Deep Learning for Multi Disease Classification in Chest Imaging

Zina Faisal and Jamal Salahaldeen Majeed Alneamy
Department of Software Engineering, College of Computer Science and Mathematics, University of Mosul, 41002 Mosul, Iraq
Zina.24csp16 @student.umosul.edu.iq, jamal_alneamy @uomosul.edu.iq

Keywords: Chest X-ray, Multi-Label Classification, Deep Learning, Convolutional Neural Networks, Transformer, Medical Image Analysis.

Abstract: X-ray imaging is used extensively in detecting diseases of the chest; however, this is affected by factors that reduce classification accuracy, such as blurred images and significant class imbalance. In this paper, a deep learning model is proposed for classification of images containing multiple chest diseases. The images used were a combination of images from the NIH ChestX-ray14 dataset, containing 112,120 images of 14 diseases, and images containing only COVID-19 cases. Three baseline architectures were used, namely, ResNet50 as a reference model, EfficientNet-B5, and CoAtNet-0-rw. The class imbalance was reduced using the Asymmetric Loss (ASL) algorithm, as well as other training optimization techniques. In addition, a data purification phase was included to correct any tags that were inconsistent, thus preventing model confusion. The best model was EfficientNet-B5, which had an AUROC of 0.862. The maximum accuracy was obtained by using CoAtNet-0-rw, which had an accuracy of 95%. In addition, a hybrid model was proposed, which used a combination of deep model features and an SVM classifier. The model had good generalizability, as observed from the evaluation using an external data source, VinDr-CXR.

1 INTRODUCTION

The importance of chest X-ray (CXR) imaging in the diagnosis of thoracic diseases has not diminished over time, given its accessibility and rapid clinical integration into healthcare systems across the globe. In routine medical practice, CXR imaging is one of the first diagnostic tools employed by clinicians in the detection of thoracic and pulmonary diseases. The increase in thoracic diseases across the globe, especially in recent times due to COVID-19, has emphasized the urgent need for computer-assisted diagnostic systems that can aid clinicians in their practice, especially when they are under heavy pressure of workload [1]. Advancements in deep learning have revolutionized medical image analysis and diagnosis in recent times. Convolutional neural networks (CNNs) and their variants have shown promising results in thoracic disease detection and diagnosis from chest radiographs [2], [3]. Although several studies have attempted multi-disease classification frameworks and their implementation in thoracic disease diagnosis, radiographic analysis is far more challenging in routine practice than in experimental scenarios. In routine practice, chest X-ray images often involve multiple concurrent abnormalities rather than an isolated disease state. The multi-label nature of thoracic disease diagnosis is more challenging and complicates the development of automated classification systems. In addition, datasets like ChestX-ray14 have a long-tail distribution, where a few classes dominate the dataset, while others remain highly underrepresented. The development of effective and principled strategies to handle data imbalance is one of the major challenges in thoracic disease diagnosis from radiographs. Another practical constraint is related to label-image inconsistency, which is often introduced during automated dataset development, as discussed in reference [4]. The presence of noise in weakly labeled medical datasets may have an adverse effect on model reliability and practicality. Additionally, existing studies often do not report results of external validation, making it difficult to compare and analyze generalization across independent clinical environments. Therefore, external validation of models using datasets like VinDr-CXR is essential.

To address these challenges and limitations, this study proposes an effective and comprehensive framework for multi-label classification of 15

thoracic diseases and radiographic findings. The proposed system is designed by incorporating state-of-the-art deep learning models like ResNet, EfficientNet, and CoAtNet, as discussed in references [5]-[7]. Additionally, an effective data purification strategy is incorporated into the framework to address label-image inconsistency, as discussed in this study. The framework is trained and validated by employing internal testing and external validation, which is essential for assessing the reliability of model generalization. The contributions of this study can be summarized as follows:

1) The proposed framework is effective and scalable enough to classify 15 thoracic diseases and radiographic findings into multiple labels;
2) The framework is effective in handling long-tail class imbalance problems by employing Asymmetric Loss Optimization;
3) The framework incorporates an effective data purification strategy that is essential for addressing label-image inconsistency;
4) The framework is effective in comparing and analyzing the performances of state-of-the-art models like ResNet, EfficientNet, and CoAtNet;
5) The framework is validated by employing internal testing and external validation, which is essential for assessing model reliability and generalization.

