CLN: A multi-task deep neural network for chest X-ray image localisation and classification

Gabriel Iluebe Okolo[a], Stamos Katsigiannis[b], Naeem Ramzan[a,*]

[a] *University of the West of Scotland, School of Computing, Engineering and Physical Sciences, High St., Paisley, PA1 2BE, UK*
[b] *Durham University, Department of Computer Science, Upper Mountjoy Campus, Stockton Road, Durham, DH1 3LE, UK*

Abstract

Chest X-ray (CXR) imaging is a widely used and cost-effective medical imaging technique for detecting various pathologies. However, accurate interpretation of CXR images is a challenging and time-consuming task that requires expert radiologists. Although deep learning methods have demonstrated high performance in CXR image classification, concerns over interpretability limit their clinical adoption. Localising pathologies on chest X-rays could improve interpretability and trust in these systems. In this work, we propose the Chest X-ray Localisation Network (CLN), a multi-task deep neural network designed to localise and classify pathologies in CXR images. Our proposed architecture was trained and evaluated on a subset of the ChestX-ray14 CXR data set, which included bounding box annotations of eight different pathologies from expert radiologists, achieving a maximum classification mean AUC score of 0.918 and a maximum localisation mean IoU accuracy of 0.855 for the eight examined pathologies (atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax). Our approach outperformed state-of-the-art methods, demonstrating its potential as a reliable solution for computer-aided CXR image diagnosis, offering notable advantages over existing methods, including superior classification and localisation accuracy, reduced performance decay with increased IoU thresholds, and an overall simpler architecture.

Keywords: chest radiography, X-rays, localisation, image classification, deep learning

1. Introduction

Chest X-ray (CXR) imaging is one of the most widely used medical imaging tests in clinical practice (Kelly et al., 2016). In comparison to other medical imaging techniques, chest radiography is widely accessible and reasonably priced (Raoof et al., 2012; Chandra et al., 2022). Chest X-rays are frequently used as a frontline diagnostic imaging modality, particularly in underdeveloped areas of the globe dealing with a high burden of infectious illnesses (Dhoot et al., 2018), as they can be portable and have reduced setup and operating expenses. CXR scans provide vital information about a patient's health, and skilled radiologists can manually inspect them and detect illness markers and indicators. However, this is a challenging process that is arduous and time-consuming, requiring expert radiologists in order to avoid misinterpreting or overlooking important markers on the CXR. In addition, this is a manual process that is prone to fatigue-related errors.

The burden on radiologists can be decreased by using automated CXR analysis to help with population screening, triaging and interpretation (Doi, 2007). Also, automated CXR analysis can offer the frontline practitioner a useful visual help for diagnosing an illness, and can reduce the variation in reading across radiologists, better identify aberrant cases for expert interpretation, and even act as a second reader throughout the diagnostic decision-making process (Yu et al., 2011). It can also be considerably helpful in remote areas that lack sufficient medical personnel. Recent research on Computer Aided Diagnosis (CAD) systems has focused on the application of machine learning techniques for automating CXR-based diagnosis (Litjens et al., 2017; Ker et al., 2017), with various deep learning methods being proposed in the literature for CXR image classification and/or localisation, e.g. (Lakhani and Sundaram, 2017; Annarumma et al., 2019; Chandra et al., 2022; Brunese et al., 2020; Okolo et al., 2022; Pereira et al., 2020; Zhang et al., 2023). The public release of the large-scale NIH Chest X-ray14 (Wang et al., 2017) and CheXpert (Irvin et al., 2019) data sets, each with more than 100,000 CXR images encouraged further research in the area.

Deep learning (DL) methods have recently achieved impressive performance in CXR image classification, but physicians are reluctant to adopt and trust such systems due to their lack of interpretability. A potential solution to this problem lies in the localisation of detected pathologies within CXR images. Classification and localisation are two different but related problems in computer vision. Classification is the task of assigning a label to an image, while localisation is the task of identifying the location of an object or anomaly in an image. In the case of CXR images, the later is of utmost importance to physicians as it would enable them to establish trust in deep learning models and enhance their interpretability. Low abnormality lo-

*Corresponding author
Email addresses: gabriel.okolo@uws.ac.uk (Gabriel Iluebe Okolo), stamos.katsigiannis@durham.ac.uk (Stamos Katsigiannis), naeem.ramzan@uws.ac.uk (Naeem Ramzan)

calisation accuracy and poor model interpretability have thus emerged as major roadblocks to the widespread implementation of these methods in clinical practise. To this end, visual attention is increasingly being used for model interpretability and explainability as a result of recent developments in convolutional neural network (CNN) attention modelling and learning (Zhou et al., 2016; Selvaraju et al., 2017), whereas more recent approaches for CXR image localisation employ transformers and self-attention mechanisms (Ouyang et al., 2021; Han et al., 2023).

Many older works on automated CXR image analysis focused solely on classification using CNNs. Bar et al. (Bar et al., 2015) used a DL approach that combined features retrieved by a deep CNN model with low-level features for diagnosing pleural effusion, cardiomegaly, and normal vs. abnormal cases. Their method achieved an area under the curve (AUC) score of 93%, 89%, and 79% for pleural effusion, cardiomegaly, and normal vs. abnormal, respectively. Cicero et al. (Cicero et al., 2017) utilised the GoogleNet model, achieving AUC scores between 85%-96.4% for six different pathologies. Wang et al. (Wang et al., 2017) employed a weak-supervised technique for the classification of eight pathologies on the ChestX-ray8 CXR data set, achieving an average AUC of 80.3%. A DenseNet-121 model (CheXNet) was used in (Rajpurkar et al., 2017) on the ChestX-ray14 data set that contains 14 pathologies, reaching an average AUC of 84.11%. A cascading neural network was utilised by Kumar et al. (Kumar et al., 2018) on the same data set. They used under-sampling and over-sampling strategies to reduce bias resulting from unbalanced data, achieving an average AUC of 79.5%. Majdi et al. (Majdi et al., 2020) proposed a custom DenseNet-121 for identifying cardiomegaly and pulmonary nodules, reaching an AUC of 73% for pulmonary nodule identification, and 92% for cardiomegaly detection. Zhao et al. (Zhao et al., 2021) also worked on the Chest-Xray14 data set and proposed a deep CNN model with attention mechanism (AMDenseNet), achieving an average AUC of 85.37%. Blais and Akhloufi (Blais and Akhloufi, 2021) used various models with binary relevance to identify chest pathologies. When combined with the Adam optimiser, the Xception deep CNN model outperformed other models, reaching a mean AUC of 95.87% on 6 pathologies and 94.90% on the CheXpert data set's 14 pathologies. More recent approaches for CXR image classification employed transformers, such as the work by Okolo et al. (Okolo et al., 2022), which proposed IEViT, an enhanced version of the Vision Transformer for CXR image classification and evaluated it on four different CXR data sets for turberculosis, pneumonia, and COVID-19, achieving a maximum average F1-score between 96.39% and 100%.

CXR classification and localisation are two interconnected tasks. CXR classification aims to classify a given CXR image into different predefined classes, typically referring to specific pathologies, whereas localisation focuses on identifying the regions of interest (ROIs) within the CXR image that refer to specific abnormalities. Accurate localisation can help in establishing trust to a DL model as it can enhance its interpretability, thus potentially enabling its use in clinical practice. Zhou et al. (Zhou et al., 2016) introduced the concept of Class Activation Mapping (CAM) for localising ROIs in an image. CAM provides a way to generate heat-maps that highlight the discriminative regions contributing to the classification decision, providing insights into the ROIs for abnormality detection. By applying a thresholding technique to separate ROIs, the localised regions are extracted, providing a visual explanation of the areas that significantly contribute to the target class predicted by a CNN model (Selvaraju et al., 2017; Gascoigne-Burns and Katsigiannis, 2022).

Multiple CNN-based CXR localisation methods have been proposed in the literature. Among them, Rajpurkar et al.'s (Rajpurkar et al., 2017) CheXNet DL model achieved performance comparable to expert radiologists in pneumonia diagnosis. Despite focusing on classification, the model implicitly learns to identify and localise pneumonia regions within CXR images. Wang et al. (Wang et al., 2017) showed that a combined weakly supervised multi-label image classification and disease localisation framework is capable of detecting and even spatially localising thoracic diseases by using the activation and weights acquired from the network. Li et al. (Li et al., 2018) proposed a unified classification and localisation approach that leverages both class information and limited location annotation by first using a CNN to process the input image, and then divide the image into a grid of patches to capture the local information specific to the disease.

More recent CXR localisation techniques focused on the use of transformers and attention-based mechanisms. Han et al. (Han et al., 2023) proposed the Radiomics-Guided Transformer (RGT) model that combines global image information with local radiomics-guided auxiliary data for accurate pathology localisation and classification without requiring bounding box annotations. RGT consists of image and radiomics Transformer branches, fusion layers for aggregating information, and cross-attention layers for interaction between image and radiomics features. Ouyang et al. (Ouyang et al., 2021) introduced a novel attention-driven weakly supervised algorithm that combines activation- and gradient-based visual attention through a hierarchical attention mining framework. In another work, Qi et al. (Qi et al., 2022) utilised a pre-trained U-Net in order to segment the lung lobes and an intra-image graph to compare different regions of the lobes to achieve weakly-supervised disease localisation. It is evident from the literature that these newer methods outperform the CAM-based ones that rely on CNNs. While these latest advancements in CXR localisation have demonstrated promising results, it is essential to consider the trade-off between complexity and practical deployment. The increased complexity comes at the cost of higher computational resource requirements, longer training times, and complex optimisation to ensure efficient inference.

