the multiple instance learning (MIL) paradigm, which considers embedding as an instance. The plain chest radiograph as a bag of instances is positive, e.g., one of the 13 different thoracic diseases, if a single instance is positive. The PCAM pooling uses the same MIL framework but produces normalized attention weights for each feature map embedding. Fig. 2 depicts the PCAM pooling mechanism in detail. When the CNN is fully trained for multiclass classification, the feature map is represented by M with shape (C, H, W), which is generated from the last fully connected layer of the CNN (CX-Ultranet). The C represents the channel dimension, while H and W represents the height and width of the feature map. We then take the disease with the highest probability from CX-Ultranet and use it to target anomaly localization. All the channels among constant embedding share constant weight. The CAM of a particular thoracic disease is given by

$$f_{i,j} \quad = w^t M_{i,j} + b | i, \quad j\epsilon \quad H, W\# \tag{1}$$

where, $M_{i,j}$ is feature embedding of the feature map M with shape *(C,H, W)* at point *(i,j)*. *w, b* are the weights and bias of the CX-Ultranet from its last fully connected layer. $f_{i,j}$ is the logit function before the sigmoid operation under multiclass classification setting which monotonically calculates the disease likelihood over $M_{i,j}$. Later it is used to guide the PCAM pooling method to generate the heatmap and then refine the target anomaly location. To measure the contribution of each embedding the normalized attention weights are assigned. Now since $f_{i,j}$ is unbounded and exists in (-∞, +∞) so we bound it with a sigmoid function $p_{i,j}$.

$$p_{i,j} \quad = \quad sigmoid(s_{i,j})\# \tag{2}$$

Now we normalize it to get attention weights. Hence, we formulate PCAM as

$$\mathrm{x} = \sum_{i,j}^{H,W} w_{ij} M_{i,j} w_{i,j} = \frac{sigmoid(w^t M_{i,j} + b)}{\sum_{i,j}^{H,W} sigmoid(w^t M_{i,j} + b)} \tag{3}$$

The x is the pooled feature embedding layer and $w_{(i,j)}$ are the extracted attention weights. H and W represents the height and width of the feature map. These are utilized to generate a heatmap for the input images. Subsequently, we demonstrated that the generated heatmaps precisely covers the anomalous region. The threshold is set to 95 % of the threshold map for better visual clarity. Results from other compared state of the art (SOTA) pooling methods are also given same probability thresholding so that the visual comparisons are on equivalent.

3.3. Anomaly detection network

The anomaly detection network (ADN) consists of a multi layered perceptron with 100-neuron hidden layer and one output layer as shown in Fig. 1. It generates an anomaly score for the input image x and is written as

$$\alpha = \varphi(x, M, t)\# \tag{4}$$

Where t is the weakly supervised trainable parameter, M is the feature extracted by the backbone CNN and t is the trainable parameter. According to [17] Gaussian distributions match well with such scores in plenty of datasets. We define a univariate gaussian distribution as $r_1, r_2, r_3 \ldots r_l \sim \mathcal{N}(\mu, \sigma^2)$. The corresponding reference score is given by

$$\mu_R = \quad \frac{1}{l}\sum_{i=1}^{l} r_i\# \tag{5}$$

$$\sigma_R^2 = \quad \frac{1}{l}\sum_{i=1}^{l} (r_i - \quad \mu_R)^2\# \tag{6}$$

We set the values of σ=1, μ=0 and l=5000. For optimizing the anomaly detection module, we have used the following contrastive loss [18–20]

$$\lambda_{anomaly}(\alpha, \beta, \mu_r \sigma_r) = (1-\beta)\left|\frac{\alpha - \mu_R}{\sigma_R}\right| + \beta \max\left(0, margin - \quad \frac{\alpha - \mu_R}{\sigma_R}\right) \tag{7}$$

where σ_R is standard deviation of anomaly scores for the used normalized data generated by the univariate gaussian distribution. The β is the ground truth label where '0' indicates the neutral case and '1' indicates positive case belonging to any of the 13 different classes, and the *margin* represents the Z-score confidence. The empirical value was set to 5 in this work. The anomaly detection network, when trained independently, functions as a binary classifier to differentiate between positive and neutral classes, thus constituting a standalone binary classification model. However, in our work, it is utilized as a post-processing technique after acquiring disease probabilities from CX-Ultranet. Furthermore, CX-Ultranet is a multiclass classifying deep convolutional neural network that addresses any class imbalance through a novel approach. This renders the anomaly detection network intrinsically robust to the presence of class imbalance in the dataset.

3.4. Confidence aware network (CAN)

We considered a total of 13 major thoracic conditions and aimed to localize the region of interest for each one of them. It is not uncommon to have more than a single finding in a given chest x-ray. The CX-Ultranet can simultaneously detect various thoracic diseases and give the percentage probabilities of all the thoracic diseases. However, if we only compare the anomaly scores for localization then the conclusion is insignificant. Recent work depends on either probability thresholding or attention weights for the localization of multiple diseases. Therefore, we used the attention weights in the PCAM map to adhere to the MIL framework. Our confidence-aware network is based on the shared feature extractor. For a strong confidence prediction, it consists of the confidence aware module [16] with four 100-neuron hidden layers, as shown in Fig. 2. The computation of the confidence-aware network is formulated as:

$$\iota = \zeta(x; \theta, \gamma)\# \tag{8}$$

where ι represents the confidence score generated corresponding to the localized anomaly, γ represents the ensembled parameters of the confidence aware module. The ι lies between [0,1] therefore the standard l2 loss is used to optimize the confidence prediction module which is formulated as a regression task.

$$\lambda_{confidence} = \quad |\iota - \delta|^2\# \tag{9}$$

where δ is the anomaly probability which is defined as

$$\delta = \quad \begin{cases} prob, \& if\ \beta = 0 \\ 1 - prob, \& if\ \beta = 1 \end{cases}\# \tag{10}$$

The complete algorithm is written as follows.
Algorithm for the Proposed Model

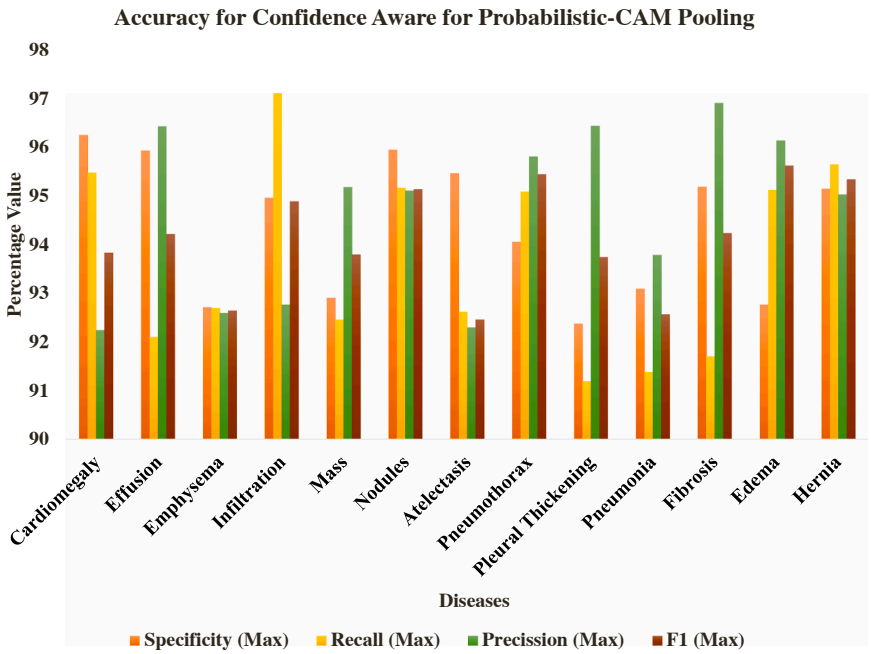

Fig. 3. The performance metrics of CAPCAM. The Specificity, Recall, Precision and F1 scores achieved by confidence aware probabilistic CAM for the thirteen thoracic diseases.

Algorithm for the Proposed Model

Inputs: *Input features* x_i, *labels* v

Outputs: - $\lambda_{anomaly}$, ι , $\lambda_{confidence}$

Intermediate Steps

Step 1 CX-Ultranet –

- *Predicted Result:* $g(u)$ *is the predicted result of the model*
- *Loss function* = $CI_{ce}^{w}(u) = -\big(w_p v\, log\big(g(u)\big) + w_n(1-v)\, log(1-g(u))\big)$
- *Extracted Feature Map* - $M(C,H,W)$
- *Extracted weights and bias* - (w,b)
- *Generate* β *(class with highest percentage) from Multiclass percentage (* $g(u)$*)*

Step 2 CAM (Class Activation Mapping) -

- *CAM Calculation :* $(f_{i,j}) = f_{i,j} = w^t M_{i,j} + b | i,\ j \epsilon H,W$
- *bound* $p_{i,j}$ - $p_{i,j} = sigmoid(f_{i,j})$
- *Pooled features* - $x = \sum_{i,j}^{H,W} w_{i,j} M_{i,j}$, $w_{i,j} = \frac{sigmoid(w^t M_{i,j}+b)}{\sum_{i,j}^{H,W} sigmoid(w^t M_{i,j}+b)}$
- *Extracted attention weights* - $w_{i,j}$

Step 3 ADN (Anomaly Detection Network) -

- *Calculate anomaly score* - $\alpha = \varphi(x,M,t)$
- *Generate univariate gaussian dist.* - $r_1, r_2, r_3 \ldots r_l \sim \mathcal{N}(\mu,\sigma^2)$
- *Generate reference score* - $\mu_R = \frac{1}{l}\sum_{i=1}^{l} r_i$

$$\sigma_R^2 = \frac{1}{l}\sum_{i=1}^{l}(r_i - \mu_R)^2$$

- *Set values of* - $\sigma = 1, \mu = 0$ *and* $l = 5000$
- *Optimize ADN after obtaining :*

  - $\lambda_{anomaly}(\alpha,\beta,\mu_r\sigma_r) = (1-\beta)\left|\frac{\alpha-\mu_R}{\sigma_R}\right| + \beta\, max\left(0, margin - \frac{\alpha-\mu_R}{\sigma_R}\right)$
- *Anomaly score and reference Score with Contrastive Loss Generate Anomaly Probability* (δ) -

$$\delta = \begin{cases} prob & if\ \beta = 0 \\ 1 - prob & if\ \beta = 1 \end{cases}$$

Step 4 CAN (Confidence Analysis Network) -

- *Generate confidence score* - $\iota = \zeta\,(x;M,\gamma)$
- *optimize with L2 loss* - $\lambda_{confidence} = |\iota - \delta|^2$

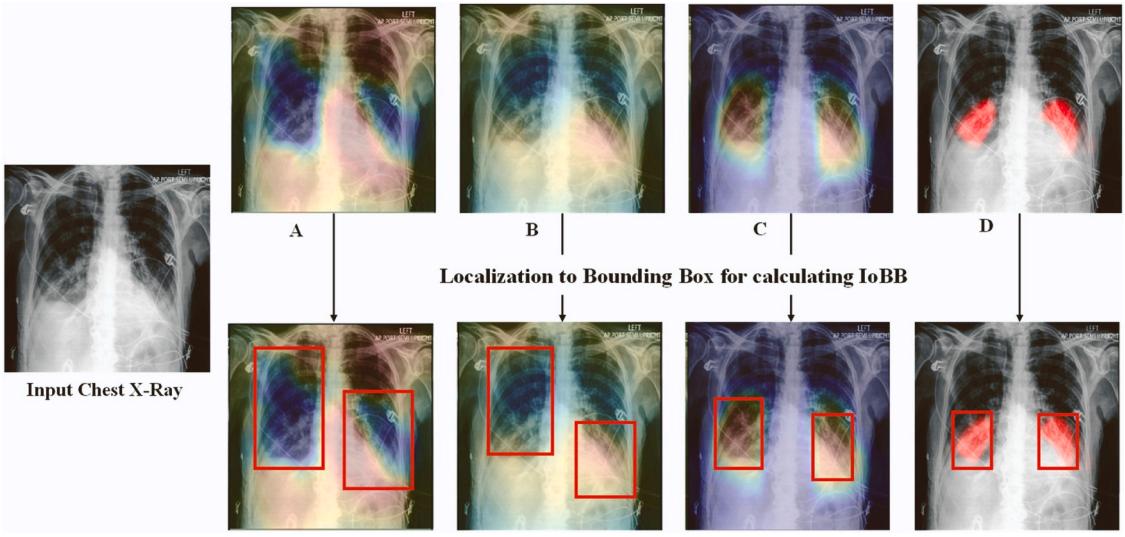

Fig. 4. Diagnosis of Atelectasis by four different approaches in NIH dataset. A) Attention Pooling, B) LSE Pooling, C) CAPCAM pooling, and D) CAPCAM (threshold).