2 RELATED WORK

The researcher [8] built a Federal Deep Learning system that divided the data across three hospitals and adopted several models, achieving an accuracy rate of (98.3%) for binary classification. He also used techniques to identify the locations of lesions, but the researcher limited his study to one disease and limited images. A system was developed by the researcher mentioned in [9] that focused on the diagnosis of six types of lung diseases. The model was trained on the Xception architecture, utilizing 11,716 images, along with data augmentation techniques and noise removal. The model achieved an accuracy of 97.3%. However, the model utilized multi-class classification, rather than multi-label classification. A model was developed by the researchers mentioned in [3] that focused on the classification of the X-ray images utilizing the convolutional neural network, which was sensitive to the lesions. The model was presented to the data panel of the experts, who determined the location and type of the lesions. The results were then tested, utilizing the AUROC metric, which achieved a result of 0.888. A model was developed by the researcher mentioned in [10] that focused on the diagnosis, utilizing the EfficientNet model, along with the Weighted Multiclass Cross-Entropy, which was trained on the NIH ChestX-ray14 dataset. The model achieved an accuracy of 88%. An external evaluation of the model, focusing on pneumonia and ignoring the other 14 diseases, did not confirm the model's ability to generalize. The study by [2] uses two advanced models, namely, EfficientNet-b5 and CoAtNet-0-rw, and compares them with a baseline model using ResNet-50 after integrating a novel loss function, Lours, which balances the classes. The best results are obtained using the CoAtNet-0-rw + Lours model, where the accuracy is 0.8157 and AUROC is 0.842. Group-CAM is also applied to identify the diseased regions in the image. However, the authors also acknowledge the problems associated with false tags. The study conducted by the researcher [11] suggested the development of the ChestXRay6 model and the integration of three datasets, aggregating 9,514 images. The class imbalance issue was addressed by using data augmentation, and the binary classification accuracy reached 97.94%. On the other hand, the accuracy of multi-class classification decreased to 80%. The occurrence of overfitting was observed when the training accuracy was 98%, while the validation accuracy was 77%. In another study, the researcher [12] grounded his study on the basis of the comparative evaluation of the different attention units. Among the units, the best performance was achieved by the Pyramid Squeeze Attention unit, though the best performance was achieved by the integration of the units, resulting in an AUROC of 0.898. In another study, the researcher [13] proposed the evaluation of the performance of the models based on the comparison of the different EfficientNet B0-B7 models, with the addition of the technique that enhanced the model's performance. The best performance was achieved by the EfficientNet-B7 model, with an accuracy of 95.83% and an AUROC of 0.8309. However, the model was not evaluated for generalization. Study [14] proposed the model known as the CXR-MultiTaskNet, which is capable of carrying out two different tasks, including disease classification and pest location, using the ResNet50 architecture. The model was able to achieve an accuracy of 0.989, and the pest location was also able to achieve an IoU of 0.851, but the model had the disadvantage of requiring the positioning tags, which are not usually available. For clarity, a comparison of the related studies is provided in Table 1.

Table 1: Comparison of previous studies.

Researcher	Year	Algorithm	Dataset	Data Size	Evaluation Metric
Sourasekhar et al.	2020	ResNet18 (Federated CNN)	Mendeley	5232	Accuracy = 98.3%
Yimer et al.	2021	Xception	NIH+Jimma	11,716	Accuracy = 97.3%
Li et al.	2021	Lesion-Aware Convolutional Neural Network (LACNN)	NIH ChestX-ray14	112,120	AUROC = 0.888
Kabiraj et al.	2022	EfficientNet (CX-Ultranet)	NIH ChestX-ray14	112,120	Accuracy = 88%
Chen et al.	2023	EfficientNet-B5, CoAtNet-0-rw	NIH ChestX-ray14	112,120	AUROC = 0.842
Nahiduzzaman et al.	2023	ChestX-Ray6 CNN	Kaggle – Chest X-ray Pneumonia COVID-19 Image Data CollectionNeo X-ray	9,514	Accuracy = 97.94%
Gu et al.	2024	DenseNet121 + Attention	CheXpert	224000	AUROC = 0.898
Ucan et al.	2025	EfficientNet-B7 + Coordinate Attention	NIH ChestX-ray14	112,120	AUROC = 0.8309
Reddy & Patil	2025	ResNet-50 (Multi-task)	VinDr-CXR	15000	Accuracy = 0.989, IoU = 0.851

3 DATASET DISEASE DESCRIPTION

Atelectasis signifies a decrease in lung volume due to an obstruction of the airway that leads to alveolar collapse. Radiographically, atelectasis appears as a homogeneous opacity with signs of volume loss, which include the mediastinal shift and fissure displacement as in Figure 1 [15], [16].

