

Reducing False Alarms in Automated Chest X-ray Diagnosis: A Two-Stage System that Outperforms YOLOv5 and Faster R-CNN

Amar Debnath
Optimizely Bangladesh
Dhaka-1212, Bangladesh

Sarker T. Ahmed Rumeen
Department of Computer Science and Engineering
University of Dhaka
Dhaka-1000, Bangladesh

Eity Modhu
Data and Design Lab
University of Dhaka
Dhaka-1000, Bangladesh

M. Murshida Mahub
Department of Microbiology
University of Dhaka
Dhaka-1000, Bangladesh

ABSTRACT

Chest X-ray diagnosis is delayed because of the increasing workload faced by radiologists and the intrinsic tendency of deep learning models to produce false positive detections. This paper proposes a lightweight multi-modal deep learning framework to address these challenges. Potential disease regions are identified using a YOLOv5 object detector, and a binary classifier based on an EfficientNet architecture with transfer learning is employed. The main role of the classifier is to remove the detector's false positive predictions. On the VinDr-CXR benchmark (18k chest X-rays, 14 abnormality types, expert bounding boxes), the proposed system obtains a mean average precision (mAP) of 0.246. This results in a relative improvement of 69% over the baseline YOLOv5 detector (mAP 0.145) and is significantly better than the performance of the standalone YOLOv5 (mAP 0.145) and Faster R-CNN (mAP 0.142). Using a two-stage cascade of a YOLOv5 detector and an EfficientNet-based classifier with transfer learning, the accuracy of disease localization in chest X-rays is significantly improved. This approach improves the reliability of computer-aided diagnostic workflows by reducing false positives and does not require large-scale pretraining or complex transformer architectures.

Keywords:

Thoracic Abnormality Localization, Multi-Modal Deep Learning, False Positive Pruning, Chest X-Ray Analysis, Transfer Learning

1. INTRODUCTION

The COVID-19 pandemic resulted in millions of deaths worldwide, and many survivors experienced long-term respiratory complications. Patients' lungs were the most severely affected organs, a leading cause of COVID-related mortality. Chest X-rays became a primary diagnostic tool for assessing lung damage. However, the surge in patient volumes overwhelmed healthcare systems [31], and physicians reported inadequate care due to shortages of staff, protective equipment, and testing facilities [16, 35]. Since the pandemic, the use of computer-aided diagnostic (CAD) tools—particularly those based on artificial intelligence—has increased. These methods have proven effective

for disease prediction, enabling preliminary triage and rapid isolation of critical cases.

Despite these advances, automated analysis of medical images (X-ray, MRI, CT, etc.) remains highly challenging. Achieving performance comparable to radiologists requires sophisticated algorithms and extensive domain knowledge. Deep learning has paved the way for accurate detection of chest X-ray abnormalities, supported by several state-of-the-art datasets. While disease identification aids early treatment, precise localization of abnormalities can further reduce radiologists' workload by highlighting regions of interest for review.

Existing approaches primarily focus on classifying chest X-rays as normal or abnormal, but rarely address localizing the affected region. This work aims to localize disease areas with high precision. Localization requires datasets with reliable ground truth bounding boxes. Among publicly available resources, VinDr-CXR [21] provides one of the most comprehensive collections, containing 22 types of findings annotated by professional radiologists. This study focuses on 14 abnormality types and their bounding box locations.

A two-stage CNN-based system is developed to predict and localize chest diseases from X-ray images using VinDr-CXR. First, a YOLO [26]-based object detector is deployed to identify candidate disease regions. Second, a transfer learning-based binary classifier (EfficientNet [33]) is incorporated to prune false positive bounding boxes produced by the detector. This two-stage process achieves high-precision localization.

The major contributions of this work are:

- (1) Fine-tuning a YOLOv5 model to simultaneously detect and localize 14 types of thoracic abnormalities from chest X-rays.
- (2) Developing a false-positive pruning filter using an EfficientNet-based binary classifier, which significantly reduces spurious bounding box predictions on healthy X-rays.
- (3) Employing transfer learning from the ChestX-Ray8 dataset to improve binary classifier accuracy with minimal additional training.
- (4) Achieving a mean average precision (mAP) of 0.246, outperforming standalone YOLOv5 (mAP 0.145) and Faster R-CNN (mAP 0.142) on the VinDr-CXR benchmark.

- (5) Analyzing predicted disease area distributions and heat maps, which demonstrate substantial overlap with ground truth locations, confirming the accuracy of the proposed approach.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents a high-level overview of the proposed approach. Section 4 describes the proposed methodology in detail. Section 5 reports experimental results. Section 6 concludes with a summary and directions for future work.

2. RELATED WORK

This section reviews prior work in chest X-ray analysis, organized by theme: classification-only methods, datasets and the annotation gap, image augmentation limitations, object detection and localization, false positive reduction, and recent transformer or foundation models. The gap addressed by the proposed multi-modal verification pipeline is then identified.

Early deep learning efforts focused on classifying chest X-rays for specific diseases. [13] proposed a CNN-based model to identify coronavirus using region and edge-based operations. [39] compared VGGNet [30], ResNet [10], and EfficientNet [33] for COVID-19 detection, providing coarse localization via class activation maps. Moses et al. [18] surveyed deep neural networks for disease detection in chest X-rays, covering COVID-19 pneumonia, tuberculosis screening, and pneumothorax detection. Similarly, [27] developed a CNN classifier that detected pneumonia with accuracy exceeding several state-of-the-art algorithms. While these works demonstrate high classification performance, they do not produce bounding boxes for abnormality localization—a key requirement for clinical interpretability. Similar localization challenges exist in other modalities (e.g., brain MRI [20] or foreign object detection [43]), but the focus of this work is thoracic diseases.

A major barrier to localization research has been the scarcity of large, expert-annotated datasets with bounding box information. CheXpert [11] contains 224,316 chest X-rays from approximately 65,000 patients, with a labeler identifying 14 common anomalies. However, it lacks bounding box annotations needed to train localization models. Similarly, ChestX-Ray8 [36] from the National Institutes of Health includes a small number of bounding boxes, used only to validate radiologist statements rather than for localization training.

The largest publicly available dataset with bounding boxes for multiple thoracic abnormalities is VinDr-CXR [21], comprising 18,000 chest X-rays with 14 types of abnormalities manually annotated by 17 expert radiologists. This dataset is used for training and evaluation in the present study. More recently, Fan et al. [7] constructed CXR-AL14, containing over 165,000 chest X-rays with more than 253,000 bounding boxes, and developed a deep learning framework that simultaneously identifies and localizes 14 common thoracic abnormalities while computing the cardiothoracic ratio, achieving mean average precision (mAP) values of 0.572–0.631 at an IoU threshold of 0.5. This represents a substantial improvement over prior localization benchmarks and underscores the value of large-scale, human-in-the-loop annotation pipelines. Several studies have applied image augmentation to boost deep learning performance [14, 19, 41]. However, augmentation is not universally beneficial. [4] showed that applying multiple augmentation techniques reduces F1 score in ECG-based medical data. In chest X-ray analysis, [6]

demonstrated that geometric transformations such as rotation and translation actually decrease classification performance for COVID-19 diagnosis. These findings motivated the targeted use of augmentation (e.g., preserving anatomical orientation) rather than aggressive geometric transforms in the proposed approach.

Beyond classification, researchers have explored object detection architectures to localize abnormalities. Cheng et al. [5] investigated deep learning-based object detection strategies for chest X-rays and explored few-shot object detection to address limited training examples for rare diseases. Tang et al. [34] introduced an attention-guided multiple instance learning framework that enables lesion localization without bounding box annotations during training. Zhang et al. [40] worked on semi-supervised object detection, leveraging both labeled and unlabeled chest X-rays to mitigate data scarcity. Specific to individual diseases, [8] developed a CNN-based model that detects and localizes tuberculosis in chest X-rays, but the work is limited to a single pathology.

A notable gap in the literature is the systematic application of YOLO-family detectors (e.g., YOLOv5) to thoracic abnormality localization. While YOLOv5 has been used in general medical imaging, only a few recent studies have applied it to chest X-rays [9, 42, 1], and none have combined it with a separate binary classifier for false positive pruning, a gap addressed in this work.