4. Experiment details

4.1. Datasets

The NIH chest X-ray dataset [21] consists of 112,120 X-rays from 30, 805 unique patients with 14 diseases. We strictly follow the official split of NIH, 70 % for training, 10 % for validation, and 20 % for testing to conduct experiments and fair comparison with previous works. We also tested (no training- only used for testing) our model on CheXpert dataset [24] to evaluate the performance of our model. This CheXpert testing was done to ensure generalizability of the method on the other dataset. The images were manually annotated around 200 samples and tested on the same, the number of classes remains same.

4.2. Implementation details

Preprocessing of Plain Chest Radiographs with CX-Ultranet– The CX-Ultranet compresses the 1024×1024 radiographs of the NIH dataset to 512×512, which is an optimal size for detecting anomalous regions and finer feature recognition. We did not perform any image enhancement because it can lead to the detection of superficial anomalies and can lead to incorrect findings. The novel CX-Ultranet extractor addresses the problem of class imbalance. The image resolution is preserved by using feature maps from the last fully connected convolution layer. After the compression a three-channel image is fed to the network.

1) Deconvolutions and Saliency Maps – Saliency maps are plotted for each convolutional layers from the feature extractor to aid better understanding of how the model targets particular areas over time. Regions with high probability are highlighted in red. Both of these

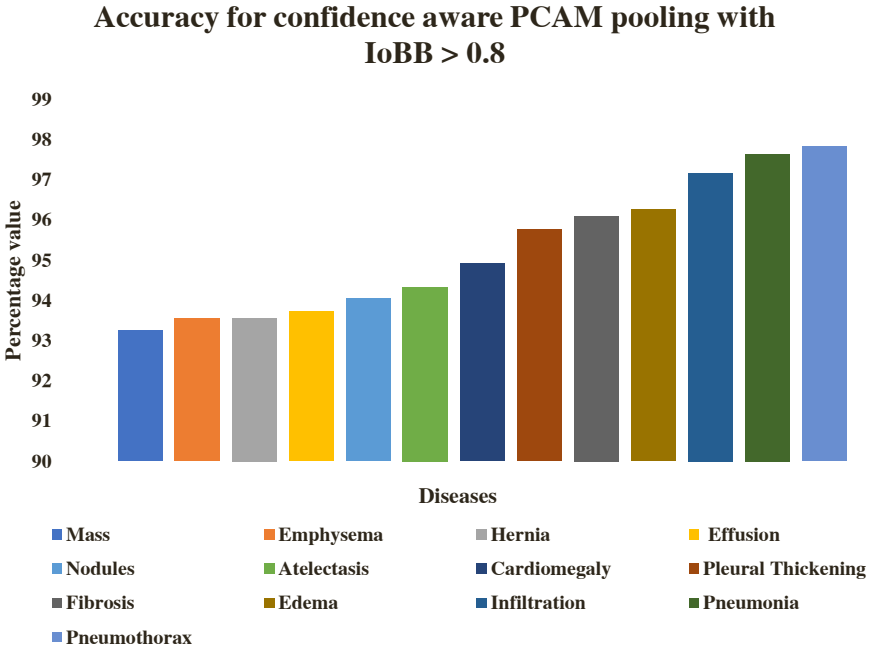

Fig. 5. Accuracy for confidence aware PCAM pooling of NIH dataset with IoBB > 0.8.

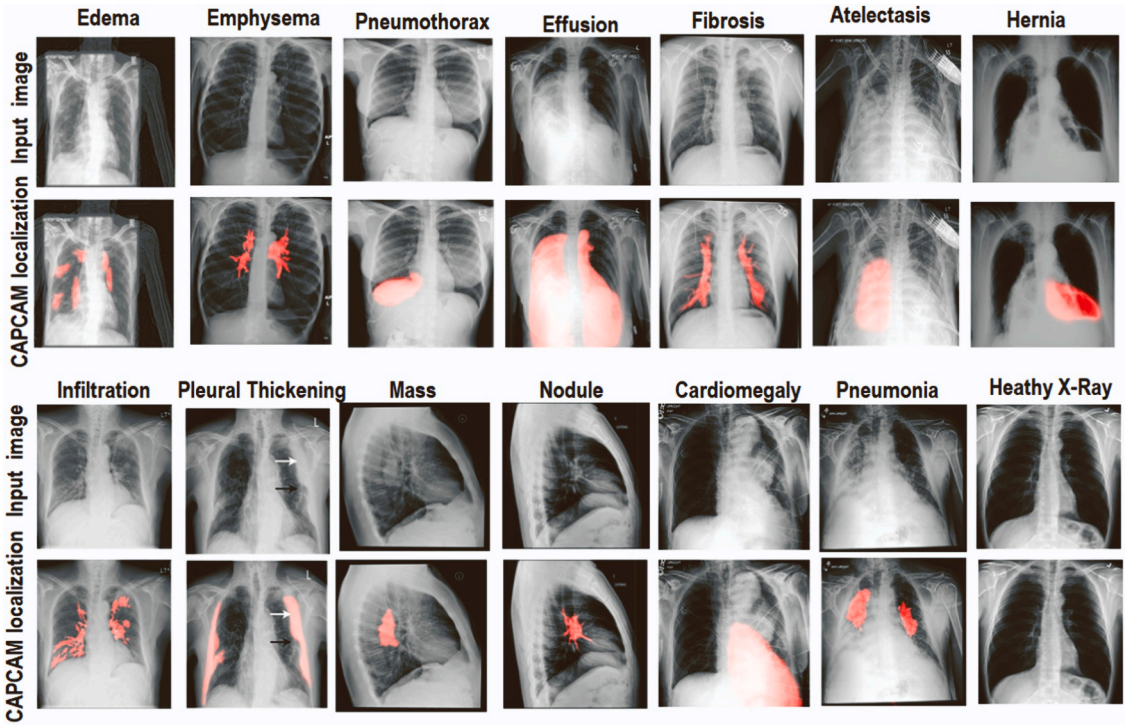

Fig. 6. Anomaly localization using CAPCAM for multiple different diseases on NIH dataset.

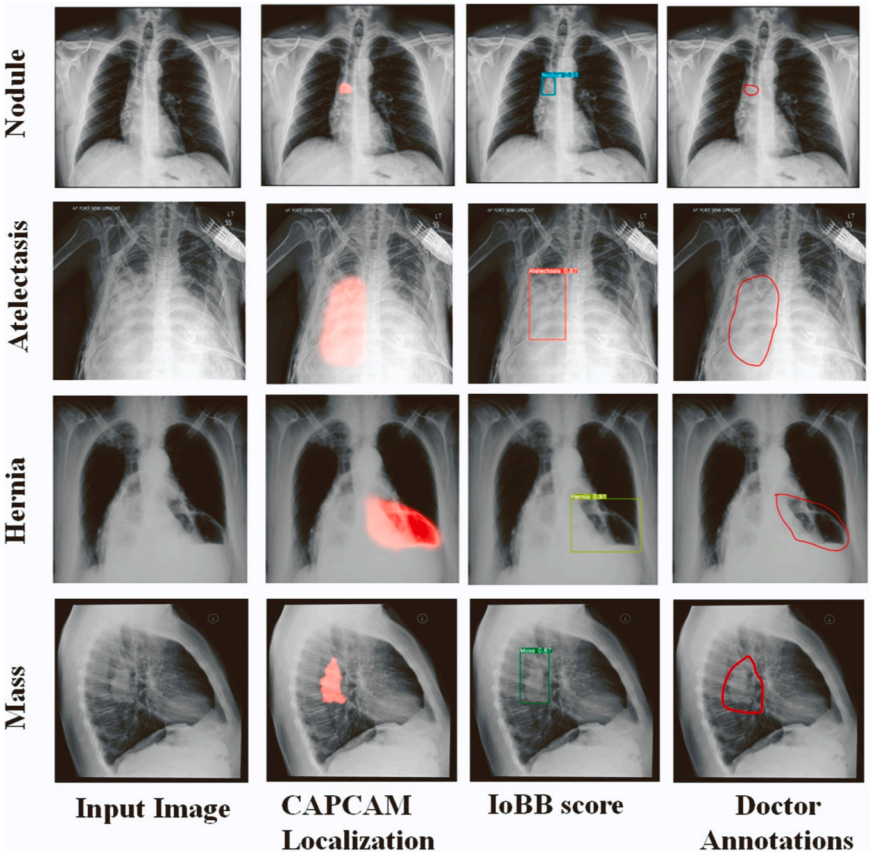

Fig. 7. Comparison of CAPCAM with doctors' annotation in terms of IoBB score and localization in CheXpert dataset.

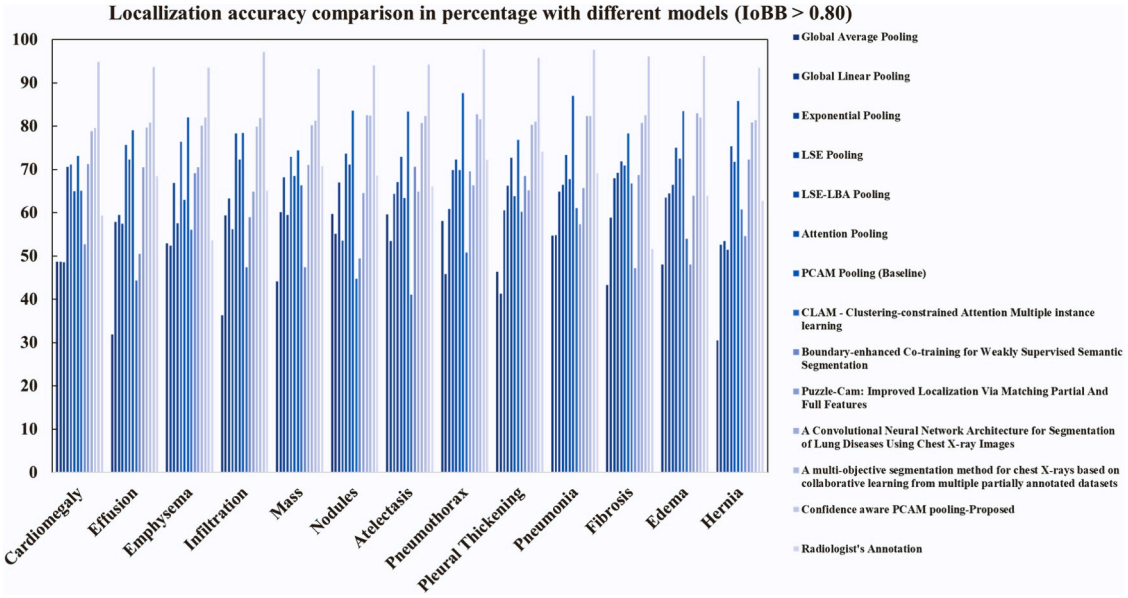

Fig. 8. Accuracy for Different Pooling with IoBB > 0.8. The comparison with other method and expert Radiologist.

steps are performed only for the feature map from the last FC layer of CX-Ultranet but for better understanding we show the outputs from all deconvolutions.