In this work, we propose the *CXR Localisation Network* (CLN), a multi-task deep neural network for the task of CXR image classification and localisation. CLN relies on a pre-trained CNN backbone that is then fine-tuned on a large CXR data set with 14 pathologies. The backbone is then followed by two parallel network branches, one for pathology classification, and one for pathology localisation by predicting a bounding box through regression. CLN was trained and evaluated on a subset

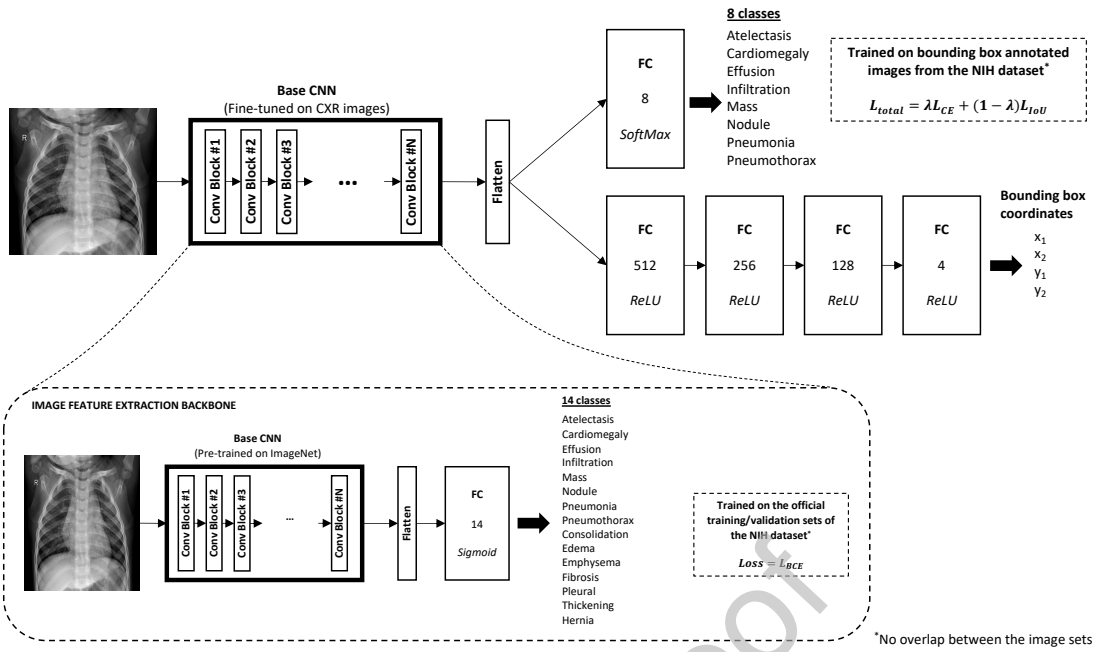

Figure 1: An overview of the proposed CXR Localisation Network (CLN) method. CE: Cross-entropy, BCE: Binary Cross Entropy, IoU: Intersection over the Union, FC: Fully-connected, ReLU: Rectified Linear Unit.

of the NIH ChestX-ray14 (Wang et al., 2017) data set that contains CXR images annotated with bounding boxes for 8 different pathologies. Our experimental evaluation showed that the proposed CLN architecture outperformed state-of-the-art methods in both the localisation and classification tasks, achieving a maximum classification mean AUC score of 0.918 and a maximum localisation mean Intersection over Union (IoU) accuracy of 0.855.

The motivation behind our proposed Chest X-ray Localisation Network (CLN) stems from three key challenges in adopting AI for medical imaging. First, localisation of pathologies enhances explainability, a crucial factor in increasing trust in AI-based diagnostic systems. By visualising the specific regions associated with each diagnosis, the model can provide interpretable results that are more accessible to radiologists and healthcare providers. Second, establishing trust among doctors is essential for real-world deployment; by clearly indicating areas of interest within the CXR, our model addresses the hesitation to rely on "black-box" AI systems. Third, while existing methods show promising accuracy, many have high computational demands, which can hinder their practicality in clinical environments with limited processing capabilities. Our approach aims to balance performance and computational efficiency, offering an interpretable, high-performing model that requires lower computational resources than more complex networks. Furthermore, contrary to most state-of-the-art CXR localisation methods, our approach utilises localisation annotations during training and learns to directly predict the location of abnormalities, instead of relying purely on the classification labels.

The contributions of this work can be summarised as follows: (i) A novel multi-task architecture for CXR image localisation and classification that outperforms the state of the art. (ii) Relative simplicity compared to state-of-the-art methods by utilising CNNs and fully-connected layers, instead of transformers, attention mechanisms, and complex network structures. (iii) Significantly improved stability in terms of localisation performance with regards to the Intersection over Union (IoU) threshold, compared to state-of-the-art methods, with mean IoU accuracy decaying much slower as the threshold increases.

Our source code is avalable to facilitate further research[1].

2. Material and methods

In this work, we propose the *CXR Localisation Network* (CLN), a multi-task deep neural network for the task of CXR image localisation and classification. The proposed architecture relies on a backbone that is fine-tuned for CXR image classification on a large CXR dataset that contains images with 14 pathologies. The network is then divided into two branches, one designed to classify the input CXR image into one of the available pathology classes, and one designed for the task of localisation by performing regression in order to estimate the coordinates of the bounding box that denotes the region of the CXR that contains the signs of the detected pathology. An overview of the proposed architecture is provided in Figure 1. The proposed architecture was trained to support the following 8 pathologies: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax.

2.1. CXR data set

To develop and evaluate the proposed architecture, we used the National Institute of Health's (NIH) ChestX-ray14 CXR

[1] A download link will be provided upon acceptance of the paper

data set (Wang et al., 2017). The data set contains 112,120 CXR images of 30,805 subjects. Images are in PNG format with a resolution of 1024×1024, and are associated with one or more of 14 different pathologies, i.e. atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia. Furthermore, the data set contains 984 bounding box coordinates for 880 of the CXR images that were hand-annotated by board-certified radiologists for 8 pathology classes: atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax. The NIH data set comes with an official training/validation/test split, with all the 880 bounding-box annotated images belonging to the test set.

In this work, the data set was utilised in two different manners. The official training and validation sets were used for fine-tuning the feature extraction backbone of the proposed architecture, whereas the subset of 880 bounding box annotated CXR images was used for training and evaluating the final multi-task network, which required bounding boxes for its training. Random stratified sampling was used to split the subset of 880 images into a training (80%) and test (20%) set, with the training set being further divided into a training (90%) and a validation (10%) set. Furthermore, all images were resized to 224×224 pixels. It must be noted that there was no overlap between the images used to fine-tune the feature extraction backbone and the ones used to train and evaluate the final multi-task network, thus there is no danger of data leakage and overfitting.

2.2. The CXR Localisation Network (CLN)

As illustrated in Figure 1, the proposed CLN architecture takes as an input a CXR image. The CXR image is then fed into a CNN-based image feature extraction backbone that has been fine-tuned on CXR images for the classification of various pathologies. The output of the convolutional base of the backbone is then flattened and passed to two separate branches of the architecture, one for CXR image classification and one for CXR image localisation. Thus, the features extracted by the backbone are shared across the two branches of the architecture.

2.2.1. Image feature extraction backbone

During the first stage of the proposed multi-task neural network architecture, the input CXR image is passed through a feature extraction module in order to compute an appropriate representation of the image. This feature extraction module constitutes the backbone of the proposed architecture and relies on a convolutional neural network (CNN) that has been pre-trained on the task of image classification. For our network's backbone, we opted to use pre-trained models trained on the extensive ImageNet (Krizhevsky et al., 2017) data set. This data set consists of 1.4 million annotated images belonging to 1,000 different classes, and these models have demonstrated exceptional efficiency in extracting image features, as evidenced by their outstanding classification performance in various applications. The models examined in this work as the backbone of the proposed architecture were CheXnet (i.e. a DenseNet121) (Rajpurkar et al., 2017), EfficienetB4 (Tan and Le, 2019), InceptionV3 (Szegedy et al., 2016), Resnet50V2 (He et al., 2016), Xception (Chollet, 2017), and MobileNetV3Small (Howard et al., 2019). It must be noted that the Keras implementations of the pre-trained models were used for this work.

The pre-trained feature extraction backbone was fine-tuned on CXR images by training it using the full NIH CXR data set that contained 14 pathologies. To this end, as shown in Figure 1, a flatten layer was added after the convolutional base of the backbone network, followed by a fully connected layer of size 14. Given that each image can be associated with multiple labels (multi-label classification), a Sigmoid activation function was used in the output layer for the final classification. The network was then trained end-to-end on the training set using binary cross-entropy as the loss function. The final model was then selected as the model from the training epoch that provided the best classification performance on the validation set.

Given the small number of bounding-box annotated CXR images in the data set compared to the total number (880 vs. 112,120), as well as the smaller number of pathologies (8 vs. 14), we opted to fine-tune the feature extraction backbone on the full NIH data set in order to create a more efficient and generalisable CXR image feature extractor by exploiting the significantly large number of images in the full data set. Furthermore, it must be noted that the annotated images were not included in the training and validation sets of the full data set.

As illustrated in Figure 1, the convolutional base of the fine-tuned model was then used as the first stage of the proposed multi-task neural network architecture, followed by a flattening layer. After the flattening layer, the network is subsequently split into two parallel branches, one targeting CXR image classification, and one targeting CXR image localisation. Considering that various CNNs can be used as the image feature extraction backbone of the proposed CLN architecture, we utilise the following naming convention for clarity: CLN-BackboneCNN. For example, CLN-ResNet50V2 refers to our proposed CLN architecture using a ResNet50V2 model as its image feature extraction backbone.