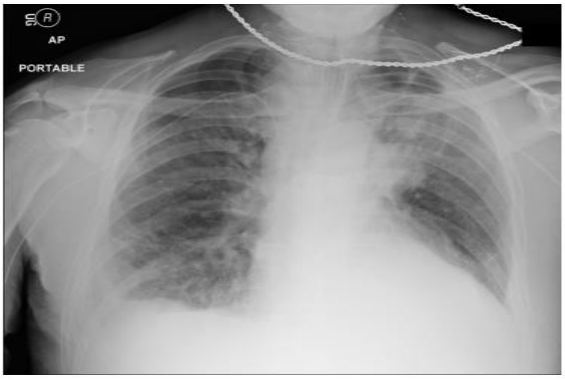

Figure 1: Example of atelectasis in a chest X-ray image (NIH ChestX-ray14 dataset).

Cardiomegaly is diagnosed if the cardiothoracic ratio exceeds 0.5 on a PA chest radiograph [15]. It manifests as an enlarged cardiac silhouette [16] as in Figure 2.

Pleural effusion is the accumulation of fluid in the pleural cavity [15]. It manifests as a basal opacity with blunting of the costophrenic angle and a meniscus sign [16] as in Figure 3.

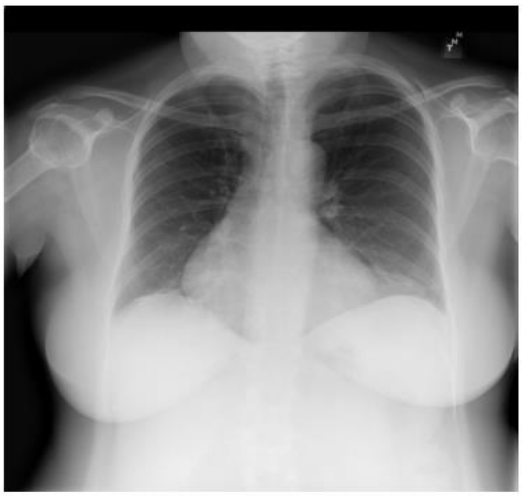

Figure 2: Example of cardiomyopathy in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

Pulmonary infiltration is an abnormal pneumonic pattern resulting from infection, inflammation, or malignancy [15]. Pulmonary infiltration is characterized by the presence of patchy or reticular opacities [16] as in Figure 4.

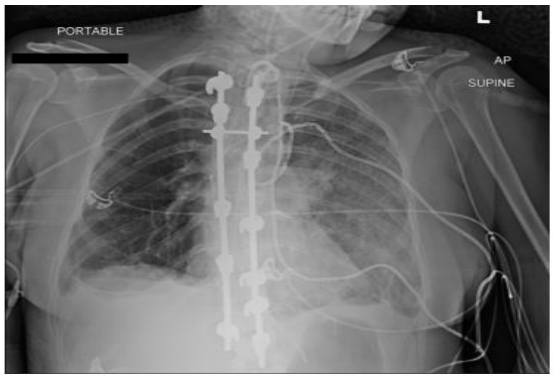

Figure 3: Example of effusion in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

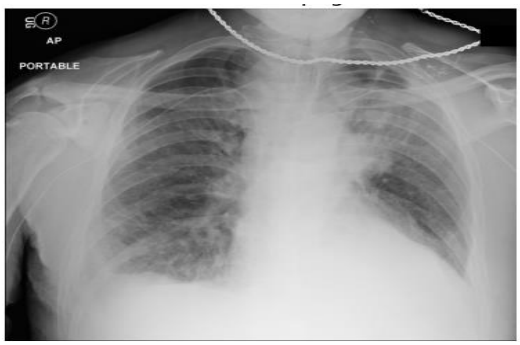

Figure 4: Example of infiltration in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

Pulmonary masses are abnormal opacities that are greater than 3 cm in diameter, with high suspicion for malignancy . Pulmonary masses are characterized by rounded or irregular opacities as in Figure 5 [15], [16].

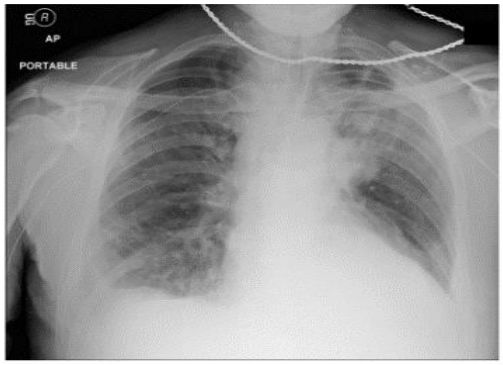

Figure 5: Example of mass in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

A pulmonary nodule is a rounded opacity less than 3 cm in size [16]. Evaluation is based on size, growth, and patient risk factors [15] as in Figure 6.