2) Training scheme for guiding different disease localization- The trainable parameter t in ADN is trained explicitly and assisting localisation of multiple diseases. We trained the same using the weights and biases from CX-Ultranet, which eliminates the requirement for a distinct training model. The weights and biases of only the disease with highest likelihood are taken into account.
3) Hyperparameters - The hyperparameters were carefully configured to optimize the performance of the proposed model. The network is trained for 20 epochs with a batch size of 64 to ensure a balance between training efficiency and model convergence. We use the dynamic learning rate which is monitored by Reduce Learning Rate on Plateau to keep an eye on the model performance. To stabilize the overall model, we chose the Adam Optimizer with a learning rate of 1e-3, a widely used and effective optimization algorithm.
4) System and Configuration- For the setup we have High performance Computers with the following configurations - Processor Intel(R)

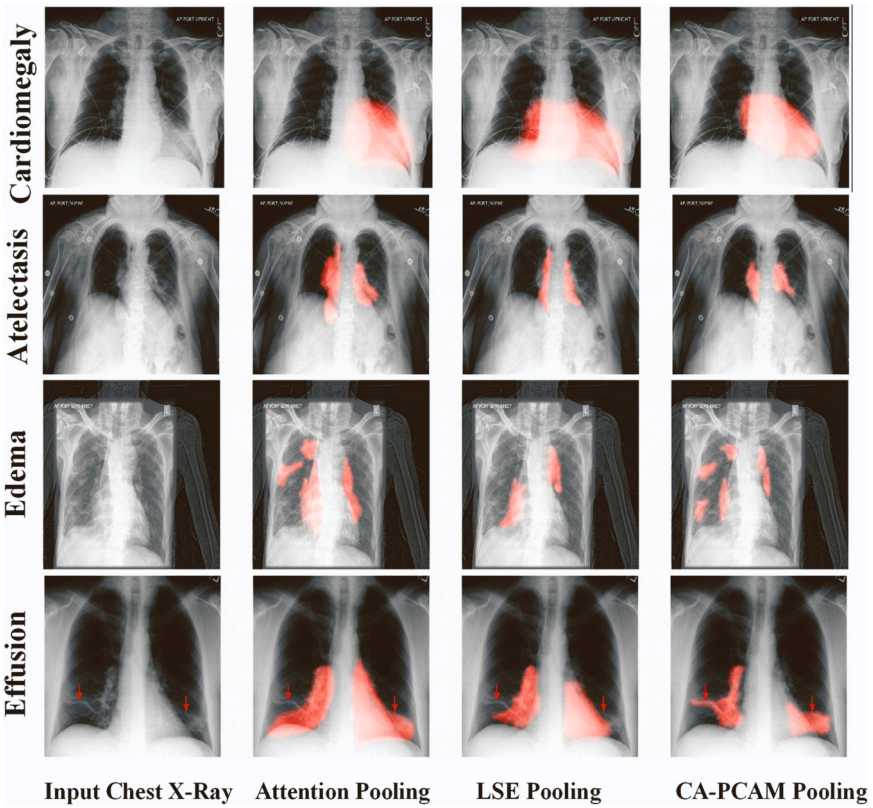

Fig. 9. Visualizations of diseases generated by the three best models including attention pooling, LSE pooling and the proposed CAPCAM pooling.

Table 1
The performance metrics (Precision, Recall, Specificity, and F1 score) for different pooling of all 13 diseases on NIH dataset.

Methods	Metrics	Card.	Effu.	Emph.	Infi.	Mass	Nodu.	Atel.	Pntx.	PT.	Pneu.	Fibr.	Edema	Hernia	Overall Accuracy
Global Average Pooling [20]	**Specificity**	66.6	68.7	52.0	64.9	52.5	61.2	66.2	50.0	55.1	65.0	52.4	57.7	58.3	56.34
	Recall	54.3	61.4	68.1	58.2	60.5	59.7	51.0	67.4	68.3	64.9	60.1	52.3	50.1	58.83
	Precision	67.9	54.1	51.4	62.3	68.3	56.6	67.2	56.8	55.2	59.2	68.3	51.2	65.1	59.23
	F1	60.3	57.5	58.6	60.2	64.2	58.1	58.0	61.6	61.0	61.9	63.9	51.7	56.6	58.32
Global Linear Pooling [20]	**Specificity**	58.3	57.0	66.2	59.6	65.1	60.7	62.3	65.9	60.2	63.8	61.9	64.1	62.7	62.95
	Recall	59.5	62.1	65.1	68.3	55.5	66.1	64.2	59.4	57.2	60.1	67.9	66.7	60.9	62.73
	Precision	60.6	64.4	62.1	61.6	56.9	67.6	63.4	64.7	57.9	61.7	58.6	58.9	57.7	63.29
	F1	60.0	63.2	63.6	64.8	56.2	66.8	63.8	61.9	57.6	60.8	62.9	62.6	59.2	60.37
Exponential Pooling [20]	**Specificity**	71.8	57.5	56.0	68.3	71.0	74.3	74.6	66.9	63.9	67.7	73.2	64.8	57.2	68.82
	Recall	73.7	63.7	62.9	70.5	67.1	56.0	60.2	61.0	69.9	70.7	65.9	69.9	65.6	64.14
	Precision	63.7	63.7	62.4	67.2	65.7	67.5	67.2	70.2	66.8	73.3	73.0	67.5	72.7	69.79
	F1	68.3	63.7	62.7	68.8	66.4	61.2	63.5	65.3	68.3	72.0	69.2	68.7	69.0	64.53
LSE Pooling [21]	**Specificity**	73.0	75.1	59.1	55.8	62.4	56.7	72.5	73.5	66.5	76.5	78.0	75.5	66.7	61.21
	Recall	59.2	70.8	72.2	56.5	71.3	56.6	59.9	74.6	67.8	79.0	66.0	61.4	72.5	65.95
	Precision	63.0	56.5	71.9	76.7	77.6	78.8	66.4	72.3	70.4	72.5	61.8	70.9	55.3	68.68
	F1	61.0	62.8	72.0	65.1	74.3	65.9	63.0	73.4	69.1	75.6	63.8	65.8	62.8	69.42
LSE-LBA Pooling [22]	**Specificity**	62.2	75.8	77.6	64.2	73.1	58.5	55.9	56.4	55.2	63.9	78.6	74.1	66.6	61.09
	Recall	66.3	78.4	56.9	68.7	71.2	67.1	56.1	70.7	59.8	57.7	55.9	65.8	66.3	64.92
	Precision	70.6	65.9	77.5	58.0	56.0	69.8	76.7	65.5	71.2	69.5	76.7	69.6	74.3	66.65
	F1	68.4	71.6	65.6	62.9	62.7	68.4	64.8	68.0	65.0	63.0	64.6	67.6	70.1	67.31
Attention Pooling [23]	**Specificity**	64.3	67.4	64.1	69.9	76.0	72.7	69.0	75.5	75.2	55.6	74.7	59.3	65.9	69.88
	Recall	63.0	75.2	59.0	59.5	67.3	61.6	68.9	63.3	60.2	78.6	75.3	74.1	65.0	67.18
	Precision	77.6	69.4	67.6	67.7	72.0	76.9	68.5	64.9	73.6	65.1	67.7	75.6	73.7	70.75
	F1	69.5	72.2	63.0	63.3	69.6	68.4	68.7	64.1	66.2	71.2	71.3	74.8	69.1	69.5
PCAM Pooling (Baseline)	**Specificity**	80.4	77.2	81.9	81.8	78.1	70.7	71.4	71.0	77.4	66.2	78.3	74.3	81.8	76.27
	Recall	79.3	71.1	75.1	81.7	72.8	77.5	71.3	80.2	75.3	82.5	72.2	80.7	82.3	78.99
	Precision	75.3	72.6	80.1	81.8	78.8	72.8	77.4	75.3	82.7	74.6	81.3	82.5	72.0	74.62
	F1	77.2	71.8	77.5	81.8	75.7	75.1	74.2	77.6	78.9	78.4	76.5	81.6	76.8	77.45
CLAM - Clustering-constrained Attention Multiple instance learning [27]	**Specificity**	83.7	83.2	74.4	86.6	83.2	68.2	75	81	80.7	87	77.6	65.2	82.5	90.11
	Recall	68.6	85.2	72.1	82.1	81.2	72.5	85.1	70.9	68	82.4	81	85.9	75.5	89.85
	Precision	86.1	65.9	86.8	70.8	66.2	78.1	76.2	68.9	82.9	72	74.1	85.7	83.1	90.53
	F1	76.4	74.3	78.8	76	72.9	75.2	80.4	69.9	74.7	76.8	77.4	85.8	79.1	90.26
Boundary-enhanced Co-training for Weakly Supervised Semantic Segmentation [28]	**Specificity**	81.2	81	86.8	83.7	68.9	85.1	65	80.9	69.8	71	65.1	69.2	65.4	56.34
	Recall	78.3	78	80.4	67.6	85.8	81.6	70.2	66.2	68.2	82.5	65.3	77.1	81.3	58.83
	Precision	87.1	77.2	66.2	66.7	78.1	80.3	66.5	67.4	69.5	82	75.8	69.2	86.7	59.23
	F1	82.5	77.6	72.6	67.1	81.7	80.9	68.3	66.8	68.8	82.2	70.2	72.9	83.9	58.32
Puzzle-Cam: Improved Localization Via Matching Partial and Full Features [29]	**Specificity**	77.9	80.6	68	74.1	85.3	78.3	77.6	72.6	71.8	68.6	82	78.4	74.6	62.95
	Recall	86.7	81	76.7	65.5	75.1	71.8	66.2	80.3	76.4	85.1	81.2	78.7	75.5	62.73
	Precision	87.2	86.1	74.4	75.4	83.7	66.2	75.6	65.3	77.3	86.2	84.4	80.5	66.5	63.29
	F1	86.9	83.5	75.6	70.1	79.1	68.9	70.6	72	76.8	85.6	82.8	79.6	70.7	60.37
A Convolutional Neural Network Architecture for Segmentation of Lung Diseases Using Chest X-ray Images [37]	**Specificity**	77.23	80.35	83.46	78.77	81.90	79.32	82.46	80.12	83.79	79.01	82.11	79.79	83.11	78.46
	Recall	84.57	78.12	77.65	83.11	77.41	84.79	77.97	84.32	77.23	85.12	77.65	84.90	77.46	81.43
	Precision	80.12	81.46	82.79	81.05	83.21	80.79	83.54	81.23	84.01	80.43	83.90	81.72	84.32	80.99
	F1	78.90	79.73	80.11	79.87	80.28	82.54	80.71	82.77	80.39	82.38	80.79	83.01	80.87	81.18

(*continued on next page*)

Core(TM) i9–10900 K CPU @ 3.70 GHz 3.70 GHz, Installed RAM 128 GB (128 GB usable) NVIDIA RTX A4000, Dedicated GPU memory 0.0/16.0 GB, Shared GPU memory 0.0/63.9 GB, GPU Memory 0.0/79.9 GB. We have used PyCharm and Data Spell (version – 2022) for the python development process.