2.2.2. CXR image classification module

For the CXR image classification module of the proposed architecture, the output of the flattening layer after the feature extraction backbone is passed to a fully-connected layer of size 8 that uses a SoftMax activation function for the final classification into the 8 pathologies included in the annotated subset of the NIH data set, i.e. atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax. Given that each annotated image is associated with a single label (multi-class classification), the Softmax activation function was used in the output layer for the final classification. It must be noted that, since the loss function that will be used for training will take into consideration both classification and localisation performance, the network can only be trained with CXR images for which localisation annotations exist. Consequently, out of the 14 pathologies included in the NIH dataset, only the 8 included in the annotated subset can be supported by the CXR image classification module.

2.2.3. CXR image localisation module

For the CXR image localisation module of the proposed architecture, the output of the flattening layer after the feature extraction backbone is passed through a series of four fully-connected layers. The first, second, and third fully-connected layers have a size of 512, 256, and 128, respectively, and they all use a rectified linear unit (ReLU) activation function. The final fully-connected layer has a size of 4 and also uses a ReLU activation function in order to ensure that the output values cannot be negative. The four outputs of the final fully-connected layer correspond to the values x_1, x_2, y_1, y_2 of a bounding box defined by the coordinates (x_1, y_1), (x_1, y_2), (x_2, y_1), (x_2, y_2). The CXR image localisation branch in our proposed architecture utilises the CXR image features extracted by the feature extraction backbone module. Its purpose is to estimate the region (bounding box) within the CXR image that corresponds to the pathology prediction made by the CXR image classification branch, employing a regression approach.

2.3. Loss function & Training

To train the proposed architecture, we propose the use of a loss function that combines two loss metrics, one for classification and one for localisation, with their combination controlled by a hyperparameter λ that defines the contribution of each loss to the total loss. For the classification metric, the Cross-Entropy loss (L_{CE}) was selected:

$$L_{CE} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}), \tag{1}$$

with $M = 8$ the number of classes, $y_{o,c} \in \{0, 1\}$ denoting whether observation o belongs to class c, and $p_{o,c}$ the predicted probability that o is of class c. For the localisation metric, we selected the *Intersection over Union* (IoU) metric, which describes the extent of overlap between the predicted bounding box and the ground-truth bounding box, with higher overlap leading to higher IoU values. IoU is defined as

$$IoU(B_{pred}, B_{gt}) = \frac{|B_{pred} \cap B_{gt}|}{|B_{pred} \cup B_{gt}|}, \tag{2}$$

where B_{pred} is the predicted bounding box and B_{gt} is the ground-truth bounding box. IoU ranges between 0 and 1, with 0 denoting no overlap and 1 denoting perfect overlap. Higher overlap corresponds to better localisation and leads to higher IoU values, thus we define the IoU loss as:

$$L_{IoU} = 1 - IoU(B_{pred}, B_{gt}) \tag{3}$$

The final loss function for the training of the proposed architecture is then defined as:

$$L_{total} = \lambda L_{CE} + (1 - \lambda) L_{IoU} \tag{4}$$

The contribution of the classification loss (L_{CE}) to the total loss (L_{total}) is controlled by the hyperparameter λ, while the contribution of the localisation loss (L_{IoU}) is controlled by $(1 - \lambda)$. By combining these two losses, the overall loss function aims to capture both classification and localisation errors, with the trade-off between the two losses controlled by the hyperparameter λ. The value of the hyperparameter λ is critical to the performance of the proposed method. Its effect on localisation and classification performance is thoroughly examined in Section 3.4.2.

Based on preliminary experimentation for selecting the training hyperparameters, the proposed multi-task neural network architecture was trained using the proposed loss function and the Adam (Kingma and Ba, 2014) optimiser with a batch size of 16 and an initial learning rate of 0.001 that was reduced by a factor of 2 after every 4 epochs, with a lower limit of 0.000001. All experiments were conducted using the TensorFlow library and the Keras API on a GeForce RTX 4090 24GB GPU. Early stopping was also applied, with training stopping after 20 epochs with no improvement in validation loss.

3. Results

3.1. Evaluation Protocol & Metrics

The proposed multi-task neural network architecture for CXR image localisation and classification underwent training and evaluation using the subset of 880 annotated CXR images, encompassing 8 different pathologies. The performance of the network was then compared against state-of-the-art methods. Six variants of the proposed CLN architecture were evaluated, CLN-ResNet50V2, CLN-EfficientNetB4, CLN-InceptionV3, CLN-Xception, CLN-MobileNetV3Small, and CLN-CheXNet, each differing in terms of the image feature extraction backbone used. To ensure a fair performance evaluation, localisation and classification performance was evaluated according to the performance metrics utilised in the compared works. To this end, classification performance was evaluated in terms of the area under the ROC curve (AUC), whereas localisation performance was evaluated using the IoU metric, computed using the predicted and the ground truth bounding boxes.

As is common across the literature (Han et al., 2023; Wang et al., 2017; Li et al., 2018; Ouyang et al., 2021), IoU accuracy is defined by considering the localisation to be correct when the computed IoU is larger than a fixed IoU threshold $T(IoU)$. Consequently, the reported localisation accuracy for a given threshold $T(IoU)$ denotes the percentage of images for which the IoU between the predicted and the ground truth bounding box was higher than the threshold $T(IoU)$. The proposed CLN architecture was evaluated for seven different IoU thresholds, i.e. $T(IoU) \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. Evaluating IoU at various thresholds is a standard practice in object detection to comprehensively assess a model's localisation performance. IoU measures the overlap between the predicted bounding box and the ground truth, providing a quantitative metric for localisation accuracy (Zheng et al., 2020; Luo et al., 2024). By analysing performance across a range of IoU thresholds, we can determine how well the model balances precision and recall at different levels of overlap. This approach is particularly relevant in medical imaging, where precise localisation is crucial for accurate diagnosis and treatment planning.

Table 1: Pathology localisation results using state-of-the-art methods and the proposed method for six different backbone models on the NIH chest X-ray dataset. Results are measured by IoU accuracy at a fixed threshold ([0.1,0.7]).