Minimizing false positives is critical for clinical deployment of object detectors. [15] developed a two-stage framework that employs separate verification networks to refine detection outputs. [23] introduced confidence calibration methods tailored to medical object detection, addressing overconfidence in deep learning models. However, these verification approaches are generic and not specifically optimized for thoracic abnormalities in chest X-rays. This work builds on this idea by using a lightweight EfficientNet binary classifier trained with transfer learning to prune false positives from YOLOv5 proposals.

Recent years have seen a shift toward transformer architectures. Statheros et al. [32] introduced CLARiT_y, a vision transformer-based model for joint multi-label classification and weakly-supervised localization, using class-specific attention tokens and a SegmentCAM module. It outperforms prior methods by over 50% on small pathologies (nodules, masses) using only image-level labels. Przulj et al. [25] presented ChestGPT, integrating a vision transformer with the Llama-2 large language model, achieving an F1 score of 0.76 for global classification on VinDr-CXR and generating localization outputs via prompt-guided inference. More broadly, Yang et al. [38] proposed CheXFound, a self-supervised vision foundation model pretrained on over one million chest X-rays, designed to generalize across classification, detection, and localization tasks without task-specific retraining. Wang et al. [37] unveiled a foundation model for chest radiography with impressive zero-shot localization performance from multi-institutional pretraining. While these transformer- and foundation-based approaches are powerful, they often require massive pretraining data, complex attention mechanisms, or high inference latency. In contrast, the proposed YOLOv5 + EfficientNet pipeline is lightweight and specifically targets the false-positive challenge, complementing these large models for practical clinical workflows.

At the clinical level, Anderson et al. [3] presented an FDA-cleared AI system that achieved an AUC of 0.976 for abnormality detection on chest X-rays, and showed that physicians aided by the system had significantly improved accuracy (AUC difference of 0.101) compared to unaided interpretation. Although that system does not

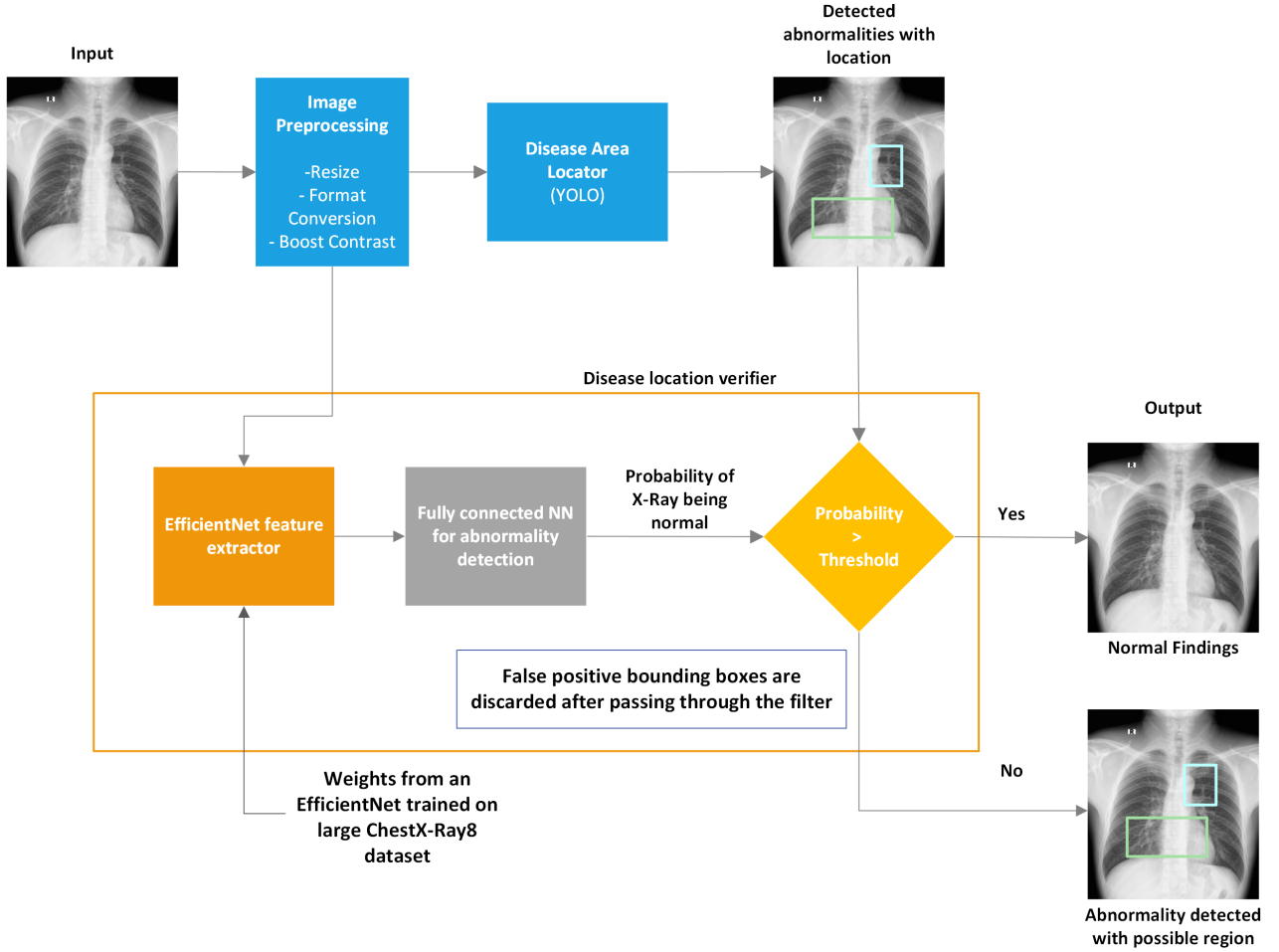


Fig. 1: Multi-modal chest X-ray abnormality detector. The YOLOv5 detector produces candidate bounding boxes; the EfficientNet binary classifier outputs a healthy/unhealthy probability; the filter module prunes false positives based on two thresholds.

provide bounding boxes, its clinical validation highlights the importance of minimizing false positives—a goal directly addressed by the verification module proposed in this work.

In summary, prior work has advanced chest X-ray classification, dataset availability, and—more recently—localization using transformers or foundation models. However, three key gaps remain. First, few studies apply YOLOv5 to thoracic abnormality localization, and none combine it with a dedicated false-positive pruning network. Second, existing verification methods [15, 23] are generic; they are not tailored to the specific false-positive patterns of YOLOv5 on chest X-rays. Third, despite the high mAP reported by Fan et al. [7] (0.572–0.631) on CXR-AL14, their dataset and annotation pipeline are not publicly available in the same form as VinDr-CXR, and their method’s computational cost is higher.

In contrast, the proposed multi-modal verification pipeline directly addresses these gaps, achieving a 69% relative improvement in mAP over standalone YOLOv5 (0.246 vs. 0.145) on the standard VinDr-CXR benchmark using a lightweight, reproducible two-stage design. This demonstrates that a simple detector-classifier cascade can substantially improve localization precision without large-scale pretraining or transformer complexity.

3. APPROACH OVERVIEW

The proposed system localizes thoracic abnormalities in chest X-rays by combining an object detector with a false-positive pruning module. The detector (YOLOv5) generates candidate bounding boxes around potential disease regions. However, YOLOv5 produces many false positives. To suppress these, a binary classifier (EfficientNet) is trained to predict whether an input X-ray is healthy or unhealthy. A filtering module then uses the classifier’s predicted probability to accept or reject the detector’s bounding boxes. This two-stage design improves localization precision without requiring large-scale pretraining or transformer architectures. Figure 1 provides a high-level illustration of the complete pipeline.

The proposed approach proceeds in four sequential stages: preprocessing, candidate bounding box generation using YOLOv5, whole-image abnormality classification using EfficientNet, and final pruning of false positives via dual thresholds.