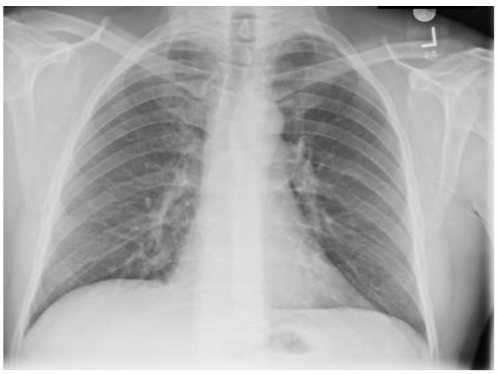

Figure 6: Example of nodule in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

Pneumonia is a pulmonary parenchymal infection causing alveolar filling with exudates [15]. It is usually visualized as lobar consolidation with possible presence of air bronchograms [16] Figure7.

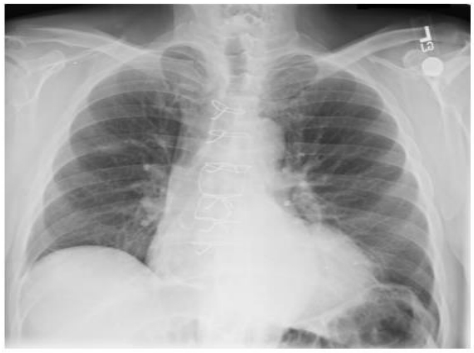

Figure 7: Example of Pneumonia in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

Pneumothorax is the accumulation of air in the pleural space leading to lung collapse [15]. It is radiographically seen as a radiolucent area in the periphery of the chest, without any vascular markings and a pleural line [16] as in Figure 8.

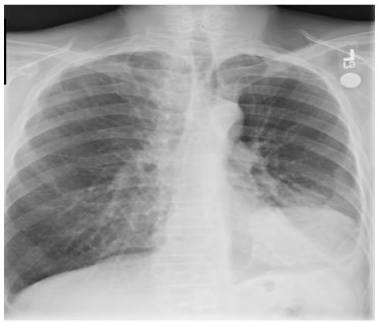

Figure 8: Example of Pneumothorax in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

Consolidation is the replacement of air within the alveoli by fluid and/or cellular material [15]. It is radiographically seen as a homogeneous opacity, usually with an air bronchogram sign [16] as in Figure 9.

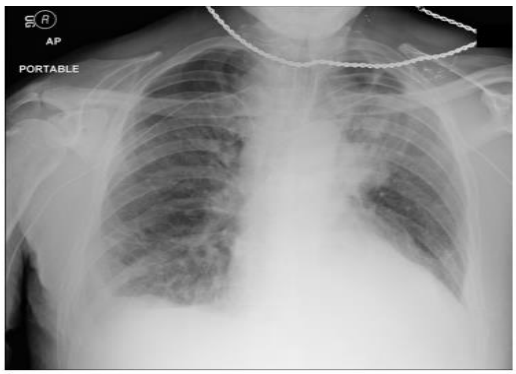

Figure 9: Example of Pulmonary Consolidation in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

Pulmonary edema is commonly due to left heart failure [15]. It shows as bilateral perihilar opacities with possible interstitial markings [16] as in Figure 10.

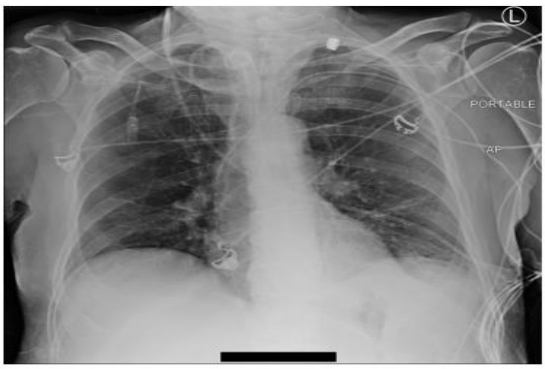

Figure 10: Example of edema in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

Emphysema is characterized by permanent destruction of alveolar walls with consequent trapping of air [15]. It appears as hyperlucent lungs with flattened diaphragms [16] as in Figure 11 [15].

Pleural thickening reflects fibrosis of the pleura [15]. It manifests as linear pleural thickening along the chest wall [16] Figure 13.