T(IoU)	Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Mean
0.1	ViT-based (Han et al., 2023)	0.58	0.91	0.61	0.77	0.44	0.11	0.75	0.25	0.553
	Wang et al. (Wang et al., 2017)	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38	0.569
	RGT (Han et al., 2023)	0.61	0.95	0.65	0.82	0.50	0.13	0.79	0.28	0.591
	Li et al. (Li et al., 2018)	0.71	0.98	0.87	0.92	0.71	0.40	0.60	0.63	0.728
	Ouyang et al. (Ouyang et al., 2021)	0.71	**1.00**	**0.89**	0.88	0.76	0.65	0.91	0.78	0.820
	CLN-ResNet50V2 (ours)	0.76	0.93	0.73	0.92	0.65	**0.94**	0.92	0.72	0.822
	CLN-MobileNetV3Small (ours)	**0.89**	**1.00**	0.75	0.84	0.67	0.64	0.96	**0.95**	0.837
	CLN-EfficientNetB4 (ours)	0.84	0.97	0.73	**1.00**	0.56	**0.94**	0.92	0.79	0.844
	CLN-InceptionV3 (ours)	**0.89**	**1.00**	0.75	0.86	**1.00**	0.77	0.91	0.62	0.850
	CLN-Xception (ours)	**0.89**	**1.00**	0.79	0.83	0.73	0.77	**1.00**	0.81	0.852
	CLN-CheXNet (ours)	0.79	**1.00**	0.83	0.84	0.88	0.75	**1.00**	0.75	**0.855**
0.2	ViT-based (Han et al., 2023)	0.38	0.85	0.39	0.55	0.24	0.01	0.51	0.15	0.385
	Wang et al. (Wang et al., 2017)	0.47	0.68	0.45	0.48	0.26	0.05	0.35	0.23	0.371
	RGT (Han et al., 2023)	0.41	0.91	0.41	0.59	0.26	0.05	0.57	0.19	0.424
	Li et al. (Li et al., 2018)	0.53	0.97	0.76	0.83	0.59	0.29	0.50	0.51	0.623
	Ouyang et al. (Ouyang et al., 2021)	0.54	**1.00**	0.75	0.79	0.67	0.53	0.86	0.60	0.720
	CLN-ResNet50V2 (ours)	0.68	0.93	0.63	0.92	0.65	**0.94**	0.89	0.72	0.794
	CLN-MobileNetV3Small (ours)	0.81	0.97	0.75	0.84	0.56	0.57	0.96	**0.95**	0.800
	CLN-EfficientNetB4 (ours)	0.81	0.97	0.73	**1.00**	0.56	0.88	0.92	0.74	**0.826**
	CLN-InceptionV3 (ours)	0.78	**1.00**	0.68	0.82	**1.00**	0.77	0.77	0.62	0.805
	CLN-Xception (ours)	**0.87**	**1.00**	0.73	0.83	0.67	0.65	0.88	0.81	0.806
	CLN-CheXNet (ours)	0.79	0.97	**0.80**	0.84	0.82	0.69	**1.00**	0.70	**0.826**
0.3	ViT-based (Han et al., 2023)	0.20	0.45	0.19	0.32	0.06	0.00	0.21	0.02	0.181
	Wang et al. (Wang et al., 2017)	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.221
	RGT (Han et al., 2023)	0.28	0.79	0.22	0.38	0.12	0.01	0.41	0.05	0.283
	Li et al. (Li et al., 2018)	0.36	0.94	0.56	0.66	0.45	0.17	0.39	0.44	0.496
	Ouyang et al. (Ouyang et al., 2021)	0.40	**1.00**	0.52	0.68	0.58	0.46	0.69	0.43	0.600
	CLN-ResNet50V2 (ours)	0.60	0.82	0.57	0.92	0.65	**0.88**	0.89	0.72	0.755
	CLN-MobileNetV3Small (ours)	0.76	0.97	0.69	0.72	0.56	0.50	0.82	**0.95**	0.744
	CLN-EfficientNetB4 (ours)	0.76	0.97	0.73	**1.00**	0.56	**0.88**	0.77	0.74	**0.800**
	CLN-InceptionV3 (ours)	0.78	0.97	0.68	0.82	**0.95**	0.65	0.77	0.57	0.773
	CLN-Xception (ours)	**0.83**	**1.00**	0.70	0.83	0.67	0.35	0.76	0.81	0.744
	CLN-CheXNet (ours)	0.70	0.97	**0.77**	0.80	0.77	0.69	**1.00**	0.65	0.792
0.4	ViT-based (Han et al., 2023)	0.10	0.21	0.03	0.05	0.02	0.00	0.04	0.00	0.056
	Wang et al. (Wang et al., 2017)	0.09	0.28	0.20	0.12	0.07	0.01	0.08	0.07	0.115
	RGT (Han et al., 2023)	0.17	0.54	0.13	0.18	0.07	0.01	0.26	0.02	0.173
	Li et al. (Li et al., 2018)	0.25	0.88	0.37	0.50	0.33	0.11	0.26	0.29	0.374
	Ouyang et al. (Ouyang et al., 2021)	0.26	**1.00**	0.29	0.56	0.40	0.35	0.50	0.32	0.460
	CLN-ResNet50V2 (ours)	0.61	0.82	0.50	**0.92**	0.65	**0.75**	0.89	0.72	0.731
	CLN-MobileNetV3Small (ours)	0.76	0.97	0.63	0.72	0.56	0.43	0.82	**0.95**	0.728
	CLN-EfficientNetB4 (ours)	0.76	0.97	0.60	0.91	0.56	0.63	0.73	0.68	0.730
	CLN-InceptionV3 (ours)	**0.78**	0.90	0.54	0.77	**0.90**	0.59	0.77	0.48	0.716
	CLN-Xception (ours)	0.74	**1.00**	**0.67**	0.83	0.53	0.29	0.76	0.71	0.693
	CLN-CheXNet (ours)	0.70	0.97	**0.67**	0.76	0.71	0.63	**1.00**	0.60	**0.753**
0.5	ViT-based (Han et al., 2023)	0.05	0.15	0.01	0.04	0.02	0.00	0.03	0.00	0.034
	Wang et al. (Wang et al., 2017)	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.061
	RGT (Han et al., 2023)	0.08	0.32	0.05	0.09	0.05	0.00	0.12	0.01	0.090
	Li et al. (Li et al., 2018)	0.14	0.84	0.22	0.30	0.22	0.07	0.17	0.19	0.269
	Ouyang et al. (Ouyang et al., 2021)	0.15	0.99	0.14	0.33	0.27	0.22	0.35	0.22	0.330
	CLN-ResNet50V2 (ours)	0.58	0.74	0.50	**0.92**	0.59	**0.69**	0.89	0.72	0.703
	CLN-MobileNetV3Small (ours)	0.70	0.97	**0.63**	0.64	0.44	0.21	0.82	**0.90**	0.664
	CLN-EfficientNetB4 (ours)	0.70	0.97	0.47	0.82	0.50	0.50	0.73	0.68	0.671
	CLN-InceptionV3 (ours)	**0.76**	0.84	0.39	0.77	**0.90**	0.41	0.77	0.48	0.665
	CLN-Xception (ours)	0.71	**1.00**	0.58	0.75	0.47	0.18	0.72	0.67	0.634
	CLN-CheXNet (ours)	0.70	0.91	0.60	0.72	0.71	0.63	**0.96**	0.55	**0.720**
0.6	ViT-based (Han et al., 2023)	0.01	0.03	0.01	0.01	0.01	0.00	0.01	0.00	0.010
	Wang et al. (Wang et al., 2017)	0.02	0.08	0.05	0.02	0.00	0.01	0.02	0.03	0.029
	RGT (Han et al., 2023)	0.02	0.15	0.03	0.04	0.03	0.00	0.06	0.00	0.041
	Li et al. (Li et al., 2018)	0.07	0.73	0.15	0.18	0.16	0.03	0.10	0.12	0.193
	Ouyang et al. (Ouyang et al., 2021)	0.08	0.97	0.05	0.18	0.14	0.15	0.27	0.11	0.240
	CLN-ResNet50V2 (ours)	0.58	0.74	0.47	**0.88**	0.53	0.31	0.89	0.61	**0.626**
	CLN-MobileNetV3Small (ours)	0.60	0.90	**0.59**	0.56	0.33	0.07	0.82	**0.80**	0.584
	CLN-EfficientNetB4 (ours)	0.60	0.90	0.33	0.77	0.25	**0.44**	0.73	0.53	0.569
	CLN-InceptionV3 (ours)	**0.65**	0.84	0.29	0.64	**0.79**	0.35	0.68	0.43	0.583
	CLN-Xception (ours)	0.43	**1.00**	0.39	0.63	0.27	0.00	0.60	0.48	0.474
	CLN-CheXNet (ours)	0.52	0.81	0.50	0.72	0.53	0.31	**0.92**	0.40	0.588
0.7	ViT-based (Han et al., 2023)	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.001
	Wang et al. (Wang et al., 2017)	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.011
	RGT (Han et al., 2023)	0.01	0.04	0.01	0.02	0.01	0.00	0.03	0.00	0.015
	Li et al. (Li et al., 2018)	0.04	0.52	0.07	0.09	0.11	0.01	0.05	0.05	0.118
	Ouyang et al. (Ouyang et al., 2021)	0.02	0.77	0.01	0.12	0.08	0.10	0.06	0.03	0.150
	CLN-ResNet50V2 (ours)	0.47	0.74	0.43	**0.80**	0.29	0.13	**0.85**	0.44	**0.520**
	CLN-MobileNetV3Small (ours)	**0.51**	**0.90**	**0.56**	0.48	0.33	0.00	0.82	**0.70**	0.538
	CLN-EfficientNetB4 (ours)	0.46	0.77	0.17	0.73	0.19	0.19	0.69	0.26	0.432
	CLN-InceptionV3 (ours)	0.35	0.84	0.25	0.64	**0.42**	**0.24**	0.64	0.33	0.463
	CLN-Xception (ours)	0.11	0.85	0.21	0.33	0.13	0.00	0.28	0.38	0.288
	CLN-CheXNet (ours)	0.33	0.81	0.43	0.60	0.24	0.13	0.71	0.35	0.450

*λ set to 0.1 for CLN-ResNet50V2 and CLN-EfficientNetB4, to 0.2 for CLN-Xception, and to 0.4 for CLN-InceptionV3, CLN-MobileNetV3Small, and CLN-CheXNet.

Table 2: AUC scores for the annotated subset of the NIH dataset for the task of CXR image classification. Results ordered by ascending mean AUC. Methods in bold are also included in the localisation performance comparison.

Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Mean
Wang et al. (Wang et al., 2017)	0.72	0.81	0.78	0.61	0.71	0.67	0.63	0.81	0.718
ViT-based (Han et al., 2023)	0.74	0.78	0.81	0.72	0.70	0.66	0.65	0.76	0.728
CrossViT (Han et al., 2023)	0.69	0.71	0.72	0.72	0.74	0.79	0.82	0.88	0.759
PS-ViT (Han et al., 2023)	0.75	0.81	0.82	0.73	0.79	0.73	0.69	0.81	0.766
DNetLoc (Gündel et al., 2018)	0.77	0.88	0.83	0.71	0.82	0.76	0.73	0.85	0.794
CAN (Ma et al., 2019)	0.78	0.89	0.83	0.70	0.84	0.77	0.72	0.86	0.799
Li et al. (Li et al., 2018)[a]	0.80	0.87	0.87	0.70	0.83	0.77	0.67	0.88	0.799
Ouyang et al. (Ouyang et al., 2021)	0.77	0.87	0.83	0.71	0.83	0.79	0.72	0.88	0.800
Liu et al. (Liu et al., 2019)	0.79	0.87	**0.88**	0.69	0.81	0.73	0.75	0.89	0.801
Seyyed-Kalantari et al. (Seyyed-Kalantari et al., 2021)	0.81	0.92	0.87	0.72	0.83	0.78	0.76	0.88	0.821
CLN-InceptionV3 (ours)	0.83	0.96	0.70	0.71	0.86	0.92	0.78	0.84	0.825
Han et al. (Han et al., 2021)	0.83	0.92	0.87	0.76	0.85	0.76	0.77	0.86	0.828
Rajpurkar et al. (Rajpurkar et al., 2017)	0.82	0.91	**0.88**	0.72	0.86	0.78	0.76	0.89	0.828
CLN-EfficientNetB4 (ours)	0.77	0.94	0.79	0.83	0.73	0.94	0.77	0.87	0.831
RGT (Han et al., 2023)	0.80	0.92	0.78	0.86	0.88	0.88	0.79	0.81	0.839
CLN-CheXNet (ours)	0.75	0.96	0.84	0.74	**0.90**	0.86	0.78	0.97	0.849
CLN-Xception	0.80	**0.99**	0.86	0.84	0.85	0.90	0.83	0.95	0.878
CLN-ResNet50V2 (ours)	0.83	0.90	0.81	**0.92**	**0.90**	**0.95**	0.91	0.92	0.893
CLN-MobileNetV3Small	**0.89**	**0.99**	**0.88**	0.89	0.85	0.93	**0.93**	**0.98**	**0.918**

[a]Results refer to the 80% annotated - 80% unannotated setting

*λ set to 0.1 for CLN-ResNet50V2 and CLN-EfficientNetB4, to 0.2 for CLN-Xception, and to 0.4 for CLN-InceptionV3, CLN-MobileNetV3Small and CLN-CheXNet.