First, preprocessing: All chest X-ray images are first resized to a fixed resolution (e.g., 640×640 pixels) and converted to the RGB format expected by YOLOv5. No aggressive geometric augmentations (rotation, translation) are applied, as prior work [6] has shown they can degrade classification performance on chest X-rays.

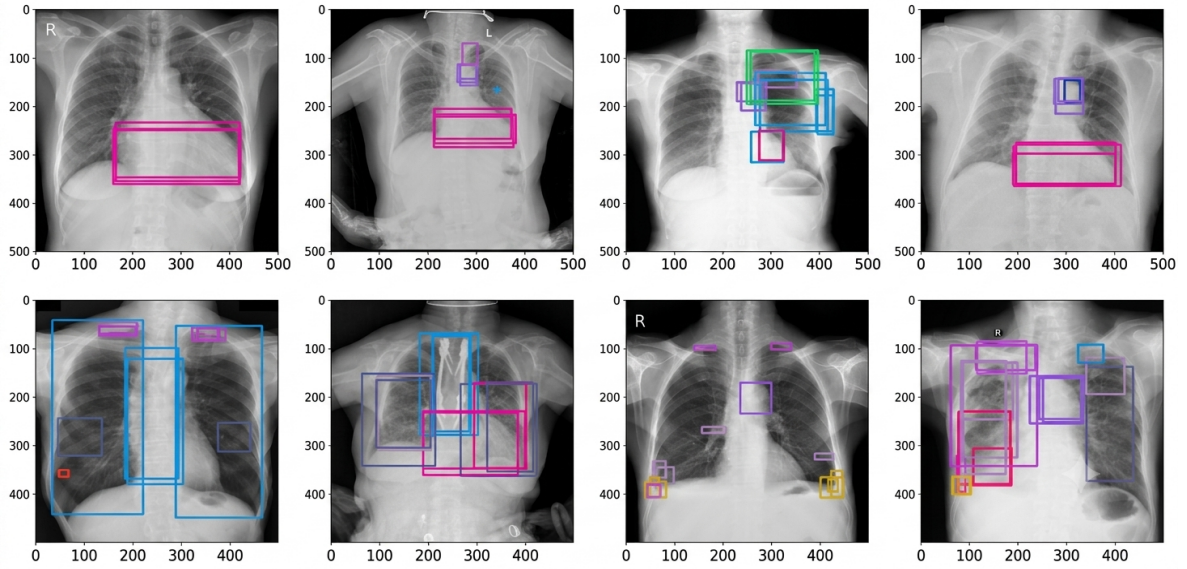


Fig. 2: Representation of X-ray images with disease locations as bounding boxes in VinDr-CXR [21] dataset.

Second, object detection with YOLOv5: A YOLOv5 model is trained on the VinDr-CXR dataset to predict bounding boxes around 14 types of thoracic abnormalities. The model outputs a set of candidate boxes, each with a confidence score. While YOLOv5 recalls many true positives, it also generates numerous false positives (bounding boxes on healthy regions).

Third, binary classification with EfficientNet: A separate binary classifier is trained using an EfficientNet backbone (pretrained on ImageNet and fine-tuned on VinDr-CXR). The classifier takes the entire chest X-ray as input and outputs a single probability p indicating the likelihood that the image contains at least one abnormality (i.e., unhealthy). Images with p close to 1 are highly likely to be unhealthy; those with p close to 0 are likely healthy.

Fourth, threshold-based pruning: The filter module combines the outputs of both models. It uses two empirically determined thresholds: a low threshold T_{low} and a high threshold T_{high} ($0 < T_{low} < T_{high} < 1$). For each X-ray image, the classifier produces a probability p (unhealthy likelihood). The pruning rules are:

If $p \leq T_{low}$: Do nothing, keep the disease prediction as is.

If $T_{low} < p < T_{high}$: Add Normal/Healthy probability and keep other disease bounding box detection prediction.

If $p \geq T_{high}$: Replace with Normal prediction with normal score 1.0, remove all other detection prediction.

Finally, T_{low} and T_{high} were tuned using a validation split of the VinDr-CXR dataset (not the test set) to avoid overfitting. For each candidate pair, the mean average precision (mAP) was computed on the validation set. The optimal values were $T_{low} = 0.25$ and $T_{high} = 0.8$, which maximized mAP. These thresholds are used for all results reported in this paper. The pruning logic is summarized within the yellow-bordered region of Figure 1.

4. SYSTEM DESIGN

This section describes the proposed system in detail. First, the data sets used and necessary pre-processing steps are

described, followed by details of the two deep learning models used in combination to localize the chest disease area with high confidence.

4.1 Data Set Description

VinDr-CXR [21] was used, a dataset containing 18,000 chest X-ray images as training data that were manually tagged by 17 radiologists. The samples were gathered from two of Vietnam's largest hospitals.

Each sample may include up to fourteen types of abnormalities or be labeled healthy. The abnormalities are: Aortic enlargement, Atelectasis, Calcification, Cardiomegaly, Consolidation, ILD, Infiltration, Lung Opacity, Nodule/Mass, Other lesions, Pleural effusion, Pleural thickening, Pneumothorax, and Pulmonary fibrosis. Not only did the radiologists classify the images, but they also personally identified the regions (in the form of rectangular bounding boxes) containing the anomaly. These bounding boxes (disease-affected areas of the lung) are given as four end coordinates: xmin, ymin, xmax, ymax, as shown in Figure 2. It is to be noted that multiple disease predictions may exist for a particular area, and each potential disease area is shown as a separate box. For that reason, the predicted disease areas often contain overlapping boxes, as depicted in Figure 2.

4.2 Data Processing and Augmentation

At first, the input images (chest X-rays) were converted to PNG format and then scaled down to (256 x 256), (512 x 512), and (1024 x 1024) dimensions and examined to determine whether the resizing affected the training model score. In general, 512px-sized images were selected as the initial experiment showed that precision didn't vary much for the difference in image sizes. Each X-ray in this dataset had numerous duplicate bounding boxes as multiple radiologists annotated them. Hence, the data was susceptible to overfitting. The intersection of those bounding boxes was calculated and the coordinates were averaged where the intersection over union (IoU) [29] reached 0.5. As a result, the duplicate boxes were merged, thereby minimizing the likelihood of the model overfitting. Figure 3 illustrates the idea of this process.

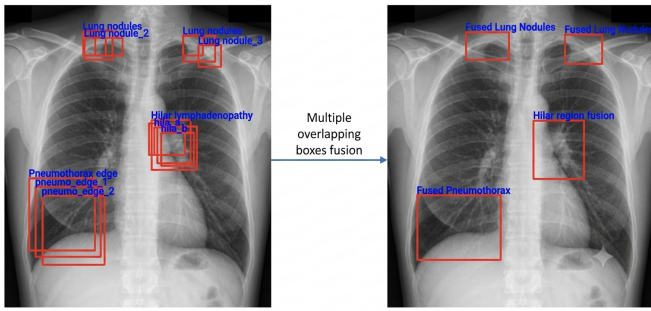


Fig. 3: Reduction of overlapping bounding boxes.

However, the primary dataset is insufficient in size to train an efficient object detector model. The dataset had only 4400 X-ray scans having an abnormality, which is extremely low for training the YOLO object detector model [12]. So, the dataset was supplemented by including pre-processed images. From the dataset, ten thousand images were randomly selected and an augmentation approach was conducted by applying a series of pre-processing operations (probabilistic selection): blurring, scaling, CLAHE [24], histogram equalization, random brightness, and contrast modification. After processing, these images were added to the original dataset, bringing the overall image count to 25,000, with roughly 7000 aberrant (i.e., having disease) images.

4.3 Deep Learning Model 1: Detect and Localize Chest Abnormalities

The primary objective of this work is to identify chest anomalies and also to determine the likely location of interest. YOLOv5, a modified version of YOLO [26] developed in PyTorch by the deep learning community, was utilized. Ultralytics' ¹ Python package was used, which implemented YOLOv5 and included utility files for researchers to use for straightforward training and inference. For training, the images were scaled to 512x512 px and fed straight into the YOLO model. Coordinates for the bounding box were transformed from absolute to relative values. The data was divided into six folds before training, with five used for training and one for validation. Pre-trained weights were used for the model to accelerate training, as the model was trained on the COCO [17] dataset, which is typically used for object detection research. These pre-trained weights helped to train the object detector models faster as it was initialized with prior knowledge. Throughout the training, hyperparameters such as learning rate and batch size were tuned for improved performance.

