AI-Assisted Chest X-ray Annotation Tool for Abnormality Classification and Localization

Kyla Joy P. Shitan
University of the Immaculate Conception
Davao, Davao del Sur, Philippines
kshitan_200000000995@uic.edu.ph

Julieza Jane Bella A. Raper
University of the Immaculate Conception
Davao, Davao del Sur, Philippines
jraper_200000000052@uic.edu.ph

Karl Vincent F. Bersamin
University of the Immaculate Conception
Davao, Davao del Sur, Philippines
kbersamin_200000000668@uic.edu.ph

Kristine Mae M. Adlaon
University of the Immaculate Conception
Davao, Davao del Sur, Philippines
kadlaon@uic.edu.ph

ABSTRACT

Accurate interpretation of Chest X-ray (CXR) images presents challenges within the medical field, prompting the integration of Artificial Intelligence (AI) to support radiologists. This study introduces a comprehensive system crafted to facilitate the annotation process for radiologists. The primary objectives entail the development of an advanced annotation model utilizing EfficientDet for precise abnormality detection and bounding box placement. Furthermore, the system incorporates a customizable radiography tool aimed at refining annotations. Developed using EfficientDet and Django frameworks, the system underscores areas for enhancing model accuracy. Future endeavors will concentrate on gathering feedback from radiologists to further optimize the system's efficacy and utility in clinical settings.

KEYWORDS

Deep learning, chest-Xray, annotation, abnormality detection

1 INTRODUCTION

Approximately one billion radio imaging examinations are performed annually worldwide. The prevalence of radiologist errors has been estimated at 4% in a typical sample of cases encountered in practice (perceptual error). However, errors may be as high as 30% when test cases all show abnormalities (perceptual and cognitive errors) [3].

Even with our current medical annotation tools, there are still problems that require better improvements. Such challenges are: the Lack of Standardization, Time Constraints, Inter-observer Variability, and Complexity of Medical Images.

This gap and leave room for more research, and with this, the researchers will conduct further study which aims to be able to optimize the annotations that exist in a Chest-X Ray. The study will utilize two publicly available datasets; The NIH CheXpert Chest X-Ray dataset, and the Vin-DR CXR Dataset.

The datasets have 14 abnormalities overall, with 9 in common; Atelectasis, Cardiomegaly, Consolidation, Infiltration, Nodule/Mass, Pleural Thickening, Pneumothorax, Pulmonary Fibrosis, and Pleural Effusion.

A study on the effects of model augmentation in the diagnosis of CXR was conducted and it was found that machine learning tools created to simplify CXR interpretation perform well, enhance clinicians' detection abilities, and boost the effectiveness of radiology workflow. Clinical engagement and expertise will be crucial to secure the adoption of high-quality CXR machine learning systems [1]. This form of integration still needs further study after its initial usage during the COVID-19 pandemic [2].

A study was conducted for HITL (Human-in-the-loop) solution to improve chest radio-graph diagnosis. The extensive implications for future clinical AI deployment and implementation strategies arise from the superior diagnostic accuracy exhibited by the combined HITL AI solution when compared to both radiologists and AI functioning independently [6].

A study focused on enhancing ICU chest X-ray classification for diagnosing pathology in critical patients. The research utilized approaches such as using manual annotations, automatically generated silver labels, or a combination of both to evaluate their impact on classification performance [8].

With limited manual annotation, models trained on silver labels notably enhanced performance. The MS model, trained solely on silver labels, achieved a 75.3% AUC score, increasing to 75.5% with transfer learning (MC+S). However, the quality of silver labels was crucial; if too erroneous, transfer learning proved a useful alternative. Combining silver labels with transfer learning and additional training on gold labels yielded optimal results.

Another study focuses on the challenging task of diagnosing chest-related diseases through chest X-ray (CXR) radiography.