5. Result and discussion

5.1. Results

5.1.1. Evaluation metrics

The proposed CAPCAM model is evaluated on multiple evaluation metrics such as specificity, recall, precision, and F1 scores. The results are illustrated in Fig. 3. As our model uses both attention weights and confidence scores to assist anomaly localization, therefore it shows a

Table 1 (*continued*)

Methods	Metrics	Card.	Effu.	Emph.	Infi.	Mass	Nodu.	Atel.	Pntx.	PT.	Pneu.	Fibr.	Edema	Hernia	Overall Accuracy
A multi-objective segmentation method for chest X-rays based on collaborative learning from multiple partially annotated datasets [38]	Specificity	84.79	82.35	79.01	81.23	83.46	80.79	79.32	82.11	84.57	81.90	78.46	80.35	83.79	81.05
	Recall	78.12	80.35	84.57	81.72	78.90	83.11	84.32	80.12	77.65	81.46	85.12	82.79	77.97	80.90
	Precision	81.05	81.46	79.73	82.11	82.79	81.90	80.43	83.21	83.79	83.46	80.12	82.35	84.01	80.72
	F1	79.57	80.89	82.01	81.92	81.28	82.50	82.31	81.61	81.11	82.40	82.54	82.06	81.35	81.79
CAPCAM Pooling (Proposed method)	Specificity	**90.9**	**91.0**	**90.3**	**90.6**	**90.1**	**91.5**	**91.6**	**90.0**	**90.5**	**90.0**	**91.5**	**91.7**	**91.2**	68.82
	Recall	**88.6**	**88.2**	**88.5**	**88.5**	**89.6**	**89.7**	**89.8**	**88.5**	**88.7**	**89.0**	**87.8**	**89.7**	**88.4**	64.14
	Precision	**91.3**	**91.9**	**92.0**	**90.4**	**91.9**	**90.2**	**90.1**	**90.8**	**91.4**	**91.9**	**91.1**	**91.3**	**91.9**	69.79
	F1	**89.9**	**90.0**	**90.2**	**89.4**	**90.7**	**89.9**	**89.9**	**89.6**	**90.0**	**90.4**	**89.4**	**90.5**	**90.1**	64.53

*Cardiomegaly, Effusion, Emphysema, Infiltration, Mass, Nodules, Atelectasis, Pneumothorax, Pleural Thickening, Pneumonia, Fibrosis, Edema, Hernia

very high degree of precision and F1 scores. We also performed an ablation study on different pooling methods to evaluate the performance of our model as compared to others. Max F1 scores is the harmonic mean of recall and precision which exceed 92 % for all thirteen thoracic diseases. The CAPCAM network has minimal class imbalance because the CX-Ultranet feature extractor handles it very efficiently. A further reason for such high results is because we incorporate the MIL framework while generating the PCAM.

5.1.2. Intersection over bounding box (IoBB) scores

In order to compare the obtained results with the ground truth, the IoBB accuracy calculated. The CAPCAM model has the ability to localize the anomalous region which is discussed in the next section. Fig. 4. shows the localization of atelectasis by attention pooling, LSE pooling and the proposed model. The CAPCAM pooling with confidence threshold set to 0.9 shows the best result than any other methods. The benefit of confidence prediction network is also clearly visible as more refined localization of the anomalous region as seen if Fig. 4 (C) to Fig. 4 (D). We next compared the bounding boxes to the ground truth annotations in the dataset, and the results are shown in Fig. 5. For this experiment, we set the threshold IoBB score to 0.8, which implies that if IoBB with the ground truth scores more than 0.8, we will consider it a success; otherwise, we will consider it a failure. We purposefully kept the confidence criterion higher in order to achieve a better IoBB score.

5.1.3. Visual observation

The visual outputs of the proposed model in combination with the feature extractor for the localization of anomalous region is shown in Fig. 6. We have evaluated the confidence score for the anomaly localized separately and then evaluated the localized regions with a confidence score greater than 0.9 as final anomalous regions. The threshold value is determined empirically in order to maximize IoBB scores. The ablation study of different thresholds is also shown in the ablation study section. As shown in Fig. 6, the network precisely targets various lesions and locations for various diseases.

We used the CheXpert dataset to evaluate the effectiveness of CAPCAM. As the dataset was not labelled, we manually annotated it with the doctor's assistance. The doctor's annotations are used as the ground truth and then tested on our model. Fig. 7 depicts various diseases in a single image and provides a confidence score based on the CheXpert dataset. This proves that our proposed CAPCAM model can also perform well on the unseen radiographs and can be used as a generalized model for thoracic diseases localization.

5.2. Discussion and comparison

5.2.1. Comparison with other models

A comparative study of accuracy with different pooling methods is performed for the evaluation of the proposed model. Fig. 8 compares our CAPCAM model to Global Average Pooling, Global Linear Pooling [20], Exponential Pooling [20], LSE pooling [21], LSE-LBA pooling [22], Attention Pooling [23], CLAM - Clustering-constrained Attention Multiple instance learning [27], Boundary-enhanced Co-training for Weakly Supervised Semantic Segmentation [28] and, Puzzle-Cam: Improved Localization Via Matching Partial and Full Features [29] and expert Radiologist as well. The CAPCAM outperforms the baseline PCAM, with a significant improvement in anomaly localization. For all methods, the same feature extractor was implemented. Other prerequisites like attention pooling and LSE-LBA pooling were recreated. We have kept the same feature extractor for all the different methods. All approaches were evaluated on the same dataset and compared using the IoBB score. The IoBB threshold for all models is fixed at 0.8, with a greater IoBB indicates a successful case and a lower IoBB indicates a failed test case. Expert Radiologist does not perform well only from images (we have considered any medical history as these are not available in the data set for radiologist assessment). The CAPCAM network surpasses all other models by more than 20 %. Finding the appropriate anomalous region with a high precision depends significantly on the generation of CAMs based on attention weights and their interpretation in a probabilistic model. This is then directed by an independent confidence prediction network for the resulting activation map. The thresholding of confidence enhances the performance of CAPCAM by 10 %-15 % compared to baseline PCAM. Fig. 9 illustrates the performance of different pooling methods. The details performance of different comparable computer aided method on both the datasets with our proposed method are illustrated in Table 1 and Table 2.

5.2.1.1. Assessment notes by the radiologist/specialist. Radiologists begin by assessing the lungs and airways for opacities, nodules, masses, infiltrates, congestion, and atelectasis that indicate potential diseases e.g., pneumonia, lung cancer, pulmonary edema among others. Next, they evaluate the heart and mediastinum for signs of cardiomegaly, widening of the mediastinum, and calcifications that could point to conditions like congestive heart failure, aortic aneurysms and dissection, infectious processes and the like. They also assess the pleura for thickening and effusions indicating pleural diseases. Radiologists then move on to bony structures, looking for bone destruction from tumors or infections as well as fractures or congenital bone anomalies. All this is then synthesized into a coherent whole to arrive at the diagnosis.

However, visual analysis of chest X-rays has limitations due to the

Table 2
The performance metrics (Precision, Recall, Specificity, and F1 score) for different pooling of all 13 diseases on Chexpert dataset.

Methods	Metrics	Card.	Effu.	Emph.	Infi.	Mass	Nodu.	Atel.	Pntx.	PT.	Pneu.	Fibr.	Edema	Hernia	Overall Accuracy
Global Average Pooling [20]	**Specificity**	59.9	52.1	57.0	54.8	50.3	58.2	53.5	51.1	56.1	50.1	55.4	52.0	57.3	57.59
	Recall	51.2	58.5	51.0	56.1	59.0	52.8	57.7	58.9	51.7	59.3	50.4	58.1	50.8	59.86
	Precision	57.3	54.3	58.1	55.8	53.9	56.4	55.0	53.1	57.9	53.0	58.8	54.1	58.5	58.77
	F1	54.2	55.3	54.0	55.5	55.9	54.5	55.8	55.5	54.7	55.2	54.0	55.6	54.4	56.33
Global Linear Pooling [20]	**Specificity**	64.1	57.1	61.0	54.3	66.4	53.0	59.3	55.5	61.7	53.1	65.8	52.4	60.1	64.70
	Recall	53.2	62.4	52.8	65.1	53.9	64.8	54.1	63.5	52.3	66.1	54.8	65.4	53.5	64.82
	Precision	60.8	58.2	63.5	59.0	61.2	58.8	63.0	59.7	64.1	58.5	60.3	59.3	63.8	62.83
	F1	56.8	60.2	57.1	61.4	57.3	61.3	58.0	61.5	57.7	61.3	57.5	61.3	58.1	63.48
Exponential Pooling [20]	**Specificity**	65.4	59.9	63.1	61.6	69.0	58.8	62.3	57.3	61.0	68.1	69.3	57.7	62.8	66.28
	Recall	61.2	67.9	58.5	68.3	59.3	64.1	60.1	66.8	59.0	61.7	58.2	69.0	60.3	67.82
	Precision	64.1	61.0	66.4	62.8	63.5	61.1	65.8	60.4	67.2	64.6	66.1	61.6	65.1	64.81
	F1	62.7	64.3	61.9	65.3	61.3	62.4	63.1	63.5	63.1	63.1	62.1	65.1	62.7	65.70
LSE Pooling [21]	**Specificity**	67.4	63.8	65.1	61.0	72.3	61.2	64.6	60.4	66.8	70.1	62.3	59.7	63.1	71.30
	Recall	62.1	69.1	61.4	70.4	60.3	68.8	62.8	71.2	61.1	63.5	69.4	72.1	61.7	69.70
	Precision	68.3	64.6	67.2	65.8	69.9	66.1	69.0	66.4	68.1	67.9	64.1	67.2	68.5	68.89
	F1	65.1	66.8	64.3	68.0	64.9	67.4	65.9	68.7	64.8	65.6	66.7	69.7	65.0	71.88
LSE-LBA Pooling [22]	**Specificity**	74.2	73.7	68.2	69.9	76.7	74.6	67.9	73.8	73.5	72.1	75.6	66.4	67.0	75.67
	Recall	66.0	72.1	69.9	75.2	69.3	74.1	67.1	68.7	72.9	68.1	68.9	71.1	72.6	72.85
	Precision	67.9	72.9	73.0	76.2	76.2	77.0	74.8	68.7	74.7	71.9	74.0	70.1	66.3	78.71
	F1	69.5	70.9	68.5	74.9	67.8	74.5	71.0	75.6	67.0	71.4	68.2	69.9	69.5	75.26
Attention Pooling [23]	**Specificity**	64.2	62.8	68.5	60.3	75.1	61.6	67.9	59.0	70.1	63.5	74.3	57.7	69.0	66.44
	Recall	69.9	72.3	63.1	73.8	61.7	71.4	62.3	74.8	61.0	70.8	61.4	75.4	62.1	71.35
	Precision	65.4	64.1	67.2	63.0	70.4	63.8	68.1	61.2	69.3	66.4	68.8	60.1	67.5	46.24
	F1	67.5	68.0	65.2	67.0	65.9	67.3	65.1	66.8	64.8	68.1	65.0	66.2	64.9	68.91
PCAM Pooling (Baseline)	**Specificity**	78.5	75.8	71.2	71.0	80.3	77.7	72.1	71.4	79.0	73.8	80.8	70.1	76.1	79.43
	Recall	70.1	73.5	79.0	79.3	70.8	72.3	78.8	80.1	71.7	77.2	70.4	81.0	73.0	78.16
	Precision	74.3	77.2	75.4	76.1	76.5	78.1	75.0	76.8	77.5	75.1	77.9	76.3	78.5	81.60
	F1	72.2	75.3	77.1	77.7	73.4	75.2	76.8	78.4	74.6	76.1	74.0	78.4	75.7	78.50
CLAM - Clustering-constrained Attention Multiple instance learning [27]	**Specificity**	71.5	73.5	76.2	77.9	74.8	77.4	72.8	71.6	75.4	77.2	78.7	77.3	71.4	76.45
	Recall	79.1	72.0	77.9	76.6	78.1	76.2	79.6	79.3	75.9	74.1	77.7	78.3	72.3	83.69
	Precision	72.7	72.8	73.7	73.3	78.2	76.4	74.2	79.5	76.9	77.1	76.9	79.4	77.2	79.55
	F1	71.3	72.5	72.7	73.8	74.4	77.0	75.2	78.4	77.9	77.1	79.3	78.8	78.0	80.72
Boundary-enhanced Co-training for Weakly Supervised Semantic Segmentation [28]	**Specificity**	80.1	77.2	69.9	78.5	68.1	75.8	70.4	79.0	68.8	76.1	79.8	69.1	77.5	76.58
	Recall	68.8	71.4	79.3	69.0	81.8	72.1	80.3	68.5	81.1	71.7	69.4	80.8	70.1	79.80
	Precision	73.5	75.1	73.0	74.3	71.2	77.7	73.8	74.8	71.0	78.1	75.4	71.4	77.9	76.71
	F1	71.0	73.2	76.0	71.6	76.4	74.8	76.9	71.5	75.1	74.7	72.3	75.7	74.1	77.38
Puzzle-Cam: Improved Localization Via Matching Partial and Full Features [29]	**Specificity**	64.3	65.4	70.8	66.1	61.6	75.8	69.3	67.6	72.7	68.3	67.5	64.8	71.2	74.56
	Recall	64.8	70.7	72.4	72.8	61.1	70.8	68.7	73.2	63.7	66.4	76.5	68.6	73.9	73.70
	Precision	65.5	65.6	61.7	71.0	63.2	71.5	72.8	67.9	66.4	69.5	62.5	67.2	76.5	68.77
	F1	74.5	73.1	61.6	74.0	65.6	74.9	71.0	73.1	64.7	69.2	71.4	69.4	69.6	76.03
A Convolutional Neural Network Architecture for Segmentation of Lung Diseases Using Chest X-ray Images [37]	**Specificity**	87.45	82.79	89.73	81.05	90.12	84.32	82.15	89.41	83.57	81.97	90.35	84.79	82.31	82.41
	Recall	81.23	87.89	82.11	88.65	81.72	87.41	88.12	81.35	88.90	87.65	82.46	88.23	89.57	86.72
	Precision	89.01	88.32	90.43	89.12	89.46	87.85	88.77	90.79	87.23	89.32	89.01	86.90	88.64	84.90
	F1	84.94	86.07	85.97	84.79	85.31	86.04	86.41	85.52	85.82	85.14	85.71	87.57	86.48	85.80