Following training, the probability distributions for the abnormalities as well as their bounding box coordinates were obtained. This model accurately predicted chest abnormalities; however, it was noted that the precision could be improved and that there were a significant number of false positives. To address them, a second deep learning model is deployed and a threshold-based pruning mechanism is then applied to filter out the false warnings generated by the YOLO-based model. The details of these steps are discussed below.

4.4 Deep Learning Model 2: Filter Out Low Confidence Disease Areas

To improve the prediction of bounding box regions generated by YOLO, a second CNN-based custom model was implemented, trained as a binary classifier to extract more knowledge from the data set. The following discussion describes this binary classifier operation in detail.

4.4.1 Restructuring EfficientNet as a Binary Classifier. EfficientNet was used as the backbone feature extractor of the binary classifier model. The EfficientNet has a combination of inverted residual CNN blocks, also referred to as MBConv blocks. These CNN blocks efficiently extract features from the 3-dimensional images and generate feature maps that are used by the classifier fully connected network to make the prediction. The structure of the custom EfficientNet used in this work is shown in Figure 4.

Images with anomalies were assigned to one category, while healthy images were assigned to the other. These images were then loaded into a fine-tuned EfficientNet [33] model, a state-of-the-art classification CNN [2] technique that outperforms comparable models in classification tasks. The classifier generates a probability distribution for each of the two classes described previously - normal and abnormal.

Additionally, the data was divided into six folds for this work, with one fold used as the validation set. After preparing the dataset with two categories, the Efficient feature extractor was trained to minimize the loss function as much as possible. The hyperparameters were tuned to get the best model with the least validation loss.

The EfficientNet backbone provides the feature map of the input X-ray image. The average pooling operation is then applied to the feature map and fed into a fully connected layer of neurons. A custom neural network block is constructed to use that feature map and predict whether the X-ray is abnormal or not. This neural network has 4 layers, including the output layer. The first layer has 1792 features (neurons), the second layer has 512 features and the third layer has 256 features. An empirical study was conducted on the count of features, and the mentioned combination gave the best training performance. The last layer of the neural network predicts the probability of chest images being normal or abnormal depending on the threshold mentioned earlier. The details of this threshold determination process are discussed next.

4.4.2 Threshold selection and Discarding False Positive Predictions. The binary classifier in the last stage of the proposed EfficientNet backbone gives the probability of an X-ray being abnormal or unhealthy. The output of this last layer indicates a probability score indicating how likely an X-ray image is abnormal (has diseases). Based on this, two threshold values are defined. If the probability score is greater than the high threshold, the image is labeled as unhealthy, and if the score falls below the lower threshold, it is labeled as healthy.

However, if the probability score lies between a low and high threshold, then the disease boxes are kept and the X-ray image is also considered to be healthy. This range is considered a low confidence range, and all possibilities (both healthy and unhealthy with associated disease area predictions) are considered.

To come up with the threshold values, an empirical study was conducted on these values, and it was found that a low threshold value of 0.25 and a high threshold value of 0.8 gave the best test mean average precision score. To illustrate the pruning mechanism concretely, consider an

¹<https://github.com/ultralytics/yolov5>

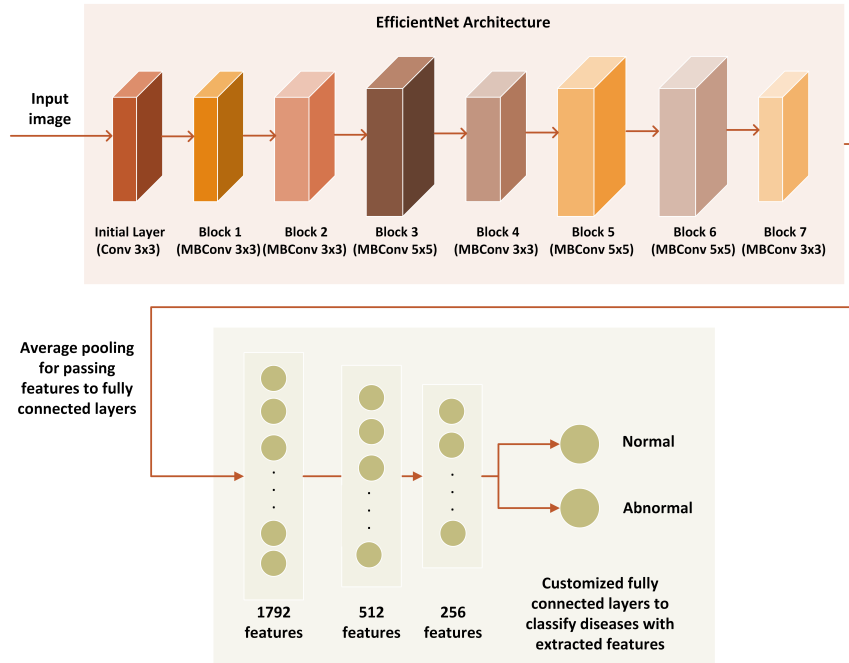


Fig. 4: Custom EfficientNet Structure for identifying abnormal X-Rays.

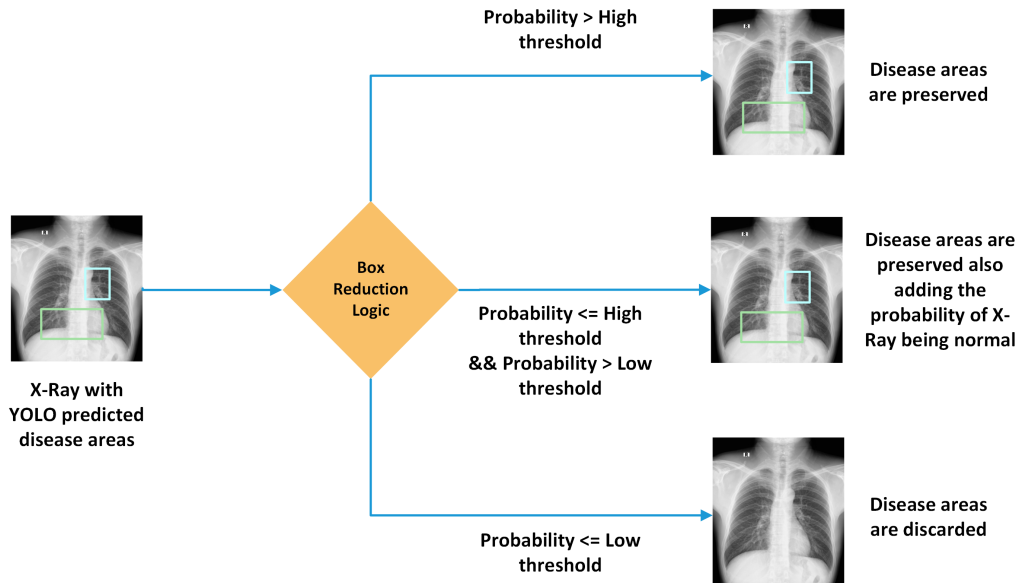


Fig. 5: Disease area pruning logic.

input chest X-ray image. Suppose the YOLO object detector predicts three disease bounding boxes with their associated probabilities: Cardiomegaly (probability 0.5), Infiltration (probability 0.89), and Nodule (probability 0.23). Concurrently, the EfficientNet binary classifier outputs a probability indicating how likely the X-ray is to be healthy (i.e., normal). The pruning filter then applies the following decision logic based on this healthy probability score:

Case 1 (Healthy probability = 0.85): Since 0.85 exceeds the high threshold of 0.8, the X-ray is confidently classified as healthy. The filter discards all disease bounding boxes and replaces them with a single Normal prediction (class label 14) with probability 1.0.

Case 2 (Healthy probability = 0.55): This value lies between the low threshold (0.25) and high threshold (0.8), indicating uncertainty. The filter retains all original disease bounding boxes and additionally appends a Normal prediction (probability 1.0) to the output, preserving both possibilities for downstream review.