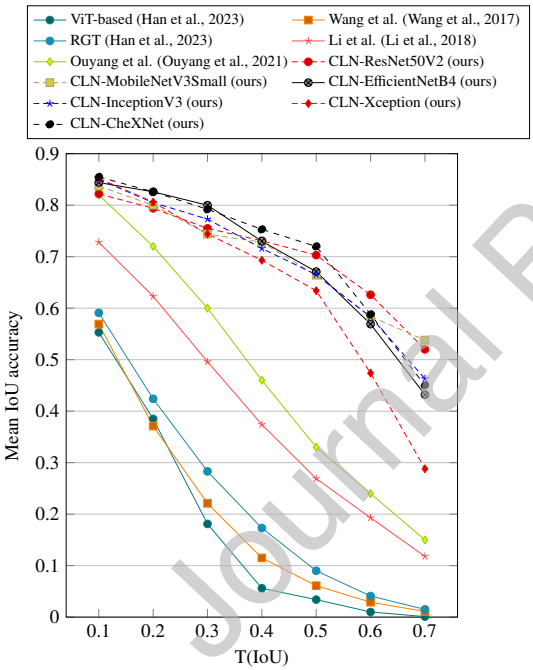

Figure 2: Pathology localisation results for the examined IoU thresholds.

3.2. Pathology Localisation

The six variants of the proposed CLN architecture were evaluated in terms of IoU accuracy against five state-of-the-art CXR image localisation methods, i.e. ViT-based (Han et al., 2023), Wang et al. (Wang et al., 2017), RGT (Han et al., 2023), Li et al. (Li et al., 2018), and Ouyang et al. (Ouyang et al., 2021). Localisation results are summarised in Table 1 for each of the 8 pathologies included in the annotated subset of the NIH data set. The mean IoU accuracy across all pathologies is also reported for all the examined methods and is also illustrated in Figure 2. From Table 1 and Figure 2, it is evident that all the six examined variants of the proposed CLN architecture outperform the compared state-of-the-art localisation methods for all the examined IoU thresholds, achieving a maximum mean IoU accuracy of 0.855 for $T(IoU) = 0.1$ using the CLN-ChexXNet variant. In comparison, the best state-of-the-art method (Ouyang et al. (Ouyang et al., 2021)) achieved a mean IoU accuracy of 0.820 for the same threshold. At $T(IoU) = 0.1$, the variants of the proposed architecture achieved an improvement in the localisation performance of individual pathologies for 7 out of the 8 pathologies, performing worse only for effusion, for which (Ouyang et al., 2021) achieved an IoU accuracy of 0.89 compared to 0.83 for the proposed method. For cardiomegaly, both the proposed method and (Ouyang et al., 2021) achieved perfect localisation performance, with IoU accuracy reaching 1.00.

The improvement in localisation performance of the proposed method compared to the state of the art becomes more substantial for higher IoU thresholds, as shown in Figure 2 and Table 1. At $T(IoU) = 0.2$, the highest mean IoU accuracy is improved by 14.7% compared to (Ouyang et al., 2021) (0.826 vs. 0.720), achieving an improvement for 7 out of the 8 pathologies, whereas both (Ouyang et al., 2021) and the proposed CLN method achieved a perfect localisation performance for cardiomegaly. At $T(IoU) = 0.3$, the improvement is 33.3% (0.800 for CLN-EfficientNetB4 vs. 0.600 for (Ouyang et al., 2021)), outperforming the other methods for all pathologies, whereas both the proposed method and (Ouyang et al., 2021) achieved perfect localisation performance for cardiomegaly (1.00). At $T(IoU) = 0.4$ the improvement is 63.7% (0.753 vs. 0.460), performing considerably better for all pathologies, whereas both the proposed method and (Ouyang et al., 2021) achieved perfect localisation performance for cardiomegaly (1.00). The improvement at $T(IoU) = 0.5$ and $T(IoU) = 0.6$ rises to 118.2% (0.720 vs. 0.330) and 160.8% (0.626 vs. 0.240), respectively, achieving better localisation for

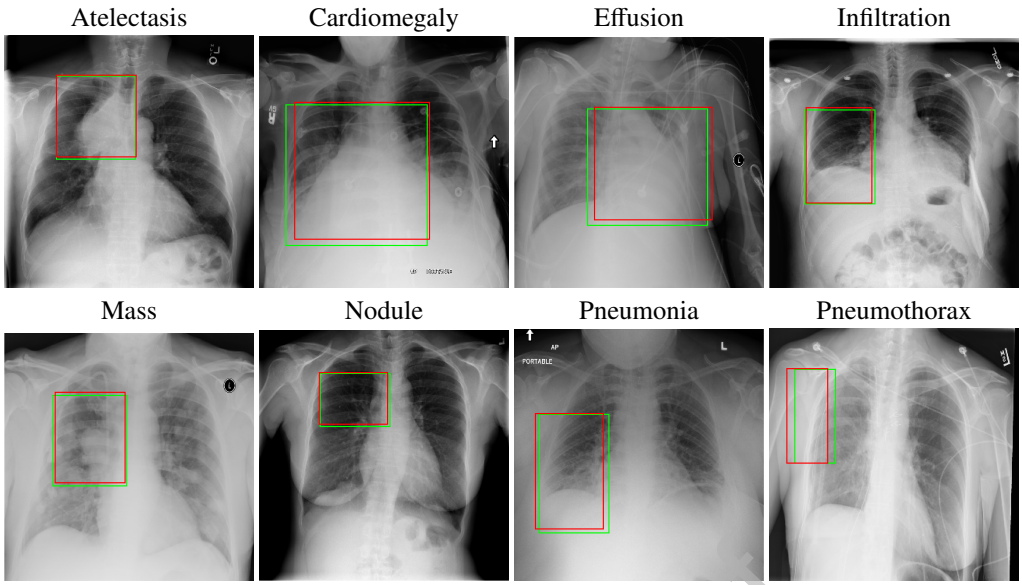

Figure 3: Visualisation of pathology localisation for a randomly selected CXR image for each pathology. Green and red denote the ground-truth and the predicted bounding box, respectively, using the proposed CLN-CheXNet model.

all pathologies. Finally, the proposed CLN method achieved a remarkable improvement of 246.7% (0.520 vs. 0.150) over the state of the art for $T(IoU) = 0.7$, achieving better localisation for all the examined pathologies.

It must be noted that for the reported results, CLN-ResNet50V2 and CLN-EfficientNetB4 were trained with $\lambda = 0.1$, CLN-Xception with $\lambda = 0.2$, whereas CLN-InceptionV3, CLN-MobileNetV3Small, and CLN-CheXNet were trained with $\lambda = 0.4$. Models were trained and fine-tuned using the training and validation subsets of the bounding box annotated CXR images from the NIH data set, and results are reported for the test subset. Examples of the predicted bounding boxes for each pathology are shown in Figure 3.

A Wilcoxon signed-rank test was used to test for significance by comparing the distribution of the IoU accuracy scores achieved for all pathologies and for all the IoU thresholds ($T(IoU)$) by our best performing CLN-CheXNet model against each of the other methods outlined in Table 1. The distribution of IoU accuracy scores was significantly different than the four compared state-of-the-art methods, with $p << 0.01$. However, a statistically significant difference could not be established against the other examined variants of our proposed method ($p > 0.05$).

3.3. Pathology Classification

We evaluated the performance of the proposed method for pathology classification of the 8 pathologies in the annotated subset of the NIH data set by comparing to state-of-the-art methods for CXR image classification. In the experimental comparison, we included the five CXR localisation methods examined in the previous section, as well as seven additional CXR image classification methods (CrossViT (Han et al., 2023), PS-ViT (Han et al., 2023), DNetLoc (Gündel et al., 2018), CAN (Ma et al., 2019), Liu et al. (Liu et al., 2019), Han et al. (Han et al., 2021), Rajpurkar et al. (Rajpurkar et al., 2017)). Classification performance is reported in Table 2 in terms of the AUC score for each individual pathology, as well in terms of the mean AUC across the 8 pathologies. From this table, it is evident that the CLN-MobileNetV3Small variant of the proposed architecture outperformed all the compared methods in terms of mean AUC, achieving a mean AUC of 0.918. The second best mean AUC (0.893) was achieved by the proposed CLN-ResNet50V2, whereas the third best (0.878) by the proposed CLN-Xception. From Table 2, it is evident that, compared to the state of the art, the variants of the proposed architecture were able to achieve improved AUC for all the examined pathologies. For effusion, the best AUC of 0.88 was achieved by the proposed CLN-MobileNetV3Small, the Liu et al. (Liu et al., 2019) and the Rajpurkar et al. (Rajpurkar et al., 2017) methods.

A Wilcoxon signed-rank test was used to test for significance by comparing the distribution of the AUC scores achieved for all pathologies by our best performing CLN-MobileNetV3Small model against each of the other methods outlined in Table 2. The distribution of AUC scores was significantly different than the compared state-of-the-art methods, with $p < 0.05$ for the Liu et al. (Liu et al., 2019), Han et al. (Han et al., 2021), Rajpurkar et al. (Rajpurkar et al., 2017), and RGT (Han et al., 2023) methods, and $p < 0.01$ for the rest. When compared against the other variants of the proposed method, the distribution of AUC scores was significantly different than CLN-InceptionV3, CLN-EfficientNetB4, and CLN-Xception ($p < 0.05$), but a statistically significant difference could not be established against CLN-CheXNet and CLN-ResNet50V2 ($p > 0.05$).