(*continued on next page*)

unavailability of relevant historical data. Radiologists can miss subtle abnormalities and misinterpret overlying structures. Experience and fatigue can impact accuracy. Examples of common errors include misdiagnosing stable nodules as new masses, failing to detect pneumothorax and subtle fractures, and misinterpreting anatomical variants as diseases.

We also evaluated the three best models visually. Fig. 9 shows the visual comparison of diseases with attention pooling, LSE pooling and the proposed CAPCAM pooling. The attention pooling typically includes more area than the actual abnormal areas, whereas LSE pooling does not. LSE pooling, on the other hand, fails in cases where several anomalous regions must be identified. In Fig. 9, we can clearly see that in the case of edema, the LSE pooling failed to localize the anomaly in the upper segment of the lungs. Our model is able to detect finer features with ease because of the attention-based CAMs. Thus, our model outperforms existing SOTA localization networks and techniques.

Table 2 (*continued*)

Methods	Metrics	Card.	Effu.	Emph.	Infi.	Mass	Nodu.	Atel.	Pntx.	PT.	Pneu.	Fibr.	Edema	Hernia	Overall Accuracy
A multi-objective segmentation method for chest X-rays based on collaborative learning from multiple partially annotated datasets [38]	**Specificity**	80.32	82.11	85.73	81.46	86.97	83.21	80.99	86.46	82.77	81.23	87.23	83.54	81.72	80.94
	Recall	86.48	84.32	81.05	85.12	80.79	83.79	85.93	81.97	84.57	86.21	80.53	83.90	86.79	86.27
	Precision	83.71	83.96	84.90	84.23	85.43	845.68	84.77	85.12	84.32	85.08	85.79	848.21	84.11	82.11
	F1	85.04	84.14	83.31	84.69	83.11	83.99	85.36	83.51	84.44	85.65	83.18	84.21	85.41	84.11
CAPCAM Pooling (Proposed method)	**Specificity**	88.0	89.5	90.8	88.4	93.1	89.8	91.1	88.0	92.5	88.9	93.8	87.7	90.3	93.89
	Recall	90.1	88.6	89.1	91.4	88.9	90.3	88.8	91.9	89.3	90.8	88.7	92.3	89.0	92.49
	Precision	93.8	92.3	91.9	93.0	94.2	93.1	92.6	93.5	94.1	93.0	92.1	93.9	92.8	94.81
	F1	91.9	90.4	90.5	90.7	91.5	91.7	90.6	91.2	91.7	91.8	90.4	91.0	90.9	97.33

*Cardiomegaly, Effusion, Emphysema, Infiltration, Mass, Nodules, Atelectasis, Pneumothorax, Pleural Thickening, Pneumonia, Fibrosis, Edema, Hernia

Table 3
IoBB and Dice Score for PCAM baseline and PCAM + ADN for NIH dataset.

PCAM (Baseline)					**PCAM + Anomaly detection network (ADN)**			
Diseases	**IoBB**		**Dice score**		**IoBB**		**Dice score**	
	Mean ± SD	**Max - Min**	**Mean ± SD**	**Max - Min**	**Mean ± SD**	**Max - Min**	**Mean ± SD**	**Max - Min**
Cardiomegaly	68.59 ± 3.60	73.10–67.21	77.24 ± 1.60	79.46–67.77	76.55 ± 5.30	83.96–59.81	78.18 ± 2.20	78.82–65.17
Effusion	67.45 ± 6.90	79.12–65.31	71.79 ± 1.70	77.86–68.30	74.90 ± 4.30	83.44–66.10	75.99 ± 4.40	81.12–69.99
Emphysema	69.00 ± 6.40	82.03–62.70	77.52 ± 2.40	82.74–72.76	74.97 ± 3.30	78.79–67.09	75.81 ± 1.40	78.65–68.66
Infiltration	76.27 ± 1.90	78.47–69.00	81.76 ± 3.40	84.35–60.26	70.26 ± 2.50	84.02–59.01	73.34 ± 5.90	75.59–67.92
Mass	65.57 ± 5.90	74.43–64.77	75.71 ± 4.20	82.30–64.85	76.22 ± 2.60	82.39–66.72	78.50 ± 6.30	79.29–69.22
Nodules	75.59 ± 1.50	83.59–59.37	75.05 ± 2.30	81.97–62.71	70.15 ± 2.20	79.20–61.26	80.70 ± 1.20	81.51–71.07
Atelectasis	73.59 ± 0.40	83.38–64.44	74.23 ± 2.60	81.07–63.90	72.90 ± 4.10	81.70–62.77	80.64 ± 5.70	83.81–68.67
Pneumothorax	71.93 ± 3.50	87.69–61.63	77.64 ± 3.20	81.34–60.99	77.34 ± 4.30	79.23–67.55	80.68 ± 3.80	82.15–70.34
Pleural Thickening	71.75 ± 3.90	76.88–65.62	78.85 ± 1.40	80.83–67.37	75.47 ± 1.60	83.43–56.73	73.37 ± 2.80	84.19–70.60
Pneumonia	65.86 ± 6.80	87.00–69.07	78.36 ± 1.50	81.00–68.82	79.37 ± 5.70	84.24–56.09	74.09 ± 5.30	82.16–70.50
Fibrosis	74.53 ± 1.50	78.37–67.66	76.49 ± 0.80	80.89–62.92	72.57 ± 1.50	78.18–59.35	80.23 ± 6.30	82.88–66.62
Edema	70.14 ± 5.10	83.55–58.72	81.57 ± 4.70	83.66–63.61	73.75 ± 1.40	84.30–59.28	74.17 ± 1.30	78.68–68.77
Hernia	68.14 ± 2.00	85.86–58.23	76.78 ± 1.10	82.81–72.30	77.70 ± 0.30	84.69–64.42	80.56 ± 5.40	82.01–70.00
Overall	68.58 ± 3.20	79.04–62.13	73.84 ± 2.17	79.65–63.88	72.87 ± 2.00	81.23–63.01	76.61 ± 3.87	77.09–68.74

Table 4
IoBB and Dice Score for PCAM + CAN and PCAM + AND + CAN for NIH dataset.

Diseases	**PCAM + Confidence Aware Network (CAN)**				**PCAM + Anomaly detection network + Confidence aware network**			
	IoBB		**Dice score**		**IoBB**		**Dice score**	
	Mean ± SD	**Max - Min**	**Mean ± SD**	**Max - Min**	**Mean ± SD**	**Max-Min**	**Mean±SD**	**Max-Min**
Cardiomegaly	75.24 ± 1.20	79.15–64.91	80.69 ± 2.30	86.83–66.83	90.90 ± 2.40	94.91–78.82	89.90 ± 1.30	93.83–85.10
Effusion	78.21 ± 5.80	78.64–62.06	78.75 ± 4.70	82.52–70.80	90.12 ± 2.80	93.72–82.00	90.02 ± 1.30	94.21–83.88
Emphysema	75.46 ± 6.90	80.10–67.78	75.24 ± 7.50	86.16–65.03	91.03 ± 3.40	93.55–81.32	90.19 ± 0.20	92.64–80.40
Infiltration	81.96 ± 2.00	78.53–64.13	81.93 ± 5.50	82.38–65.14	86.96 ± 2.20	97.17–76.81	89.43 ± 1.10	94.88–82.32
Mass	80.13 ± 1.90	78.24–68.92	81.56 ± 1.70	86.04–73.96	91.04 ± 2.90	93.25–80.30	90.74 ± 0.70	93.79–82.85
Nodules	76.51 ± 1.10	79.06–64.78	79.09 ± 1.60	84.75–69.27	89.19 ± 2.20	94.05–80.17	89.91 ± 1.30	95.13–83.01
Atelectasis	81.40 ± 3.90	79.05–63.22	77.75 ± 7.80	83.07–68.88	90.98 ± 2.80	94.33–83.94	89.92 ± 0.70	92.45–83.58
Pneumothorax	77.04 ± 6.40	79.67–65.04	74.65 ± 6.50	82.73–71.49	86.28 ± 3.20	97.83–80.62	89.62 ± 1.20	95.44–83.88
Pleural Thickening	80.37 ± 5.40	78.07–67.77	77.71 ± 7.30	82.24–68.61	86.52 ± 3.90	95.76–82.90	90.00 ± 1.70	93.74–82.78
Pneumonia	80.12 ± 0.80	80.62–67.79	80.31 ± 2.70	84.73–73.32	85.20 ± 2.80	97.64–75.87	90.41 ± 1.50	92.56–82.94
Fibrosis	77.95 ± 4.90	80.18–68.68	79.84 ± 4.80	83.56–66.75	85.54 ± 3.10	96.10–81.87	89.40 ± 1.80	94.23–84.06
Edema	75.69 ± 5.90	82.52–62.81	81.65 ± 6.40	84.66–65.89	90.54 ± 3.20	96.26–80.98	90.47 ± 0.50	95.62–83.42
Hernia	80.14 ± 2.50	78.42–64.11	81.07 ± 2.40	83.46–72.87	89.41 ± 3.10	93.56–75.89	90.08 ± 1.10	95.33–82.87
Overall	77.61 ± 3.50	78.71–66.38	77.49 ± 4.60	83.086–68.38	86.69 ± 2.82	93.87–78.11	91.00 ± 1.10	94.74–83.97

6. Ablation study

6.1. Effect of different pooling combination

The ablation study was conducted to evaluate the efficacy of the subsidiary networks in combination with PCAM pooling. The methods are divided into four categories: (i) PCAM baseline, (ii) PCAM + Anomaly Detection Network (ADN), (iii) PCAM + confidence aware network (CAN), and (iv) PCAM + ADN + CAN (proposed method). We evaluated all four approaches using the same set of data and in the same setting. The four approaches are compared to each other for all thirteen thoracic diseases based on IoBB and dice scores. The maximum (Max),

Table 5
IoBB and Dice Score for PCAM baseline and PCAM + ADN for Chexpert dataset.