Case 3 (Healthy probability = 0.12): Since this falls below the low threshold of 0.25, the X-ray is confidently classified as unhealthy. The filter retains all disease bounding boxes without adding any Normal prediction. Through systematic experimentation over the validation set, it was determined that a low threshold of 0.25 and a high threshold of 0.8 maximize the mean average precision

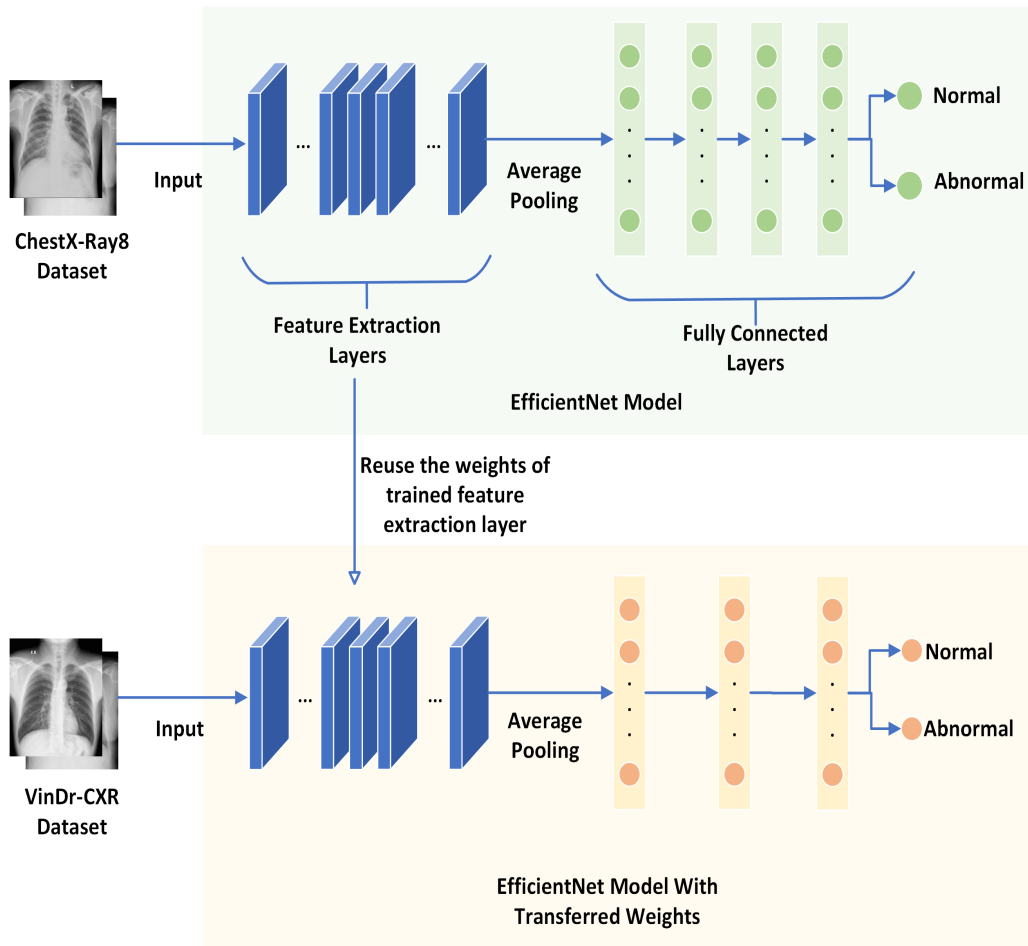


Fig. 6: Transfer pre-trained weights to binary classifier model.

on the test set. This three-way pruning strategy effectively balances sensitivity and specificity by preserving ambiguous cases while filtering clear false positives from healthy scans. The trained binary classifier provided the probability distribution of normal vs. abnormal classification of X-ray pictures. This information is used with the object detector's (earlier YOLO model) bounding boxes in this stage to eliminate false-positive disease boxes in healthy X-rays.

The weights of the models that perform the best in terms of object detection and binary classification concerning validation loss are loaded. Following that, an X-ray image is run through both models to obtain probability scores. The object detector model estimates the probability of the presence of a disease in an X-ray and also recommends bounding boxes for such diseases, whereas the EfficientNet model estimates a probability score for the normal versus abnormal class.

The binary classifier's score is then taken into account and passed through a probability score checking filter. For the filter's purpose, low and high thresholds are defined. If the abnormal class probability of the binary classifier is less than the low threshold, the X-ray is categorized as a normal X-ray and all disease bounding boxes are deleted. If the chance of the image being abnormal is between the low and high thresholds, the bounding box is retained, but the probability of the image being normal is also added. And if the chance of an abnormal class is larger than the high

threshold, the probability is retained and all abnormal class bounding boxes are displayed in the X-ray image. These threshold values were determined through experiment as they have the highest test set precision. The box pruning logic is depicted in Figure 5.

This pruning filter is quite effective at increasing the accuracy of object detection models, particularly when medical data is involved. A patient may become tense if an automated diagnostics system detects a possible abnormality in the chest X-ray image of the patient. A binary classifier is added for this reason; however, the overhead is minimal because the binary classifier is built on EfficientNet, which has fewer parameters than comparable ResNet [10] models.

4.5 Transfer Domain Knowledge from ChestX-Ray8 to Better Equip the Binary Classifier for Abnormality Detection

An EfficientNet was trained on ChestX-Ray8 [36] since it contains approximately 100k chest X-ray picture training examples. The weights were saved after fine-tuning the model to use them in the prior EfficientNet model.

The prior training model was taken and the final layer was removed, replacing it with two fully linked layers to enable it to output binary probability distributions. Training for an extended period was not required because it was previ-

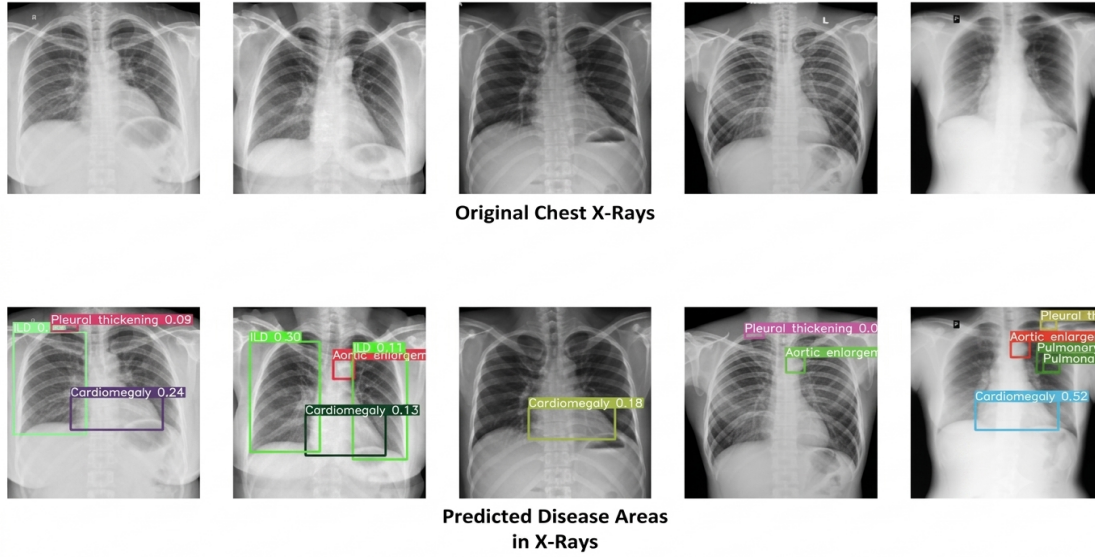


Fig. 7: Yolo predicted bounding boxes for a sample of the validation set images.

ously trained on a large chest X-ray dataset as a result of this transfer learning approach. As a result, it achieves a reasonable level of performance when distinguishing normal vs abnormal images after only a few epochs of training. This transfer-learning strategy is depicted in Figure 6. Incorporating transfer learning with the binary classifier and using it to reduce the false positive bounding boxes resulted in the maximum mean average precision for the proposed system.