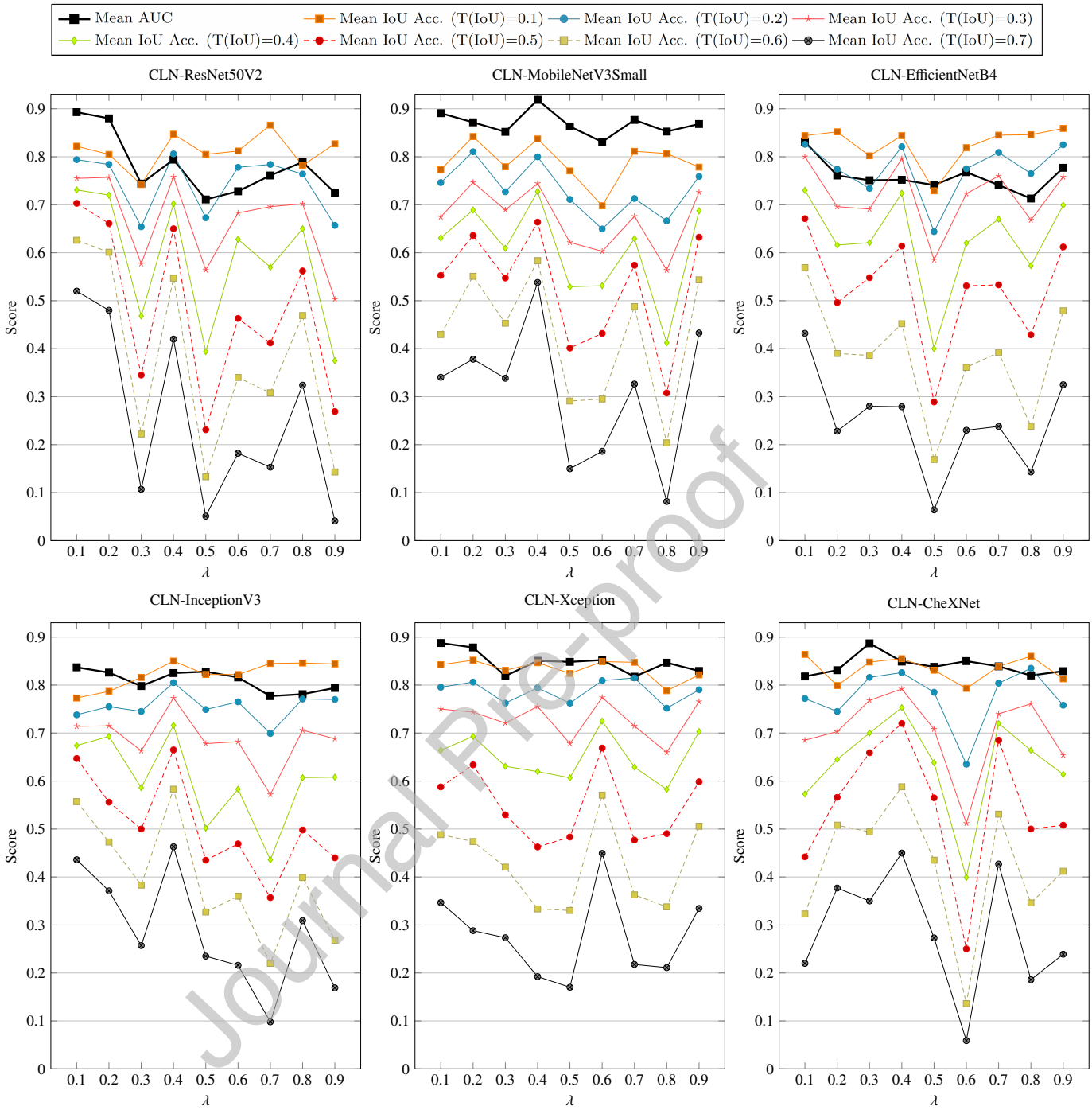

Figure 4: Pathology localisation and classification results for the six examined variants of the proposed architecture in relation to the λ hyperparameter. Values in the y axis depict the mean IoU accuracy or the mean AUC score depending on the respective series.

3.4. Ablation study

It is evident from the experimental results that the performance of the proposed CLN architecture can be affected by the model used as the image feature extraction backbone, by the value of the hyperparameter λ, and by the IoU threshold used to compute the IoU accuracy. We evaluated the effect of the IoU threshold and of λ by conducting various ablation experiments, whereas the effect of the six different models used as the image feature extraction backbone was extensively discussed in Section 3.2 and can be seen in Table 1 and Table 2. In addition, we compared the localisation performance of the proposed approach to the localisation performance achieved by using Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) on the CNN-based image feature extraction backbone of the proposed architecture.

3.4.1. Effect of IoU threshold $T(IoU)$

The effect of the IoU threshold on the mean IoU accuracy is examined in Figure 2. There is a clear negative relation between $T(IoU)$ and mean IoU accuracy for all the examined methods, with higher $T(IoU)$ leading to lower mean IoU accuracy in all cases. Nevertheless, it is evident that the proposed CLN architecture is more resistant to the increase in $T(IoU)$, exhibiting much slower decay in mean IoU accuracy compared to the state-of-the-art methods, which demonstrated rapid decay as $T(IoU)$ increased, as shown in Figure 2.

3.4.2. Effect of λ

The effect of the λ hyperparameter on the classification mean AUC score and on the localisation mean IoU accuracy is examined in Figure 4 for $\lambda \in [0.1, 0.9]$. Given the proposed loss function (Eq. 4), a higher λ implies a higher weight to the cross-entropy loss and a lower weight to the IoU loss, i.e. a higher weight for the classification metric and lower for the localisation metric. A $\lambda = 0$ or $\lambda = 1$ would lead to only the classification or the localisation metric to be considered, respectively. λ's effect on mean AUC score seems to depend on the model used as the image feature extraction backbone. Mean AUC scores for CLN-InceptionV3, CLN-CheXNet, CLN-Xception, and CLN-MobileNetV3Small are less affected by λ, with scores ranging from 0.777 to 0.837 ($\sigma = 0.022$), from 0.818 to 0.887 ($\sigma = 0.021$), from 0.817 to 0.888 ($\sigma = 0.024$), and from 0.831 to 0.919 ($\sigma = 0.025$), respectively, whereas CLN-EfficientNetB4 and CLN-ResNet50V2 are much more affected, with scores ranging from 0.713 to 0.831 ($\sigma = 0.033$) and from 0.711 to 0.893 ($\sigma = 0.066$), respectively.

From Figure 4, it is evident that λ's effect on mean IoU accuracy is much more dramatic compared to AUC. λ values at the mid of its range seem to underperform considerably. In most cases, localisation performance seems to drop as λ increases and then improves again for higher values. The observed trend for each variant is consistent across the different IoU thresholds ($T(IoU)$), but the drop in localisation performance for mid-valued λ is less prominent for $T(IoU) = 0.1$.

Considering that the proposed architecture aims at both pathology classification and localisation, the selected λ should maximise both mean AUC and mean IoU accuracy. However, as shown in Figure 4, the best mean AUC is typically achieved for a different λ than the one providing the best mean IoU accuracy, thus a trade-off must be made during hyperparameter tuning. In this work, we attempted to achieve a balance between the two metrics, leading to $\lambda = 0.1$ being selected for CLN-ResNet50V2 and CLN-EfficientNetB4, $\lambda = 0.2$ for CLN-Xception, and $\lambda = 0.4$ for CLN-InceptionV3, CLN-MobileNetV3Small, and CLN-CheXNet.

3.4.3. Comparison to Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) is a localisation method that can generate class-wise visual explanations of the classification predictions of CNN-based networks. For each possible class, Grad-CAM uses the gradients flowing into the last convolutional layer of the network to create a heatmap that indicates the image regions that were most important for predicting the respective class. Localisation bounding boxes can then be extracted from the heatmap after applying thresholding and drawing the bounding box around the largest contiguous group of selected pixels (Gascoigne-Burns and Katsigiannis, 2022). Considering that our proposed CLN architecture relies on a CNN-based backbone, localisation bounding boxes can also be computed by applying Grad-CAM on the last convolutional layer of the CNN backbone.

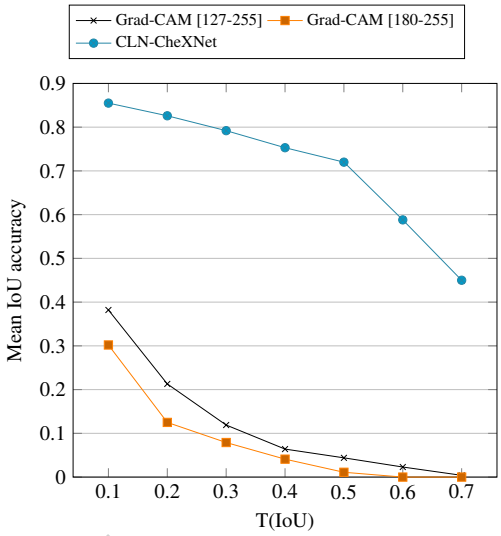

Figure 5: Localisation results of CLN-ChexNet and Grad-CAM in terms of mean IoU accuracy for the examined IoU thresholds ($T(IoU)$).

To demonstrate the effectiveness of our proposed approach, we compared the best performing CLN variant's (CLN-ChexNet) localisation performance against Grad-CAM, applied to the CNN-based image feature extraction backbone of the model. Given that the selected threshold affects localisation performance (Gascoigne-Burns and Katsigiannis, 2022), results for Grad-CAM were computed for two different thresholds, i.e. accepting pixel values within the range $[127, 255]$ and $[180, 255]$, respectively. The thresholds refer to the Grad-CAM heatmap's pixels intensity, after normalising them to the range $[0, 255]$. Localisation performance for CLN-CheXNet and the two Grad-CAM thresholds are illustrated in Figure 5. From this figure, it is evident that the proposed approach performs considerably better than Grad-CAM applied on its CNN-based backbone, with CLN-CheXNet achieving a maximum mean IoU accuracy of 0.855 for $T(IoU) = 0.1$, compared to 0.382 for Grad-CAM with the $[127, 255]$ threshold and 0.302 for Grad-CAM with the $[180, 255]$ threshold. Examples of bounding boxes created by Grad-CAM and by CLN-CheXNet in relation to the ground truth bounding boxes are shown in Figure 6.