Diseases	PCAM (Baseline)				PCAM + Anomaly detection network (ADN)			
	IoBB		Dice score		IoBB		Dice score	
	Mean ±SD	Min - Max	Mean ± SD	Min - Max	Mean ± SD	Min - Max	Mean ± SD	Min - Max
Cardiomegaly	63.62 ± 2.28	63.59–63.90	81.16 ± 2.64	77.66–81.87	65.01 ± 0.63	60.48–66.41	80.60 ± 2.73	79.95–82.51
Effusion	50.47 ± 0.50	50.47–50.79	77.14 ± 1.32	77.11–77.24	65.69 ± 1.20	63.52–69.95	79.30 ± 2.56	78.08–80.49
Emphysema	57.74 ± 2.41	52.06–60.51	77.04 ± 2.13	77.02–79.08	64.32 ± 0.97	61.79–71.64	80.88 ± 1.94	73.91–81.13
Infiltration	56.1 ± 2.329	55.89–64.25	81.71 ± 1.81	81.40–81.86	64.28 ± 1.64	63.14–67.05	76.39 ± 1.53	76.25–77.30
Mass	58.66 ± 1.25	55.51–58.88	81.52 ± 1.38	80.91–81.84	68.55 ± 2.10	68.28–68.60	81.51 ± 0.91	81.28–81.69
Nodules	63.78 ± 2.09	63.47–64.60	78.72 ± 2.92	74.70–80.27	72.93 ± 0.51	72.23–74.05	75.95 ± 1.22	75.59–76.08
Atelectasis	53.64 ± 1.58	53.49–54.33	81.49 ± 2.88	81.25–81.68	64.56 ± 1.32	61.48–65.34	74.48 ± 0.93	74.14–81.99
Pneumothorax	54.49 ± 1.48	54.09–54.98	78.92 ± 2.65	77.88–80.79	69.93 ± 2.33	68.78–71.73	76.45 ± 0.68	73.12–80.29
Pleural Thickening	59.83 ± 1.03	58.25–61.62	80.62 ± 2.58	77.99–81.87	62.32 ± 2.13	62.14–62.34	72.36 ± 0.76	72.21–72.39
Pneumonia	63.17 ± 1.72	63.02–63.21	77.19 ± 0.98	76.73–77.67	66.20 ± 0.74	62.27–70.97	75.74 ± 2.35	73.68–76.68
Fibrosis	60.65 ± 2.79	55.04–62.91	75.36 ± 2.24	73.74–75.88	71.68 ± 2.90	70.30–73.57	80.84 ± 1.37	79.99–81.14
Edema	62.72 ± 2.27	61.54–64.37	81.01 ± 2.18	80.90–81.59	65.47 ± 2.69	61.15–69.99	78.85 ± 1.51	76.97–81.95
Hernia	58.76 ± 2.30	56.89–61.05	76.93 ± 1.65	75.27–78.08	70.59 ± 1.82	70.22–71.06	82.60 ± 0.95	82.50–82.66
Overall	59.75 ± 1.81	56.77–58.41	77.13 ± 2.10	77.27–78.84	68.07 ± 1.52	67.06–69.93	78.84 ± 1.37	77.38–78.53

Table 6
IoBB and Dice Score for PCAM + CAN and PCAM + AND + CAN for Chexpert dataset.

Diseases	PCAM + Confidence Aware Network (CAN)				PCAM + Anomaly detection network + Confidence aware network			
	IoBB		Dice score		IoBB		Dice score	
	Mean±SD	Min-Max	Mean±SD	Max-Min	Mean±SD	Min-Max	Mean±SD	Min - Max
Cardiomegaly	73.54 ± 2.76	73.16–77.68	86.83 ± 1.39	80.69–86.83	86.54 ± 2.90	86.16–90.68	88.54 ± 1.41	88.16–92.68
Effusion	78.03 ± 1.25	78.03–78.04	82.52 ± 2.61	78.75–82.52	91.03 ± 0.56	91.03–91.04	93.03 ± 1.38	93.03–93.04
Emphysema	78.83 ± 2.63	78.07–78.89	86.16 ± 1.40	75.24–86.16	91.83 ± 1.78	91.07–91.89	93.83 ± 1.10	93.07–93.89
Infiltration	75.6 ± 2.762	74.19–77.22	82.38 ± 2.41	81.93–82.38	88.6 ± 2.616	87.19–90.22	90.60 ± 2.07	89.19–92.22
Mass	75.28 ± 1.77	75.08–75.87	86.04 ± 2.63	81.56–86.04	88.28 ± 1.95	88.08–88.87	90.28 ± 1.16	90.08–90.87
Nodules	74.01 ± 0.75	72.99–75.21	84.75 ± 2.37	79.09–84.75	87.01 ± 2.66	85.99–88.21	89.01 ± 1.75	87.99–90.21
Atelectasis	78.23 ± 1.77	76.93–78.84	83.07 ± 1.72	77.75–83.07	91.23 ± 1.34	89.93–91.84	93.23 ± 2.96	91.93–93.84
Pneumothorax	78.90 ± 2.37	78.89–78.90	82.73 ± 2.44	74.65–82.73	91.90 ± 2.27	91.89–91.90	93.90 ± 2.33	93.89–93.90
Pleural Thickening	78.23 ± 1.05	78.08–78.60	82.24 ± 1.70	77.71–82.24	91.23 ± 1.53	91.08–91.60	93.23 ± 1.19	93.08–93.60
Pneumonia	72.90 ± 1.46	72.60–76.47	84.73 ± 1.48	80.31–84.73	85.90 ± 1.05	85.60–89.47	87.90 ± 1.38	87.60–91.47
Fibrosis	78.05 ± 1.10	76.45–78.70	83.56 ± 2.21	79.84–83.56	91.05 ± 2.95	89.45–91.70	93.05 ± 2.70	91.45–93.70
Edema	74.87 ± 1.30	74.41–77.82	84.66 ± 1.25	81.65–84.66	87.87 ± 0.88	87.41–90.82	89.87 ± 2.08	89.41–92.82
Hernia	77.84 ± 0.62	76.83–78.88	83.46 ± 1.08	81.07–83.46	90.84 ± 1.51	89.83–91.88	92.84 ± 0.81	91.83–93.88
Overall	76.53 ± 1.23	76.12–77.09	81.69 ± 1.80	79.11–83.77	89.85 ± 1.01	88.38–91.25	91.86 ± 1.40	90.97–93.93

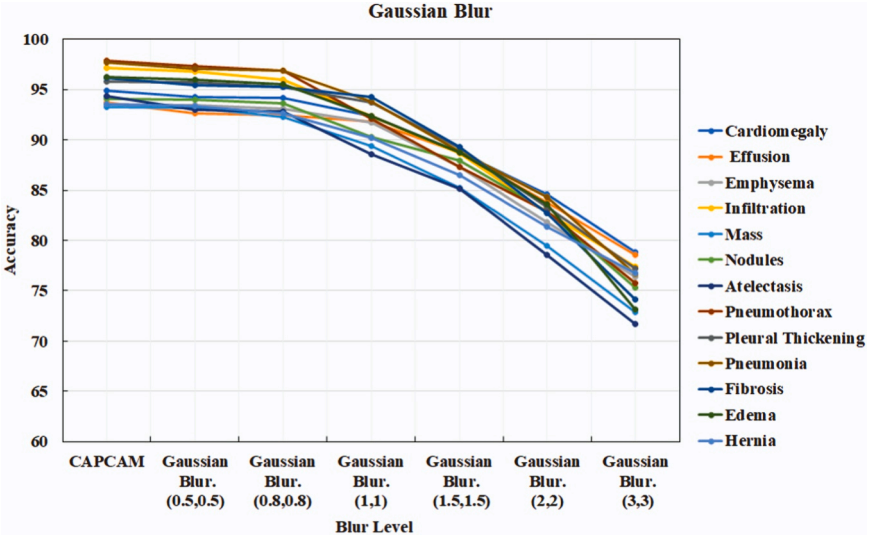

Fig. 10. Effect of Gaussian Blur on the different Kernel standard deviation along X-axis, and along Y-axis.

minimum (Min), mean and standard deviation (SD) for the IoBB as well as dice scores are also calculated. The results of the ablation study on NIH dataset are illustrated in Table 3 and Table 4. The results of the ablation study on Chexpert dataset are illustrated in Table 5 and Table 6 respectively.

Table 3 demonstrates that the baseline PCAM results achieve the highest IoBB score for Pneumothorax and the lowest for Hernia. The addition of the CAN results in the drop of average IoBB scores for various

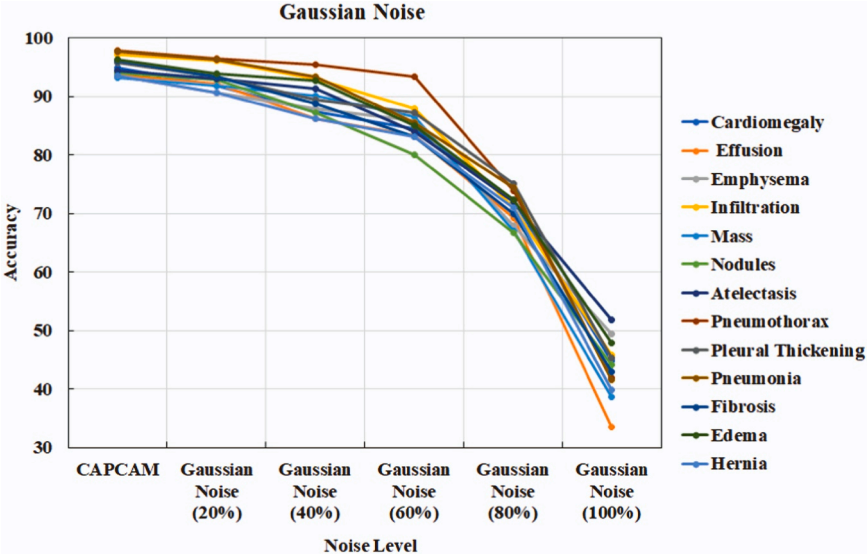

Fig. 11. Effect of Gaussian Noise different standard deviation value.