5. RESULTS AND DISCUSSION

This section reports the experimental outcomes of the proposed system. The setup and evaluation metrics are detailed, followed by individual results for the YOLOv5 detector, the EfficientNet classifier, and the pruning filter. The effect of transfer learning is then shown, along with a comparative analysis against baseline models. The section concludes with a discussion of the system's strengths, limitations, and clinical relevance.

5.1 Experiment Setup and Performance Metrics

The entire experiment was conducted in the cloud using a two-core Intel Xeon CPU and a Tesla P100 GPU. The CPU had a memory of 13 GB and the GPU had a memory of 16 GB. Python was utilized throughout the research, and PyTorch [22] was used to train the deep learning models.

The Mean Average Precision (MAP) measure was utilized to evaluate the performance of the YOLO-based model, and the normal/abnormal prediction accuracy was used to evaluate the performance of the EfficientNet-based bespoke binary classifier. The value of Intersection Over Union (IoU) [29] was required to calculate Mean Average Precision, and an IoU of 0.4 was chosen, which is recommended by the authors of the dataset. After calculating the final probability score, the projected bounding box was superimposed on the X-ray images to identify the region containing the abnormalities.

5.2 Experiment Results

The following subsections report the performance of the YOLOv5 detector, the improvement from data augmentation, the impact of the binary classifier pruning, and the final gain from transfer learning.

5.2.1 Outcome of Disease Area Localization. The objective of this research is to not only detect but also localize disease in a human chest X-ray image. A CNN-based object detector was developed for this purpose. The object (affected areas in chest image) detector (YOLO) alone did not give satisfactory results as shown in Fig. 7.

It can be seen from the figure that in some X-rays the object detector classified too many diseases and drew overlapping boxes. This object detector model gave a mean average precision of 0.145 for the test set.

A limitation that was identified is the lack of adequate training set size. Among the 15k training X-ray images, only 4400 had abnormalities in them. To improve the training of the deep learning model and increase the validation performance, the dataset was increased from 15k to 25k by randomly augmenting images. After enlarging the training set size, the object detector model was re-trained and an improved test score (MAP) of 0.164 was achieved.

5.2.2 Effectiveness of binary classifier (EfficientNet+Custom Neural Network) in improving disease area prediction. The object detector model was good at predicting abnormality with bounding boxes, but upon further investigation it was noticed that in many healthy X-rays, the model was predicting bounding boxes even though the disease probabilities were low. To remove some of the false positives from the prediction, a CNN-based binary classifier was introduced that detects if an X-ray is healthy or not.

Using the trained classifier for abnormality detection, some of the healthy X-rays that had disease bounding boxes detected by the object detector were identified. After removing the disease boxes from the healthy X-rays, a significantly improved mean average precision for the test set was obtained, which is 0.231. The effectiveness of

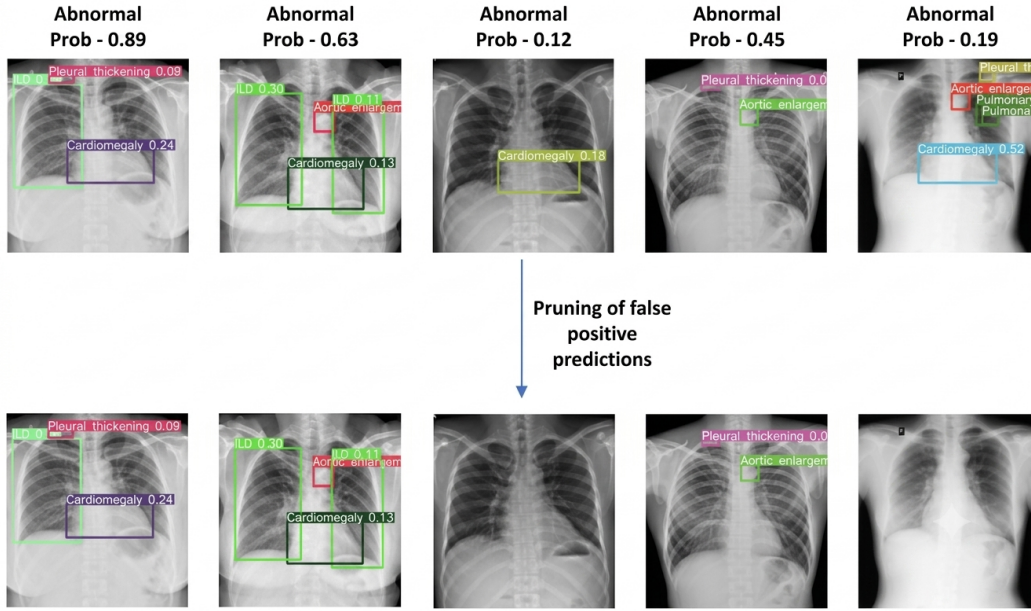


Fig. 8: Reduced Bounding Box (disease area) prediction after discarding false warnings.

Table 1. : Performance comparison between developed models and state-of-the-art models

Models	Mean Average Precision
Proposed Approach	0.246
YOLO + EfficientNet	0.231
YOLO with augmented dataset	0.164
YOLO [26]	0.145
Faster R-CNN [28]	0.142

this low-confidence disease reduction helped improve the detection of abnormalities in chest X-rays (Fig. 8).

5.2.3 Transfer Learning to improve binary classifier's accuracy. Additional data could not be added for transfer learning purposes for the disease object detector, because no other dataset provided such accurately annotated bounding boxes for localization purposes except for VinDr-CXR. However, it was possible to use the transfer learning approach for the binary classifier. ChestX-Ray8 [36] is a large and widely used chest X-ray repository for disease classification and it has around 100K chest X-ray images. This dataset was chosen for transferring the learned weights of the neurons from the last layer of the feature extractor.

After transferring the trained weights to the binary classifier, a decent validation accuracy was achieved in just 4–5 iterations. This significantly reduced the training time. The model was then fine-tuned and used for low-confidence box reduction. The test MAP score increased from 0.231 to 0.246. The increase in precision from these experiments and comparison with state-of-the-art models is shown in Table 1.

5.3 Discussion

The objective was to develop a computer-aided design (CAD) system based on deep learning that could predict and localize illness regions in human chest X-rays. For

this aim, VinDr-CXR was employed, which included radiologists' comments but had a short training sample size and duplicate illness bounding boxes. Throughout this research, a variety of strategies were used to circumvent the issue of a short dataset and maximize the accuracy of disease prediction.

The YOLO-based object-detecting model was trained. There are numerous alternative techniques for object localization problems, and Faster R-CNN [28] is one of the most widely utilized. However, YOLO was chosen due to its one-shot learning property, which saved a great deal of training and experimentation time. Additionally, the initial trial demonstrated that Faster R-CNN and YOLO provide identical precision, so the quicker method was chosen.

To evaluate and visualize the accuracy of the predicted disease bounding box areas, the area distribution was plotted as heat maps for individual diseases. Fig. 9, Fig. 10 and Fig. 11 depict the visualization of the box distributions. It can be seen from the heatmap comparison that area prediction of Calcification, Pleural Effusion, and Other Lesions had some mismatched distribution. Other diseases had almost the same distribution, which indicates that the object detector is quite accurate at suggesting the disease area boxes.

The initial assumption for Pneumothorax being less accurate is the low amount of data received for this disease. The training set did not have enough samples for the deep learning model to learn the characteristics of the disease. In the case of Pleural effusion, the object detector model seemed to have overfitted and predicted this disease only in some particular areas. Detecting other lesions is a complicated task compared to other diseases since it is spread evenly in the chest. So despite the lack of an exact match, the model still did a decent job predicting other lesions in human chest X-rays.