4. Discussion

It is evident from Table 1 and 2 that the proposed CLN architecture outperforms the examined state-of-the-art models for both CXR image localisation and classification. Among the six

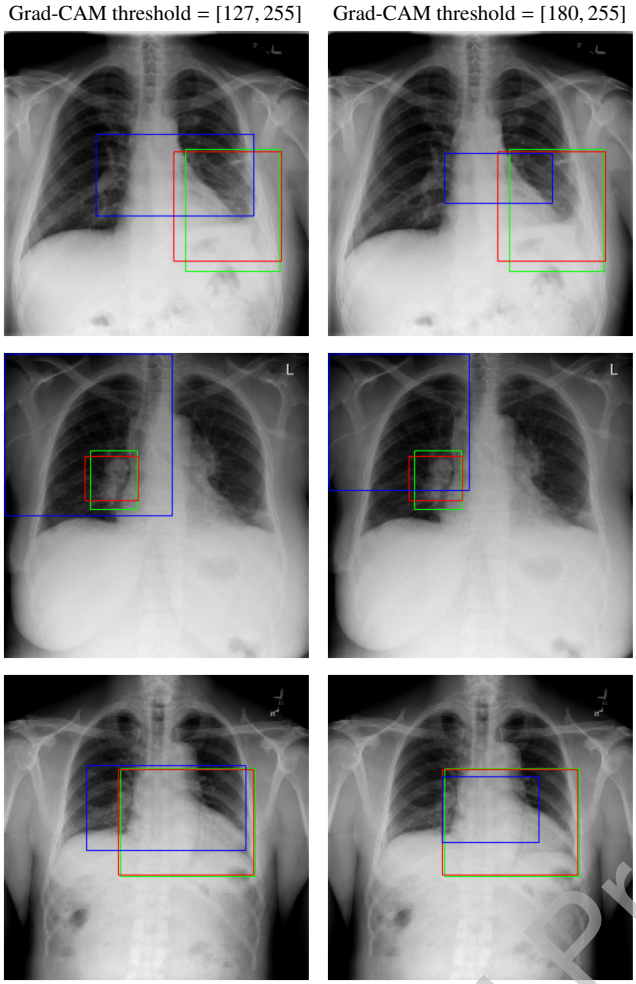

Figure 6: Visualisation of pathology localisation for randomly selected CXR images. Green denotes the ground-truth bounding box. Red denotes the predicted bounding box using the proposed CLN-CheXNet model. Blue denotes the predicted bounding box using Grad-CAM for a threshold of [127, 255] (left column) and a threshold of [180, 255] (right column).

examined CLN variants, CLN-CheXNet achieved the best balance among localisation and classification, ranking first among all models for localisation and fourth behind other CLN variants for classification. Apart from the improved performance, the proposed approach offers two significant advantages over the compared methods; better stability in localisation performance as the IoU threshold increases, and relative simplicity that leads to lower computational complexity.

Regarding the stability in localisation performance as the IoU threshold increases, it is evident from Figure 2 that mean IoU accuracy for the examined state-of-the-art localisation methods drops sharply with the increase in $T(IoU)$, whereas the accuracy for the variants of the proposed architecture drops at a much slower rate. The stability of the proposed approach with regards to the IoU threshold is further demonstrated in Figure 7, which depicts the decrease in mean IoU accuracy as a percentage of the mean IoU accuracy for $T(IoU) = 0.1$ for the proposed CLN variants and the compared CXR localisation methods. The best performing state-of-the-art Ouyang et

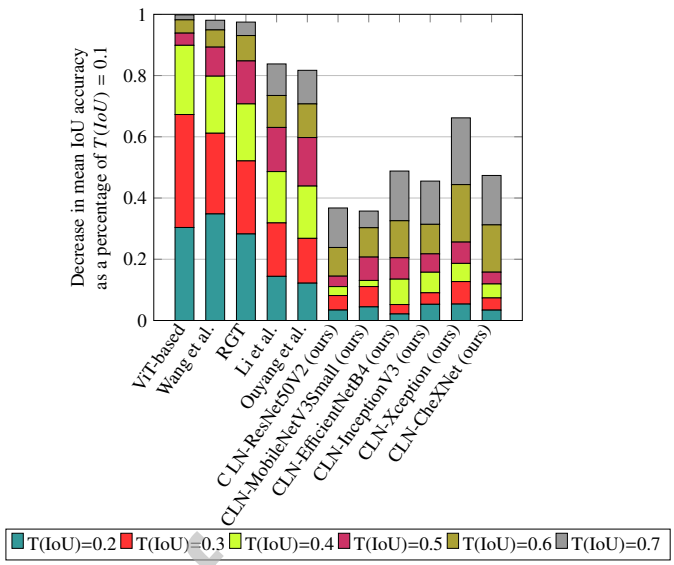

Figure 7: Decrease in mean IoU accuracy as a percentage of the mean IoU accuracy for $T(IoU) = 0.1$ as the IoU threshold increases for the proposed CLN variants and the compared CXR localisation methods.

al. (Ouyang et al., 2021) method exhibited a 81.7% decrease in mean IoU between $T(IoU) = 0.1$ and $T(IoU) = 0.7$, whereas the worst performing ViT-based (Han et al., 2023) method exhibited a decrease of 99.8%. On the contrary, the best performing variant of the proposed architecture (CLN-CheXNet) exhibited a decrease of only 48.8% in mean IoU accuracy between $T(IoU) = 0.1$ and $T(IoU) = 0.7$, whereas the worst performing variant (CLN-ResNet50V2) exhibited a decrease of only 36.7%, demonstrating better stability in localisation performance compared to the examined state-of-the-art models as the IoU threshold increases.

Regarding its complexity, the proposed architecture relies on a pre-trained CNN model and fully-connected layers for classification and for localisation via regression. The compared state-of-the-art approaches employ combinations of transformers, attention mechanisms, and complex network structures, leading to a substantial increase in complexity. The size of each examined CLN variant in terms of its number of parameters is presented in Table 3, alongside its computational cost in terms of floating point operations per second (FLOPS). From Table 3, it is evident that the size and complexity of the proposed method depends mainly on the size and complexity of the image feature extraction backbone. Smaller backbones (e.g. MobileNetV3Small) lead to smaller CLNmodels and vice versa.

We attribute the performance improvements of the proposed method to the following reasons: By utilising localisation annotations during training in the form of bounding boxes, the model learns to directly predict the location of abnormalities, which is more precise than methods that rely purely on classification labels, due to increased spatial awareness. In addition, the use of the localisation bounding boxes in the training also allows the loss function to incorporate a localisation quality metric, i.e. IoU, thus forcing the training process to optimise the proposed method for both localisation and classification. Finally, follow-

Table 3: Size and floating point operations per second (FLOPS) for the examined CLN variants.

Model	Total Parameters	Trainable Parameters	FLOPS
CLN-ResNet50V2	24,795,020	24,749,580	6.99 GFLOPS
CLN-MobileNetV3Small	1,403,900	1,391,788	0.12 GFLOPS
CLN-EfficientNetB4	18,770,916	18,645,716	3.08 GFLOPS
CLN-InceptionV3	23,033,004	22,998,572	5.69 GFLOPS
CLN-Xception	22,091,700	22,037,172	9.13 GFLOPS
CLN-CheXNet	7,735,244	7,651,596	5.70 GFLOPS

ing a multi-task approach allowed each branch of the proposed architecture (localisation and classification) to be optimised for each intended task, thus having a dedicated part of the architecture for predicting localisation bounding boxes, as opposed to various state-of-the-art methods that approach localisation as a by-product of the classification task, e.g. using Grad-CAM, attention maps, etc.

However, the proposed architecture has some limitations. Its training requires CXR images annotated with bounding boxes from expert radiologists. Annotating large numbers of CXR images is an arduous, time-consuming, and costly task, leading to limited availability of such annotated images for many pathologies. Furthermore, to the best of the authors' knowledge, there is no other CXR data set that contains bounding box annotations for the pathologies examined in this work, thus a cross-dataset evaluation is not possible. In addition, the proposed architecture does not support the prediction of multiple bounding boxes on the CXR image, thus being limited in providing localisation for a single pathology. Finally, a trade-off between classification and localisation performance must be made during hyperparameter tuning, as the ablation study showed that the best performance is achieved for different λ for each task.

5. Conclusion

In this work, we proposed the Chest X-ray Localisation Network (CLN), a multi-task deep neural network designed for the tasks of CXR image classification and localisation. Built on a pre-trained convolutional neural network backbone, the CLN architecture uses separate branches for classification and bounding box regression (localisation), achieving a mean AUC of 0.918 in classification and a mean IoU accuracy of 0.855 in localisation across eight pathologies in a publicly available annotated CXR dataset. Our proposed method provides notable advantages over existing methods, including superior classification and localisation accuracy, reduced performance decay with increased IoU thresholds, and an overall simpler architecture. The results indicate that CLN can offer an efficient solution for computer-aided CXR diagnosis, offering interpretable and effective support for radiologists. Nevertheless, the proposed method also has some limitations. In terms of practical considerations, the model's reliance on high-quality annotated data sets may limit its scalability to other imaging datasets lacking similar annotations. In addition, a trade-off between classification and localisation performance is required during hyperparameter tuning. Furthermore, the proposed architecture does not support the prediction of multiple bounding boxes on a CXR image. In terms of theoretical considerations, while CLN demonstrates interpretability, the approach could benefit from incorporating more advanced explainability techniques to increase trust in clinical applications. Future work will address these limitations by incorporating additional pathologies, supporting multiple bounding boxes per image, and further enhancing interpretability to align more closely with radiologists' needs.