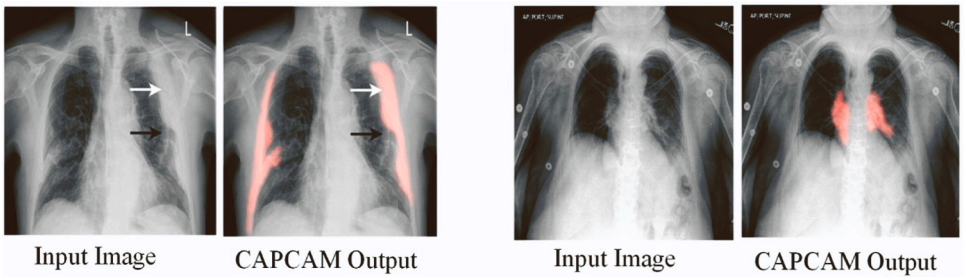

Fig. 12. The results with the presence of artifacts. The localizations are not impacted by the artifacts.

disorders, including as Infiltration, Nodules, Atelectasis, and Fibrosis. This is due to the fact that CAN is an unsupervised network that is highly dependent on the variables X, M, and from Eqs. (3), (8), and (9) respectively, but CAN cannot fetch δ without ADN and therefore $\lambda_{confidence}$ returned is highly unoptimized. The above diseases have a worse outcome since they are related with the localization of discontinuous regions, which is impossible without the probability of abnormality. However, when we compare the outcome to the mean IoBB score obtained after adding an anomaly detection network, it is always 7–12 % greater than the baseline PCAM and 4–9 % greater than PCAM + CAN. The conclusion is supported by the fact that CAN is highly dependent on δ which is generated in the ADN. ADN is a weakly supervised model with trainable parameter (t) which is crucial for localization of the target diseases. Rather than training the network on its own, the same parameter is learned using weights and biases extracted by the feature extractor i.e. CX-Ultranet. Furthermore, ADN is optimized via contrastive loss, which takes into account both the reference score μ_R and the underlying disease β into consideration. CAN is improved by L2 regression loss which depends on δ which is received from ADN. As a result, ADN surpasses CAN even without the use of δ. Our method uses the same δ in CAN resulting in high performance scores. There are no significant variations in dice scores. It rises by 3–7 % for all diseases with the introduction of CAN and ADN. ADN dice scores are higher than CAN dice scores, however the difference is about 5–10 %. Thus, we can say that the combination of our approach achieves a better result.

6.2. Effect of gaussian blur

Gaussian blur is a common image processing technique used to reduce image noise and smooth out details. In our case, we applied Gaussian blur to the testing images to deliberately reduce their quality, inorder to simulate real-world scenarios where images might be of lower quality due to various factors such as motion blur or camera limitations. Fig. 10 represents a line plot showing the relationship between different levels of Gaussian blur applied to the testing images and the corresponding IoBB accuracy achieved by the model for different diseases. From Fig. 9 we conclude that the model's performance is relatively robust to mild levels of blur but degrades rapidly as blur becomes more pronounced. As the Gaussian blur level increases, the image quality decreases, leading to a decrease in IoBB accuracy. This is because as the images become blurrier, it becomes harder for the model to accurately detect and locate the anomaly within them.

6.3. Effect of gaussian noise

In this study, we explore the impact of Gaussian noise on image processing tasks like object detection or classification, using the Signal-to-Noise Ratio (SNR) as a key parameter. Gaussian noise is introduced to the images by adding random values from a Gaussian distribution to the pixel values, and the SNR controls the level of noise added. A higher SNR corresponds to a cleaner image with a stronger signal relative to the noise, while a lower SNR results in a noisier image with weaker signal clarity. To investigate the model's performance, we conduct experiments with various SNR levels, spanning from high SNR for clean images to low SNR for noisy images. By simulating different levels of noise interference, we observe the effect on the model's ability to process the images accurately. As the SNR decreases and noise intensifies, the image quality deteriorates, making it challenging for the model to extract

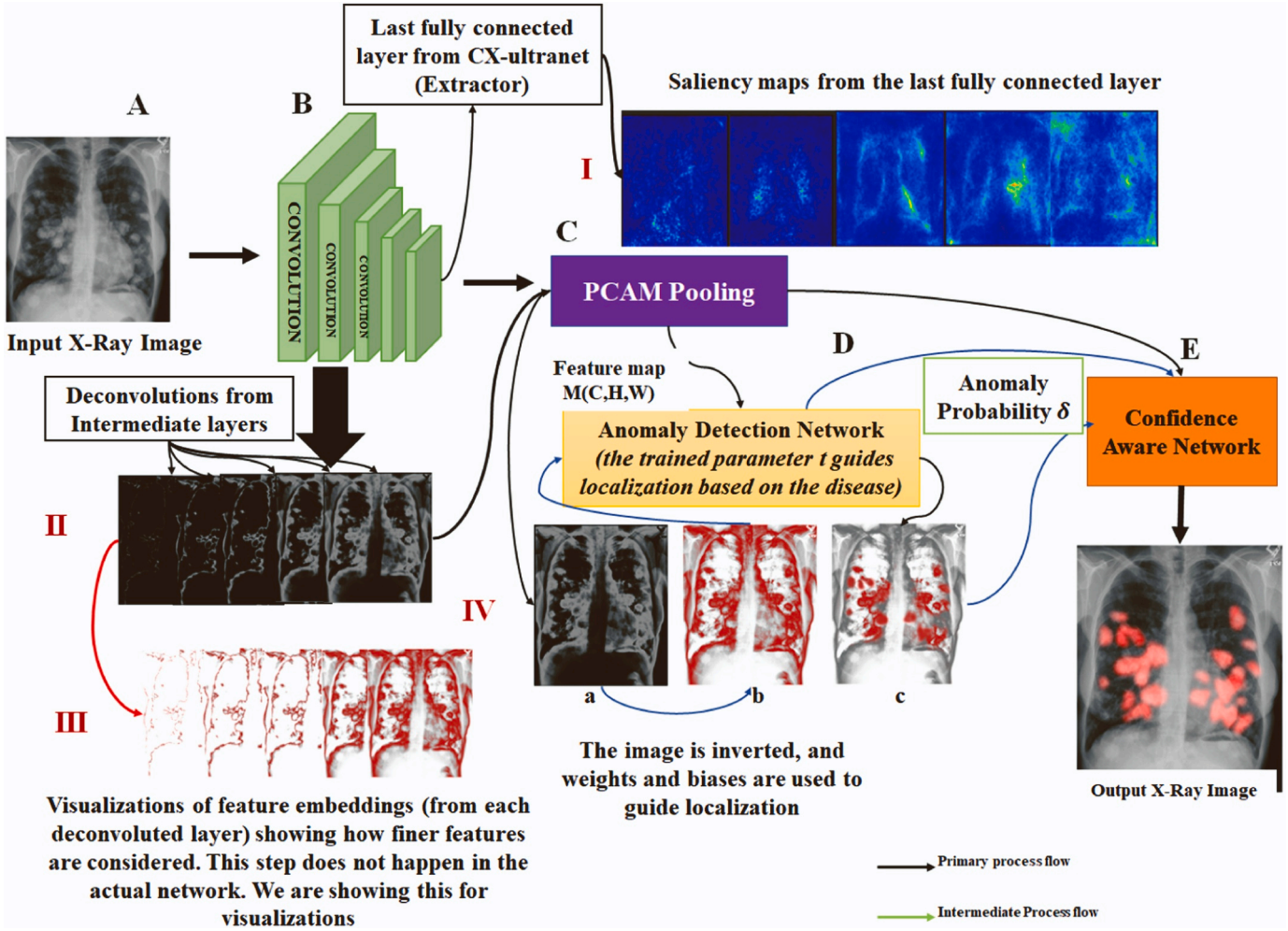

Fig. 13. The overall Explainability workflow of CAPCAM model.

relevant features and patterns, especially for anomaly localization tasks. The model's performance gradually declines as the amount of Gaussian noise increases, impacting its ability to distinguish objects from background noise. Conversely, at high SNR levels, where noise is minimal, the model is likely to perform well due to the presence of a strong and clear signal in the images. Analysing the model's performance under varying SNR levels provides valuable insights into its robustness and limitations in handling noisy image data. By fine-tuning and improving the model based on this analysis, we can enhance its performance in real-world scenarios where noisy images are common. Fig. 11 illustrates the effect of Gaussian noise on the model's performance, displaying how changes in SNR levels correspond to fluctuations in the model's accuracy in processing images. This graphical representation helps in visualizing the relationship between SNR and performance, further informing the decision-making process for optimizing the model's capabilities in image processing tasks.

The method is also tested on with (see Fig. 12) and without artifacts (see Fig. 6). The presence of artifacts does not create any problem in the disease localizations. Two random sample of output or localization on the artifacts image is shown in Fig. 12.

6.4. Explainability

The overall workflow of the explainability of the model is shown in Fig. 13. It provides an insight into the proposed model. Here, the input image is first passed through the convolutional neural network for feature extraction. The first convolutional layer is kept at 3×3 for finer feature selection after which it goes through the unique mobile inverted convolutional layer of CX-Ultranet with dimensions of 3×3 this layer is also responsible for downscaling the images from 512 × 512 × 3–256 × 256 × 32.

The image encounters further mobile inverted convolution layers of 6×6 which again downscale the image. The deconvolutions from the convolution layers are shown in (II) where we can see how the feature develops gradually as it passes through the convolution layers. If we collect the feature embeddings from the intermediate layers and visualize them in their corresponding deconvolutions, then we can visualize the features focused on by the model. The corresponding saliency maps from the intermediate layers are shown in (III) which denotes the regions of interest by the convolutional layers.