6. CONCLUSION

In this work, a multi-modal system is introduced that predicts 14 different types of anomalies from chest X-ray im-

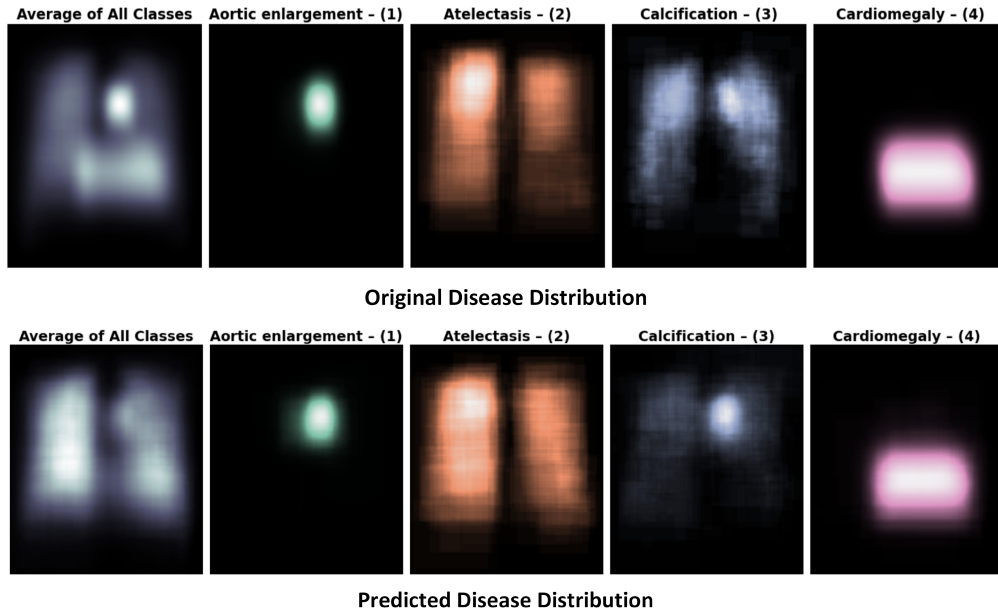


Fig. 9: Heatmap comparison of disease area distributions for Aortic enlargement, Atelectasis, Calcification, and Cardiomegaly.

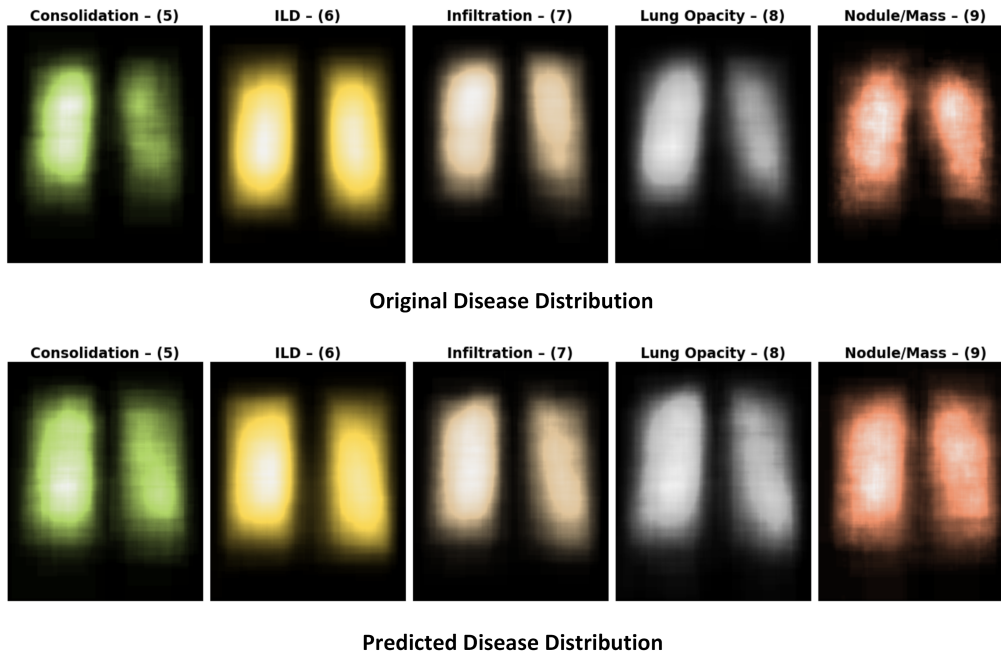


Fig. 10: Heatmap comparison of disease area distributions for Consolidation, ILD, Infiltration, Lung opacity, and Nodule.

ages using improved disease area identification. State-of-the-art methods like YOLO generate a lot of superfluous boxes (disease area) from images that could obstruct the appropriate diagnosis of a patient’s X-ray. To solve that, various data preprocessing techniques are deployed to eliminate duplicate bounding box detection (the first part of the proposed method). A binary classifier was also trained by restructuring the EfficientNet deep learning algorithm to find the likelihood that an X-ray image is abnormal or not (second part of the proposed method).

The designed pipeline could accurately generate disease areas for unhealthy X-rays and achieved a mean average precision (MAP) of 0.246, which outperformed state-of-the-art algorithms like YOLO (MAP 0.145) and Faster R-CNN (MAP 0.142). This established model can be used in computer-assisted diagnosis systems to annotate the X-rays for patients and lessen the workload for radiologists in situations like the coronavirus pandemic, where there was a sharp spike in the number of affected patients. The speed and potentially lifesaving effects of this advance-

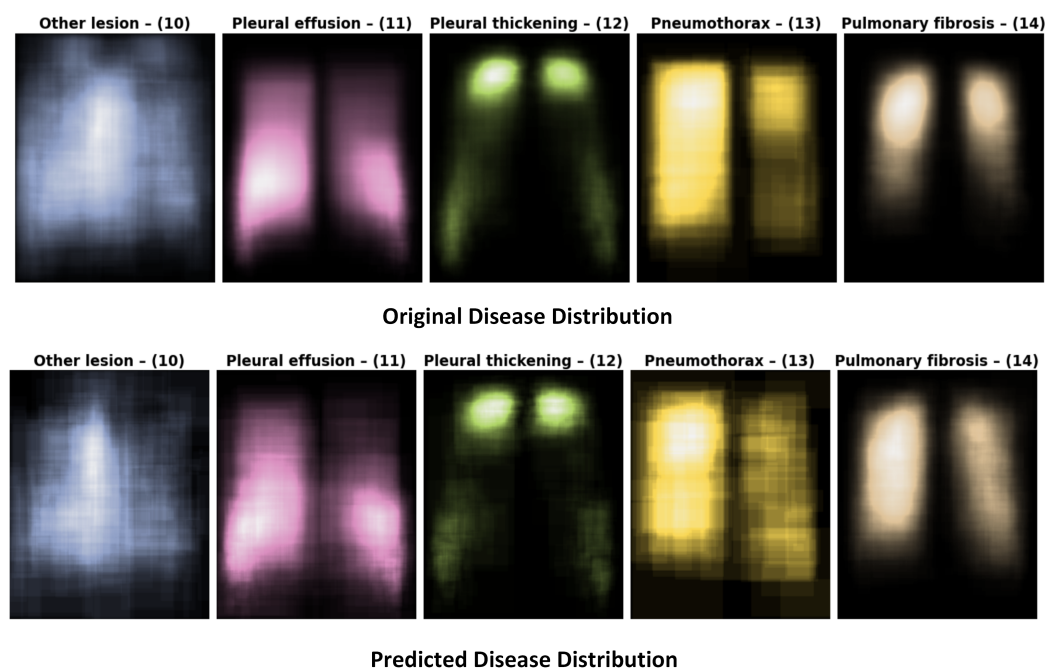


Fig. 11: Heatmap comparison of disease area distributions for Other lesions, Pleural effusion, Pleural thickening, Pneumothorax, and Pulmonary fibrosis.

ment could have a profound impact on how patients are diagnosed in hospitals.