With this dataset, the researchers employed an ensemble approach. Leveraging an ensemble of deep learning models including EfficientNet-B5, Xception, and DenseNet-201. The model first classified diseases based on infected organs (heart or lung), achieving an impressive AUC of 0.9489 for multi-classification. In the subsequent binary classification phase for specific diseases, the model demonstrated outstanding average AUC values of 0.9926 for heart diseases and 0.9957 for lung diseases. The study's innovative augmentation techniques and careful hyperparameter tuning, the research achieved superior results, surpassing previous models. Rigorous testing on various diseases, including pneumonia, edema, and consolidation, consistently demonstrated high accuracy (e.g., 0.9954 accuracy and 0.9956 AUC for pneumonia), underscoring the model's robustness and reliability across different disease categories [5]. The study has a limitation, primarily the absence of a detailed discussion on potential challenges faced during the implementation process.

Another study utilized the VinDr-CXR dataset where the researchers proposed a two-step approach; To employ the use of YOLOv5 to pinpoint abnormalities' locations, and, a binary CNN classifier, ResNet50, to classify these abnormalities [7].

The findings showed an enhancement with the two-step method, achieving a notable 77% F1 score and an mAP@0.5 score of 81.2% when YOLOv5 and ResNet50 were combined. This surpassed single-step approaches like YOLOv5, Faster R-CNN, and CheXNet.

The next study proposed a novel two-step approach for classifying chest X-ray (CXR) images. The first step involved multi-class classification, categorizing images into normal, lung disease, and heart disease. The second step focused on binary classification, identifying specific diseases within the lungs and heart. To implement the two-step classification approach, they developed two deep learning methods: DC-ChestNet, an ensemble learning of three deep convolutional neural network (DCNN) models, and VT-ChestNet, based on a modified Swin transformer architecture [13].

In the first phase, VT-ChestNet outperformed competitors with an AUC of 95.13%, followed by the average AUCs of 99.26% for heart diseases and 99.57% for lung diseases. DC-ChestNet also yielded promising results, starting with a 94.89 AUC and demonstrating notable accuracy in binary classification, achieving 99.26% AUC for heart diseases and 99.57% AUC for lung diseases.

The study by [9], which leveraged the NIH CheXpert dataset, focuses on comparing radiologists and a convolutional neural network-based AI algorithm in interpreting chest X-ray images.

Using a clinician-guided approach, the researchers categorized potential findings in chest X-rays systematically. Results show the AI algorithm achieved an AUC of 0.807 for labels and a weighted mean AUC of 0.841 after training. However, it excelled in high-prevalence findings and performed slightly less for rarer conditions. The comparative accuracy, measured using the Kappa statistic, was 0.543 for the AI algorithm and 0.585 for radiologists.

The study also details a method for labeling images based on radiological reports, achieving high precision (99.2%) and recall (92.6%). Limitations include an initially unbalanced dataset and a small number of radiology residents in the comparison.

In [4], an innovative approach using Weighted Boxes Fusion (WBF) to combine annotations from multiple radiologists is introduced. This method enhances abnormality detection in chest X-ray images by leveraging the expertise of multiple radiologists to improve deep neural network performance.

The proposed approach achieved better mean average precision (mAP) scores, indicating its effectiveness in training image detectors from labels provided by multiple radiologists.

Objectives:

Building upon relevant studies, the researchers have identified the following objectives to guide the investigation and development process:

(1) Develop an annotation model for abnormality detection and bounding box placement on chest X-ray images using EfficientDet
(2) Customize a radiography tool that allows radiologists to review, edit, and refine the annotations generated by the model, ensuring accuracy and incorporating domain expertise.
(3) Implement a method to save annotations by the radiologist and the model.
(4) Assess model performance in abnormality detection and annotation accuracy using metrics like mAP, IoU, accuracy, F1 score, confusion matrix, and inference speed. Evaluate accuracy in identifying abnormalities.

2 METHODOLOGY

In this section, researchers will detail the methodology for developing the annotation tool and detection model.