Saliency maps are generated (Fig. 13-C) for corresponding neural layers of the CX-Ultranet model. It signifies on which regions the different layers focus on as the DL model is trained. An attention module by itself will not be enough to get saliency maps, however it can be used to refine the saliency maps generated. The saliency maps are a way of interpreting the internal work of the neural layers, so no methods of improvement have been used. For localization task we have used confidence score, anomaly score, probability, and other methods of improvement. After the features are extracted the last feature map is passed on to the PCAM pooling network shown in Fig. 13 (IV a) denoted by M with shape (C, H, W) and the feature embedding is denoted by $M_{i,j}$ which also includes the weights and biases from the feature extractor. After PCAM pooling is done we receive the pooled feature $\boldsymbol{x}$ with which we can guide the first step of localization see Fig. 13(IV b). We also

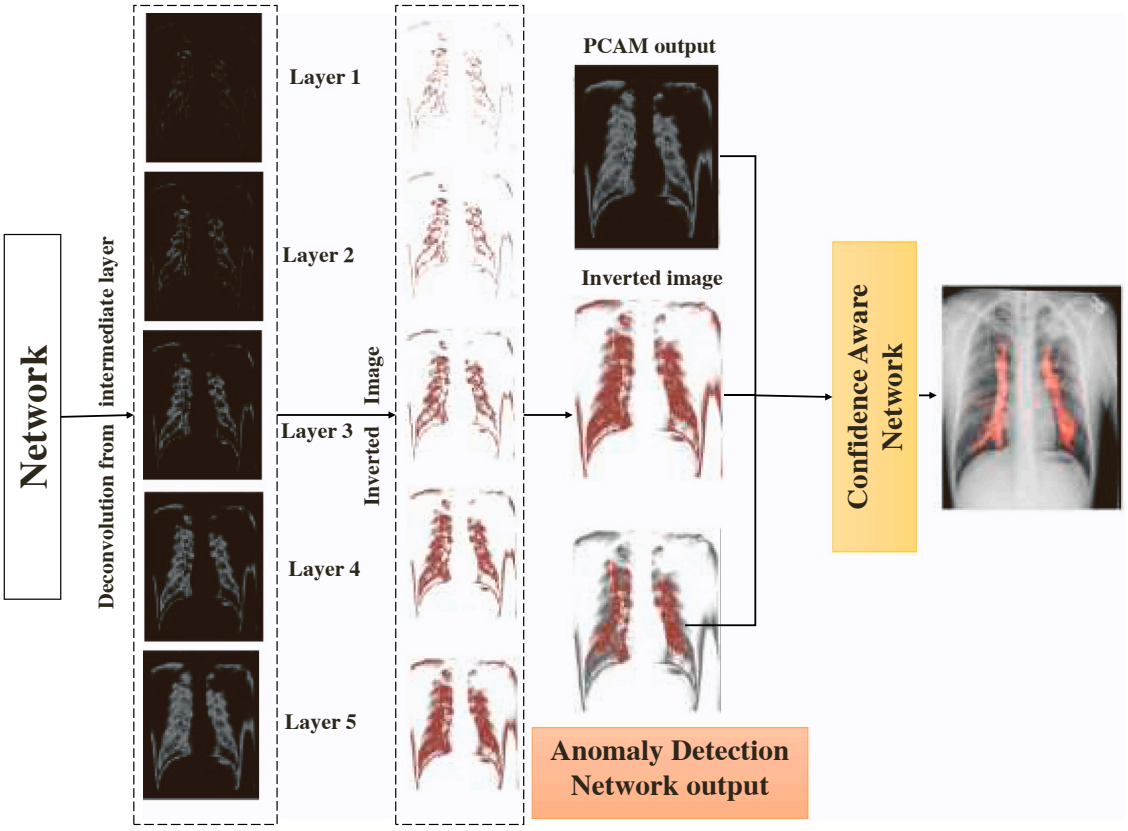

Fig. 14. Visual interpretation of different layers for fibrosis from NIH dataset.

obtain the normalized attention weights $\boldsymbol{w}_{i,j}$ for each pixel denoted by (i, j). The anomaly detection network which is a function of (x,M,t) receives $\boldsymbol{x}$ and M from the previous layer. t is the trainable parameter responsible for guiding localization for different diseases see Fig. 13 (IV c), here it is responsible for guiding localization for the disease nodules. This network explicitly trained on the disease nodules. This network explicitly trained on the weights and biases generated from the feature extractor therefore saving any further training time. The anomaly detection network is also responsible for generating anomaly probability d which is passed to the confidence aware network. The confidence aware network generates confidence scores denoted by ι based on x and M which is optimized using regression of $|\iota - \delta|^2$ based on which the final localizations are done as shown in the output image. Fig. 14 shows the output interpretation of different layers of the CAPCAM model for fibrosis. In future research, we aim to explore deeper into the interpretability of our model by identifying and understanding the weights of the contributing features.

7. Conclusion

In conclusion, this study demonstrates that the proposed CAPCAM model significantly improves weakly supervised localization of multiple thoracic diseases from chest radiographs. Specifically, CAPCAM is able to achieve an accuracy exceeding 85 % across all 13 disease classes - a substantial enhancement over previous state-of-the-art methods. The model maintains robust performance despite varying noise and image quality levels. By integrating Confidence Aware and Anomaly Detection subnetworks, CAPCAM highlights abnormal areas to provide clinical insights into its predictions. The layered architecture elucidates how the model emulates radiologists' focus on prominent regions to enable accurate localization. Translating this approach could streamline clinical workflows by supplying rapid, reliable automated second opinions to reduce misdiagnoses and unnecessary testing. The model's interpretability also promotes practitioner trust in the system. Thus, CAPCAM exemplifies deep learning's potential to complement human expertise in tackling the intricacies of chest x-ray diagnosis. Moving forward, leveraging larger datasets and flexible neural architectures to effectively incorporate clinical knowledge promises continued advances. The aim is enhancing medical decision-making to improve patient outcomes. This study provides a robust framework for weakly supervised localization that avoids intensive manual annotation. By surpassing human-level performance on public benchmarks, CAPCAM demonstrates translational viability to handle real-world clinical data.

Funding and acknowledgement

This research work is supported by the RFIER-Jio Institute's "CVMI-Computer Vision in Medical Imaging" research project fund (Grant No. 2022/33185004) under the AI for ALL" research centre. We are also grateful to Dr. Daksh Dewang Mehta, MBBS, DNB (Radiodiagnosis), Tata Memorial Hospital, Mumbai for his valuable suggestions and insights.

Source code availability

The full source code is available at GitHub link: https://github.com/bad-eastwind/Confidence-Aware-PCAM

CRediT authorship contribution statement

Sudipta Roy: Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization. **Tanushree Meena:** Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Kalyan Tadepalli:** Writing – review & editing, Validation, Project administration, Investigation, Formal analysis, Data curation. **Anwesh**

Kabiraj: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

All public data sets were used and mentioned in the dataset section.

References

[1] Roth, A. Gregory, et al., Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017, Lancet 392 (10159) (2018) 1736–1788.

[2] Yulei Jiang, et al., Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications, Radiology 220 (3) (2001) 787–794.

[3] Van Ginneken Bram, B.M. Ter Haar Romeny, Max A. Viergever, Computer-aided diagnosis in chest radiography: a survey, IEEE Trans. Med. Imaging 20 (12) (2001) 1228–1241.

[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016) 2921–2929.

[5] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. BatraVisual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[6] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C.L. ZitnickMicrosoft COCO captions: Data collection and evaluation server, arXiv preprint arXiv: 1504.00325..

[7] H. Fang, S. Gupta, F. Iandola, R.K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J.C. Platt, From captions to visual concepts and back, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1473–1482.

[8] X. Hu, J. Dai, Y. Huang, H. Yang, L. Zhang, W. Chen, G. Yang, D. Zhang, A weakly supervised framework for abnormal behavior detection and localization in crowded scenes, Neurocomputing 383 (2020) 270–281.

[9] K.Kumar Singh, Y.Jae LeeHide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3524–3533..

[10] C. Yan, J. Yao, R. Li, Z. Xu, J. HuangWeakly supervised deep learning for thoracic disease classification and localization on chest x-rays. In Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics (pp. 103-110).

[11] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. SummersSummers, Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2097–2106.

[12] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. 2015. Going deeper with convolutions. Cvpr.

[14] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

[15] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv preprint arXiv:1711.05225 (2017).

[16] J. Cai, L. Lu, A.P. Harrison, X. Shi, P. Chen, L. YangIterative attention mining for weakly supervised thoracic disease pattern localization in chest x-rays, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 589–598.

[17] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, L. Fei-FeiThoracic disease identification and localization with limited supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8290–8299.

[18] H.-P. Kriegel, P. Kroger, E. Schubert, A. ZimekInterpreting and unifying outlier scores," in Proceedings of the 2011 SIAM International Conference on Data Mining. SIAM, 2011, pp. 13-24.

[19] G. Pang, C. Shen, A. van den HengelDeep anomaly detection with deviation networks," in ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 353–362.

[20] T.Chen, S.Kornblith, M.Norouzi, and G.Hinton, "A simple framework for contrastive learning of visual representations," arXiv:2002.05709, 2020.

[21] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. SummersChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly- Supervised Classification and Localization of Common Thorax Diseases, IEEE CVPR, pp. 3462-3471, 2017.

[22] Y. Wang, J. Li, F. MetzeA comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 31–35. IEEE, 2019.

[23] P.O. Pinheiro, R. CollobertFrom image-level to pixel-level labeling with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1713–1721, 2015.

[24] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, A.Y. NgChexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 590-597).

[25] Anwesh Kabiraj, Tanushree Meena, Pailla Balakrishna Reddy, Sudipta Roy, Detection and classification of lung disease using deep learning architecture from x-ray images. International Symposium on Visual Computing, Springer International Publishing, Cham, 2022, pp. 444–455.

[26] J. Liu, G. Zhao, Y. Fei, M. Zhang, Y. Wang, Y. YuBoundary-Enhanced Co-Training for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19574-19584).

[27] M.Y. Lu, D.F.K. Williamson, T.Y. Chen, et al., Data-efficient and weakly supervised computational pathology on whole-slide images, Nat. Biomed. Eng. 5 (2021) 555–570, https://doi.org/10.1038/s41551-020-00682-w.

[28] S. Rong, B. Tu, Z. Wang, J. LiBoundary-Enhanced Co-Training for Weakly Supervised Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19574-19584).

[29] S. Jo, I.J. YuPuzzle-cam: Improved localization via matching partial and full features. In 2021 IEEE International Conference on Image Processing (ICIP) (pp. 639-643). IEEE.

[30] J. Rocha, S.C. Pereira, J. Pedrosa, A. Campilho, A.M. Mendonça, STERN: attention-driven spatial transformer network for abnormality detection in chest X-ray images, Artif. Intell. Med. vol. 147 (Jan. 2024) 102737, https://doi.org/10.1016/j.artmed.2023.102737.

[31] G. Wang, "MRChexNet: Multi-modal bridge and relational learning for thoracic disease recognition in chest X-rays," Math. Biosci. Eng., vol. 20, no. 12.

[32] Q. Xu, W. Duan, DualAttNet: synergistic fusion of image-level and fine-grained disease attention for multi-label lesion detection in chest X-rays, Comput. Biol. Med. vol. 168 (Jan. 2024) 107742, https://doi.org/10.1016/j.compbiomed.2023.107742.

[33] T. Meena, A. Kabiraj, P.B. Reddy, S. Roy, Weakly supervised confidence aware probabilistic cam multi-thorax anomaly localization network (Bellevue, WA, USA), 2023 IEEE 24th Int. Conf. Inf. Reuse Integr. Data Sci. (IRI) (2023) 309–314, https://doi.org/10.1109/IRI58017.2023.00061.

[34] T. Meena, K. Sarawadekar, An explainable self-attention-based spatial–temporal analysis for human activity recognition, IEEE Sens. J. vol. 24 (1) (2024) 635–644, https://doi.org/10.1109/JSEN.2023.3335449.

[35] Sudipta Roy, Debojyoti Pal, Tanushree Meena, Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead, Netw. Model. Anal. Health Inform. Bioinforma. 13 (1) (2023) 4, https://doi.org/10.1007/s13721-023-00437-y.

[36] A. Kabiraj, T. Meena, P.B. Reddy, et al., Multiple thoracic diseases detection from X-rays using CX-Ultranet, Health Technol. 14 (2024) 291–303, https://doi.org/10.1007/s12553-024-00820-3.

[37] A. Sulaiman, V. Anand, S. Gupta, Y. Asiri, M.A. Elmagzoub, M.S.A. Reshan, A. Shaikh, A convolutional neural network architecture for segmentation of lung diseases using chest X-ray images, Diagnostics *13* (2023) 1651, https://doi.org/10.3390/diagnostics13091651.

[38] H. Wang, D. Zhang, J. Feng, L. Cascone, M. Nappi, S. Wan, A multi-objective segmentation method for chest X-rays based on collaborative learning from multiple partially annotated datasets, Inf. Fusion *102* (2024) 102016.