Acknowledgement

The authors would like to thank all the researchers that have made their data sets publicly available, making this work possible. Furthermore, the authors declare no competing interests. No funding was associated with this work. No ethical approval was required for this work, as only publicly available data were used.

ORCID Information

Gabriel Iluebe Okolo: 0000-0002-1624-6668 Stamos Katsigiannis: 0000-0001-9190-0941 Naeem Ramzan: 0000-0002-5088-1462

CrediT author statement

Gabriel Iluebe Okolo: Conceptualization, Methodology, Software, Validation, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization

Stamos Katsigiannis: Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision

Naeem Ramzan: Conceptualization, Methodology, Validation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Declaration of interests

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.

References

Annarumma, M., Withey, S.J., Bakewell, R.J., Pesce, E., Goh, V., Montana, G., 2019. Automated triaging of adult chest radiographs with deep artificial neural networks. Radiology 291, 196–202. doi:10.1148/radiol.2018180921.

Bar, Y., Diamant, I., Wolf, L., Greenspan, H., 2015. Deep learning with non-medical training used for chest pathology identification, in: Medical Imaging 2015: Computer-Aided Diagnosis, SPIE. pp. 215–221. doi:10.1117/12.2083124.

Blais, M.A., Akhloufi, M.A., 2021. Deep learning and binary relevance classification of multiple diseases using chest x-ray images, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE. pp. 2794–2797. doi:10.1109/EMBC46164.2021.9629846.

Brunese, L., Mercaldo, F., Reginelli, A., Santone, A., 2020. Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays. Computer Methods and Programs in Biomedicine 196, 105608. doi:10.1016/j.cmpb.2020.105608.

Chandra, T.B., Singh, B.K., Jain, D., 2022. Disease localization and severity assessment in chest x-ray images using multi-stage superpixels classification. Computer Methods and Programs in Biomedicine 222, 106947. doi:10.1016/j.cmpb.2022.106947.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807. doi:10.1109/CVPR.2017.195.

Cicero, M., Bilbily, A., Colak, E., Dowdell, T., Gray, B., Perampaladas, K., Barfett, J., 2017. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. Investigative radiology 52, 281–287. doi:10.1097/RLI.0000000000000341.

Dhoot, R., Humphrey, J.M., O'Meara, P., Gardner, A., McDonald, C.J., Ogot, K., Antani, S., Abuya, J., Kohli, M., 2018. Implementing a mobile diagnostic unit to increase access to imaging and laboratory services in western kenya. BMJ Global Health 3, e000947. doi:10.1136/bmjgh-2018-000947.

Doi, K., 2007. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics 31, 198–211. doi:https://doi.org/10.1016/j.compmedimag.2007.02.002.

Gascoigne-Burns, J., Katsigiannis, S., 2022. A localisation study of deep learning models for chest x-ray image classification, in: Proc. IEEE-EMBS BHI. doi:10.1109/BHI56158.2022.9926904.

Gündel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A., Comaniciu, D., 2018. Learning to recognize abnormalities in chest x-rays with location-aware dense networks, in: Vera-Rodriguez, R., Fierrez, J., Morales, A. (Eds.), Proc. CIARP, pp. 757–765. doi:10.1007/978-3-030-13469-3_88.

Han, Y., Chen, C., Tang, L., Lin, M., Jaiswal, A., Wang, S., Tewfik, A., Shih, G., Ding, Y., Peng, Y., 2021. Using Radiomics as Prior Knowledge for Thorax Disease Classification and Localization in Chest X-rays. AMIA Annu Symp Proc 2021, 546–555.

Han, Y., Holste, G., Ding, Y., Tewfik, A., Peng, Y., Wang, Z., 2023. Radiomics-guided global-local transformer for weakly supervised pathology localization in chest x-rays. IEEE Transactions on Medical Imaging 42, 750–761. doi:10.1109/TMI.2022.3217218.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proc. IEEE CVPR, pp. 770–778. doi:10.1109/CVPR.2016.90.

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., Le, Q., 2019. Searching for mobilenetv3, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1314–1324. doi:10.1109/ICCV.2019.00140.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al., 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: Proc. AAAI, pp. 590–597. doi:10.1609/aaai.v33i01.3301590.

Kelly, B.S., Rainford, L.A., Darcy, S.P., Kavanagh, E.C., Toomey, R.J., 2016. The development of expertise in radiology: in chest radiograph interpretation,"expert" search pattern may predate "expert" levels of diagnostic accuracy for pneumothorax identification. Radiology 280, 252–260. doi:10.1148/radiol.2016150409.

Ker, J., Wang, L., Rao, J., Lim, T., 2017. Deep learning applications in medical image analysis. IEEE Access 6, 9375–9389. doi:10.1109/ACCESS.2017.2788044.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 84–90. doi:10.1145/3065386.

Kumar, P., Grewal, M., Srivastava, M.M., 2018. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs, in: Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15, Springer. pp. 546–552. doi:10.1007/978-3-319-93000-8_62.

Lakhani, P., Sundaram, B., 2017. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 284, 574–582. doi:10.1148/radiol.2017162326.

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.J., Fei-Fei, L., 2018. Thoracic disease identification and localization with limited supervision, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8290–8299. doi:10.1109/CVPR.2018.00865.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60–88. doi:10.1016/j.media.2017.07.005.

Liu, J., Zhao, G., Fei, Y., Zhang, M., Wang, Y., Yu, Y., 2019. Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision, in: Proc. IEEE/CVF ICCV, pp. 10631–10640. doi:10.1109/ICCV.2019.01073.

Luo, X., Cai, Z., Shao, B., Wang, Y., 2024. Unified-iou: For high-quality object detection. arXiv preprint arXiv:2408.06636 .

Ma, C., Wang, H., Hoi, S.C.H., 2019. Multi-label thoracic disease image classification with cross-attention networks, in: Proc. MICCAI, pp. 730–738. doi:10.1007/978-3-030-32226-7_81.

Majdi, M.S., Salman, K.N., Morris, M.F., Merchant, N.C., Rodriguez, J.J., 2020. Deep learning classification of chest x-ray images, in: 2020 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), IEEE. pp. 116–119. doi:10.1109/SSIAI49293.2020.9094612.

Okolo, G.I., Katsigiannis, S., Ramzan, N., 2022. IEViT: An enhanced vision transformer architecture for chest x-ray image classification. Computer Methods and Programs in Biomedicine 226, 107141. doi:10.1016/j.cmpb.2022.107141.

Ouyang, X., Karanam, S., Wu, Z., Chen, T., Huo, J., Zhou, X.S., Wang, Q., Cheng, J.Z., 2021. Learning hierarchical attention for weakly-supervised chest x-ray abnormality localization and diagnosis. IEEE Transactions on Medical Imaging 40, 2698–2710. doi:10.1109/TMI.2020.3042773.

Pereira, R.M., Bertolini, D., Teixeira, L.O., Silla, C.N., Costa, Y.M., 2020. Covid-19 identification in chest x-ray images on flat and hierarchical classification scenarios. Computer Methods and Programs in Biomedicine 194, 105532. doi:10.1016/j.cmpb.2020.105532.

Qi, B., Zhao, G., Wei, X., Du, C., Pan, C., Yu, Y., Li, J., 2022. Gren: Graph-regularized embedding network for weakly-supervised disease localization in x-ray images. IEEE Journal of Biomedical and Health Informatics 26, 5142–5153. doi:10.1109/JBHI.2022.3193108.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M.P., Ng, A.Y., 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. doi:10.48550/arXiv.1711.05225, arXiv:1711.05225.

Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., Rosenow III, E.C., 2012. Interpretation of plain chest roentgenogram. Chest 141, 545–558. doi:10.1378/chest.10-1302.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proc. IEEE ICCV, pp. 618–626. doi:10.1109/ICCV.2017.74.

Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M., 2021. CheXclusion: Fairness gaps in deep chest X-ray classifiers. pp. 232–243. doi:10.1142/9789811232701_0022.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proc. IEEE CVPR, pp. 2818–2826. doi:10.1109/CVPR.2016.308.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning (ICML), PMLR. pp. 6105–6114.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proc. IEEE CVPR, pp. 3462–3471. doi:10.1109/CVPR.2017.369.

Yu, P., Xu, H., Zhu, Y., Yang, C., Sun, X., Zhao, J., 2011. An automatic computer-aided detection scheme for pneumoconiosis on digital chest radiographs. Journal of digital imaging 24, 382–393. doi:10.1007/s10278-010-9276-7.

Zhang, X., Han, L., Sobeih, T., Han, L., Dempsey, N., Lechareas, S., Tridente, A., Chen, H., White, S., Zhang, D., 2023. Cxr-net: A multitask deep learning network for explainable and accurate diagnosis of covid-19 pneumonia from chest x-ray images. IEEE Journal of Biomedical and Health Informatics 27, 980–991. doi:10.1109/JBHI.2022.3220813.

Zhao, J., Li, M., Shi, W., Miao, Y., Jiang, Z., Ji, B., 2021. A deep

learning method for classification of chest x-ray images, in: Journal of Physics: Conference Series, IOP Publishing. p. 012030. doi:10.1088/1742-6596/1848/1/012030.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D., 2020. Distance-iou loss: Faster and better learning for bounding box regression, in: Proceedings of the AAAI conference on artificial intelligence, pp. 12993–13000.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: Proc. IEEE CVPR, pp. 2921–2929. doi:10.1109/CVPR.2016.319.