2.1 Conceptual Framework

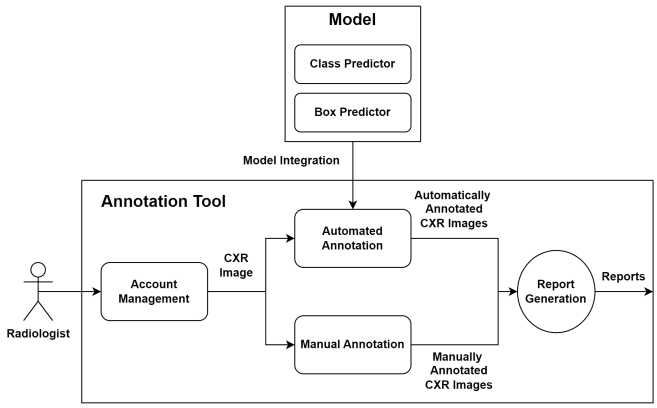

Figure 1: Conceptual Framework

As shown in the figure, this framework outlines a systematic process for abnormality detection in Chest X-ray images, combining human expertise and computational algorithms for accuracy and efficiency. The Annotation Model, EfficientDet, locates and classifies abnormalities, while radiologists refine initial annotations using the Annotation Tool.

Annotation and Classification Model:
The Annotation Model, built on EfficientDet, efficiently processes Chest X-ray images, identifying abnormalities. Integrated into the platform, it speeds up annotation by providing initial automated annotations, which radiologists can review and improve.

Annotation Tool:
The annotation tool supports radiologists in diagnostic tasks. Radiologists log in securely to upload chest X-ray images, which undergo two annotation methods: automated annotation by an EfficientDet model and manual annotation by radiologists. After annotation, the tool generates comprehensive reports, ensuring efficient workflow from image upload to report generation.

2.2 Model Preparation and Integration

This section outlines preparing, training, and integrating the pre-trained EfficientDet D0 model for annotating chest X-ray images from the VinBig and NIH datasets.

Data Preprocessing: Data preprocessing included converting images from PNG to JPG for consistency and resizing them to 512x512 pixels for uniformity and accuracy. The model focused on 9 common classes for streamlined training. Annotations for the same image were merged using Weighted Box Fusion (WBF) to improve accuracy. Out of 1908 images, 1527 were for training, and 382 for validation, with an 80% training and 20% validation split.

Model Training For model training, TensorFlow and Keras were used. The EfficientDet D0 model had pre-trained weights and was configured with 9 classes. Hyperparameters were carefully chosen, and regularization reduced errors. Data augmentation improved adaptability, and performance evaluation used COCO detection metrics.

2.3 Annotation Tool Features Implementation

This part discusses the features of the annotation tool. It goes into detail about the major and minor features that the system offers.

User Account Management User account management involves a simple registration form for new users to sign up and a secure logout function to end sessions.

Image Management Image management allows users to upload and change images. Image annotation and processing involve automatic annotation with the model and manual tools for radiologists. Users can draw boxes, add labels, and zoom for detailed inspection, aiding in categorization and reporting.

Reporting and Data Management This include generating reports summarizing key findings from annotated images and saving both images and reports for future reference.

2.4 Materials Used

In this part, the discussion shifts to the tools used to craft the system.

Software Tools The project used XAMPP with phpMyAdmin for local server development and database management. Django, Konva, and Bootstrap were also utilized . MySQL handled structured data storage.

Datasets The VinBig and NIH Chest X-ray datasets, obtained from Kaggle, supplied chest X-ray images with annotations for training the EfficientDet model.

Hardware and Computational Resources Google Colab, a cloud-based platform, was used to train the EfficientDet model. The annotation tool was developed and hosted in a local server environment provided by XAMPP.

2.5 Testing and Validation

The model will undergo thorough testing to assess capabilities and identify improvements.

Average Precision and Average Recall: These metrics provided insights into how accurately the model could identify and localize objects of interest across different levels of detection strictness and object sizes.

Area Under the Curve (AUC) Scores: These scores provided a quantitative measure of the model's ability to distinguish between different types of abnormalities.