7. REFERENCES

- [1] Sultan Ahmad, Stephen Neal Joshua Eali, Thirupathi Rao Nakka, Naceur Chihaoui, Tahani Alsubait, and Humaira Khanam. A scalable and reliable deep learning framework for enhanced brain tumor detection and diagnosis using ai-based medical imaging. *Frontiers in Medicine*, 13, 2026.
- [2] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. Ieee, 2017.
- [3] Pamela G. Anderson, Hannah Tarder-Stoll, Mehmet Alpaslan, Nora Keathley, David L. Levin, Srivas Venkatesh, Elliot Bartel, Serge Sicular, Scott Howell, Robert V. Lindsey, and Rebecca M. Jones. Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X-rays. *Scientific Reports*, 14:25151, 2024.
- [4] Talha Anwar and Seemab Zakir. Effect of image augmentation on ecg image classification using deep learning. In *2021 International Conference on Artificial Intelligence (ICAI)*, pages 182–186, 2021.
- [5] Yi-Ching Cheng, Yi-Chieh Hung, Guan-Hua Huang, Tai-Been Chen, Nan-Han Lu, Kuo-Ying Liu, and Kuo-Hsuan Lin. Deep learning-based object detection strategies for disease detection and localization in chest X-ray images. *Diagnostics*, 14(23):2636, 2024.
- [6] Mohamed Elgendi, Muhammad Umer Nasir, Qun-feng Tang, David Smith, John-Paul Grenier, Catherine Batte, Bradley Spieler, William Donald Leslie, Carlo Menon, Richard Ribbon Fletcher, et al. The effectiveness of image augmentation in deep learning networks for detecting covid-19: A geometric transformation perspective. *Frontiers in Medicine*, 8, 2021.
- [7] Weijie Fan, Yi Yang, Jing Qi, Qichuan Zhang, Cuiwei Liao, Li Wen, Shuang Wang, Guangxian Wang, Yu Xia, Qihua Wu, Xiaotao Fan, Xingcai Chen, Mi He, Jingjing Xiao, Liu Yang, Yun Liu, Jia Chen, Bing Wang, Lei Zhang, Liuqing Yang, Hui Gan, Shushu Zhang, Guofang Liu, Xiaodong Ge, Yuanqing Cai, Gang Zhao, Xi Zhang, Mingxun Xie, Huilin Xu, Yi Zhang, Jiao Chen, Jun Li, Shuang Han, Ke Mu, Shilin Xiao, Tingwei Xiong, Yongjian Nian, and Dong Zhang. A deep-learning-based framework for identifying and localizing multiple abnormalities and assessing cardiomegaly in chest X-ray. *Nature Communications*, 15:1347, 2024.
- [8] Ruihua Guo, Kalpdrum Passi, and Chakresh Kumar Jain. Tuberculosis diagnostics and localization in chest x-rays via deep learning models. *Frontiers in Artificial Intelligence*, 3:583427, 2020.
- [9] Shengnan Hao, Xinlei Li, Wei Peng, Zhu Fan, Zhanlin Ji, and Ivan Ganchev. Yolo-cxr: A novel detection network for locating multiple small lesions in chest x-ray images. *IEEE Access*, 12:156003–156019, 2024.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [11] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoos, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.
- [12] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, and Bo Ma. A review of yolo algorithm developments. *Procedia Computer Science*, 199:1066–1073, 2022.

- [13] Saddam Hussain Khan, Anabia Sohail, Muhammad Mohsin Zafar, and Asifullah Khan. Coronavirus disease analysis using chest x-ray images and a novel deep convolutional neural network. *Photodiagnosis and Photodynamic Therapy*, 35:102473, 2021.
- [14] Cherry Khosla and Baljit Singh Saini. Enhancing performance of deep learning models with different data augmentation techniques: A survey. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, pages 79–85, 2020.
- [15] H.E. Kim, H.H. Kim, B.K. Han, E.S. Ko, and K.J. Nam. Verification networks for false-positive reduction in medical object detection. *Medical Image Analysis*, 75:102272, 2022.
- [16] Victoria Kramer, Irina Papazova, Andreas Thoma, Miriam Kunz, Peter Falkai, Thomas Schneider-Axmann, Anke Hierundar, Elias Wagner, and Alkomiet Hasan. Subjective burden and perspectives of german healthcare workers during the covid-19 pandemic. *European archives of psychiatry and clinical neuroscience*, 271:271–281, 2021.
- [17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [18] Daniel A Moses. Deep learning applied to automatic disease detection using chest x-rays. *Journal of Medical Imaging and Radiation Oncology*, 65(5):498–517, 2021.
- [19] Mamillapally Nagaraju, Priyanka Chawla, and Neeraj Kumar. Performance improvement of deep learning models using image augmentation techniques. *Multimedia Tools and Applications*, 81(7):9177–9200, 2022.
- [20] Dillip Ranjan Nayak, Neelamadhab Padhy, Pradeep Kumar Mallick, Mikhail Zymbler, and Sachin Kumar. Brain tumor classification using dense efficient-net. *Axioms*, 11(1):34, 2022.
- [21] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations, 2021.
- [22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.
- [23] B.N. Patel, L. Rosenberg, G. Willcox, D. Baltaxe, M. Lyons, et al. Confidence calibration for object detection in chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 621–631. Springer, 2023.
- [24] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.
- [25] Petar Przulj. ChestGPT: Integrating large language models and vision transformers for disease detection and localization in chest X-rays. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, 2025. arXiv:2507.03739.
- [26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.
- [27] Mustafain Rehman, Qiao Pan, Dehua Chen, and Arslan Manzoor. Classification of chest x-ray images to detect pneumonia using cnn and transfer learning. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3498–3504, 2021.
- [28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.
- [29] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.
- [30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [31] Joanne Spetz. There are not nearly enough nurses to handle the surge of coronavirus patients: Here’s how to close the gap quickly. *Health Affairs Forefront*, 2020.
- [32] John M. Statheros, Hairong Wang, and Richard Klein. CLARiTy: A vision transformer for multi-label classification and weakly-supervised localization of chest X-ray pathologies. In *arXiv preprint arXiv:2512.16700*, 2024. arXiv:2512.16700.
- [33] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
- [34] Y. Tang, R. Gao, T. Han, C. Chen, S. Zhang, and S.K. Zhou. Attention-guided multiple instance learning for anomaly localization in chest radiographs. *IEEE Transactions on Medical Imaging*, 43(2):678–690, 2024.
- [35] Janice Hopkins Tanne, Erika Hayasaki, Mark Zastrow, Priyanka Pulla, Paul Smith, and Aser Garcia Rada. Covid-19: how doctors and healthcare systems are tackling coronavirus worldwide. *Bmj*, 368, 2020.
- [36] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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*, pages 2097–2106, 2017.
- [37] Z. Wang, Y. Liu, K. Chen, W. Zhang, Y. Zhou, A.L. Yuille, et al. Cxr-foundation: A general-purpose foundation model for chest radiography interpretation and localization. *Nature Medicine*, 31(2):345–356, 2025.
- [38] Zefan Yang, Xuanang Xu, Jiajin Zhang, Ge Wang, Mannudeep K. Kalra, and Pingkun Yan. Chest X-ray foundation model with global and local representations integration. *arXiv preprint arXiv:2502.05142*, 2025. arXiv:2502.05142.

- [39] Tahmina Zebin and Shahadate Rezvy. Covid-19 detection and disease progression visualization: Deep learning on chest x-rays for classification and coarse localization. *Applied Intelligence*, 51(2):1010–1021, 2021.
- [40] R. Zhang, Z. Xu, S.H. Park, K.H. Moon, and J.G. Lee. Semi-supervised object detection for chest x-ray abnormality localization. *Computerized Medical Imaging and Graphics*, 105:102201, 2023.
- [41] Weihang Zhang, Yuma Kinoshita, and Hitoshi Kiya. Image-enhancement-based data augmentation for improving deep learning in image classification problem. In *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, pages 1–2, 2020.
- [42] Xinyu Zhang, Lijun Liu, Xiaobing Yang, Li Liu, and Wei Peng. Nsec-yolo: Real-time lesion detection on chest x-ray with adaptive noise suppression and global perception aggregation. *Journal of Radiation Research and Applied Sciences*, 2025.
- [43] Fatema Tuz Zohora and K. C. Santosh. Circular foreign object detection in chest x-ray images. In K.C. Santosh, Mallikarjun Hangarge, Vitoantonio Bevilacqua, and Atul Negi, editors, *Recent Trends in Image Processing and Pattern Recognition*, pages 391–401, Singapore, 2017. Springer Singapore.