3 PRELIMENARY RESULTS

This study applies object detection to chest X-ray analysis. The aim is to identify and localize key features and abnormalities in chest X-rays. The model was trained for 1300 steps on a merged dataset of chest X-ray images.

Figure 2: Total Loss over Training Steps

After analysis, it was found that the model reached its peak at the 1200th step with the lowest training loss. However, at the 1300th step, there was a noticeable increase in total loss, indicating suboptimal learning. Therefore, training was stopped at the 1200th step for optimal model performance.

At the 1200th step, the model's loss metrics were:

- Classification Loss: 0.43003213
- Localization Loss: 0.010713379
- Regularization Loss: 0.032603912
- Total Loss: 0.47334942

During training, the loss steadily decreased, indicating progress. However, there is still room for improvement.

3.1 Performance

The object detection model's performance was evaluated using Average Precision (AP) and Average Recall (AR) metrics across various IoU thresholds and image area sizes. Higher AP and AR values indicate better performance. IoU measures overlap between predicted and ground truth bounding boxes, while 'maxDets' sets the maximum number of detections per image.

3.2 Analysis of Model Performance

1. Overall Precision and Recall:

AP @[IoU=0.50:0.95 | area=all | maxDets=100 = 0.104:]
This suggests that, on average, the model correctly identifies relevant objects with moderate accuracy when considering various IoU thresholds. It implies there's significant room for improvement in the model's precision.

2. Precision at Specific IoU Thresholds:

AP @[IoU=0.50 | area=all | maxDets=100 = 0.275:]
At an IoU threshold of 0.50, the model performs considerably better, indicating it can detect objects with a reasonable overlap with the ground truth.

AP @[IoU=0.75 | area=all | maxDets=100 = 0.050:]
At a higher IoU threshold of 0.75, the precision drops significantly, suggesting the model struggles with very accurate localization of objects.

3. Precision by Area Size:

The model shows varying performance based on the size of the detected objects. It performs best for large objects (AP = 0.146) compared to small (AP = 0.008) and medium-sized objects. (AP = 0.095)

4. Recall by Area Size:

Small Areas [IoU=0.50:0.95 | maxDets=100 = 0.057:]
The model's significantly lower recall for small areas suggests that it struggles to correctly identify smaller objects in the chest X-rays.

Medium Areas [IoU=0.50:0.95 | maxDets=100 = 0.245:]
There's a need for enhancement of the model's ability to detect medium-sized features. Considering that abnormalities usually fall into this size range.

Large Areas [IoU=0.50:0.95 | maxDets=100 = 0.349:]
This higher recall rate for large objects suggests that the model is more effective in identifying larger anomalies.

5. Recall Analysis:

The model's recall scores (AR) range from 0.174 to 0.349, showing improved detection as it makes more detections, especially for larger objects.

The model can identify chest X-ray features to some extent, but its accuracy in precisely localizing objects (higher IoU thresholds) and detecting smaller objects needs improvement. This moderate performance could be due to dataset complexity, model limitations, or the need for more training or advanced augmentation techniques.

3.3 AUC Scores

The table below displays the Area Under the Curve (AUC) scores for each class.

Table 1: AUC Scores

Abnormalities	AUC Scores
Cardiomegaly	0.54
Pleural Thickening	0.51
Pulmonary Fibrosis	0.56
Pleural Effusion	0.63
Nodule/Mass	0.54
Infiltration	0.58
Atelectasis	0.46
Consolidation	0.47
Pneumothorax	0.5

The AUC score evaluates how well a model can distinguish between classes, derived from the ROC curve. While the model shows potential in spotting some chest X-ray issues, its AUC scores aren't optimal, particularly for certain conditions. This highlights the need for further model refinement and investigation into areas of poor performance.

3.4 Overall Performance

The model demonstrates a moderate level of accuracy in detecting various chest abnormalities, as indicated by the AUC scores. However, there is room for improvement, especially in classes with AUC scores closer to 0.5, such as Pneumothorax and Consolidation. The model shows relatively better performance in detecting Pleural Effusion and Infiltration, as evidenced by higher AUC scores. While the model shows promise, its current level of accuracy necessitates further refinement before it can be reliably used in clinical settings. Improvements could include additional training data, further hyperparameter tuning, or exploring more complex model architectures.

4 FURTHER WORK

The evaluation of the AI-assisted chest X-ray abnormality classification model reveals promising yet moderate performance across various metrics. Integrated into the annotation system, the model can offer valuable assistance to radiologists by simplifying the detection and annotation process. While demonstrating an ability to identify abnormalities within chest X-ray images, there are notable areas for improvement, particularly in achieving higher precision and recall rates, especially for smaller abnormalities and precise localization tasks.

Currently, our efforts are focused on refining the model, representing just the initial step in this endeavor. Addressing data preparation and quality issues is important to enhancing model performance. Achieving acceptable metric results is crucial for future use, as there is still ample room for improvement.

In the future, we aim to assess the annotation tool's performance in clinical settings. Upon achieving satisfactory results, we'll integrate radiologists' feedback through the tool. Their input is crucial for correcting model errors and improving workflow efficiency.

Ongoing refinement and optimization efforts, including additional training data and fine-tuning of hyperparameters, are essential to enhance the model's capabilities and meet the rigorous standards required for clinical deployment. Despite current limitations, the integration of the model represents a significant advancement in AI-assisted radiology, with the potential to improve diagnostic accuracy and efficiency in clinical practice. Looking ahead, future iterations will focus on further improving the system and allowing radiologists to actively test and provide feedback, ensuring continuous enhancement and adaptation to clinical needs.

REFERENCES

[1] Hassan K Ahmad, Michael R Milne, Quinlan D Buchlak, Nalan Ektas, Georgina Sanderson, Hadi Chamtie, Sajith Karunasena, Jason Chiang, Xavier Holt, Cyril HM Tang, et al. 2023. Machine learning augmented interpretation of chest X-rays: a systematic review. *Diagnostics* 13, 4 (2023), 743.

[2] Harrison X Bai, Robin Wang, Zeng Xiong, Ben Hsieh, Ken Chang, Kasey Halsey, Thi My Linh Tran, Ji Whae Choi, Dong-Cui Wang, Lin-Bo Shi, et al. 2020. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. *Radiology* 296, 3 (2020), E156–E165.

[3] Warren B Gefter, Benjamin A Post, and Hiroto Hatabu. 2023. Commonly missed findings on chest radiographs: causes and consequences. *Chest* 163, 3 (2023), 650–661.

[4] Khiem H Le, Tuan V Tran, Hieu H Pham, Hieu T Nguyen, Tung T Le, and Ha Q Nguyen. 2023. Learning from multiple expert annotators for enhancing anomaly detection in medical image analysis. *IEEE Access* 11 (2023), 14105–14114.

[5] Adnane Ait Nasser and Moulay A Akhloufi. 2022. Classification of CXR chest diseases by ensembling deep learning models. In *2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 250–255.

[6] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. 2019. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine* 2, 1 (2019), 111.

[7] Vu-Thu-Nguyet Pham, Quang-Chung Nguyen, and Quang-Vu Nguyen. 2023. Chest x-rays abnormalities localization and classification using an ensemble framework of deep convolutional neural networks. *Vietnam Journal of Computer Science* 10, 01 (2023), 55–73.

[8] Helen Schneider, David Biesner, Sebastian Nowak, Yannik C Layer, Maike Theis, Wolfgang Block, Benjamin Wulff, Alois M Sprinkart, Ulrike I Attenberger, Rafet Sifa, et al. 2022. Improving Intensive Care Chest X-Ray Classification by Transfer Learning and Automatic Label Generation.. In *ESANN*.

[9] Joy T Wu, Ken CL Wong, Yaniv Gur, Nadeem Ansari, Alexandros Karargyris, Arjun Sharma, Michael Morris, Babak Saboury, Hassan Ahmad, Orest Boyko, et al. 2020. Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. *JAMA network open* 3, 10 (2020), e2022779–e2022779.