Pulmonary fibrosis is a condition of chronic interstitial scarring with reduced compliance [15]. Reticular opacities may be seen on radiographs with possible honeycombing [16] as in Figure 12.

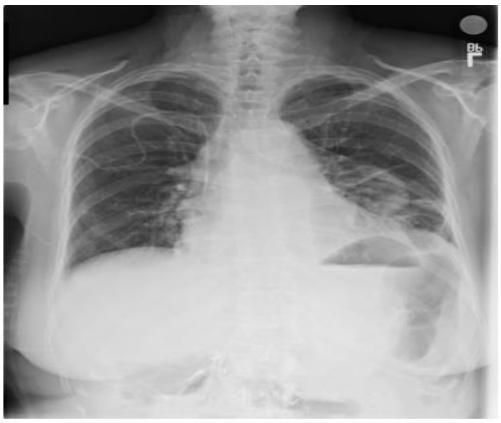

Figure 11: Example of Emphysema in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

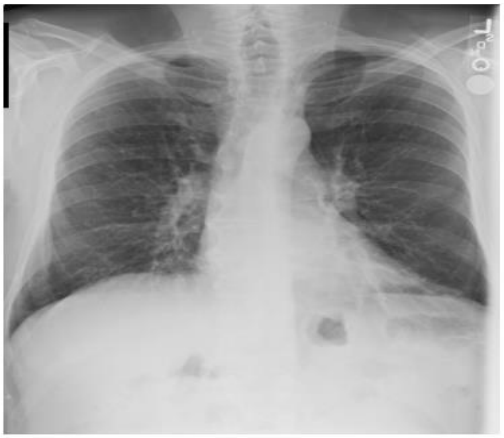

Figure 12: Example of fibrosis in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

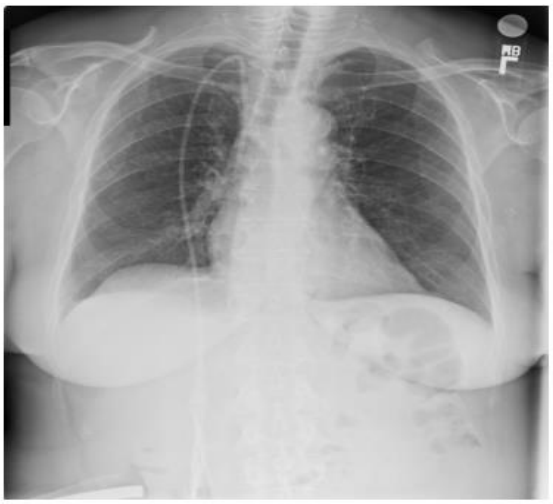

Figure 13: Example of Pleural thickening in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

Hiatal hernia is the protrusion of the stomach into the thoracic cavity [15]. It may manifest as a retrocardiac mass with an air-fluid level [16] as in Figure 14.

COVID-19 is a viral respiratory illness that can progress to acute pneumonia [15]. Chest imaging reveals bilateral peripheral ground-glass opacities [16] as in Figure15.

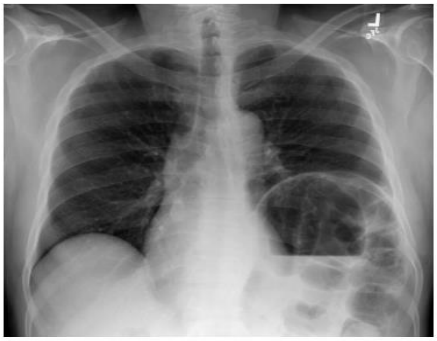

Figure 14: Example of diaphragmatic hernia in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset.

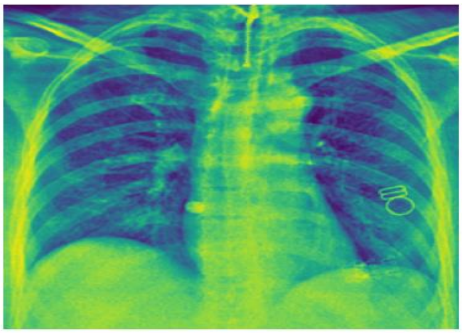

Figure 15: Example of COVID-19 in a chest X-ray image (NIH ChestX-ray14 dataset). Dataset

4 METHODS

4.1 Dataset

In this research, two sources of data were used. The first source of data is NIH ChestX-ray14 [4], which includes 15 classes, 30,000 patients, and 112,120 images. Of the 15 classes, 14 are considered pathological conditions, namely atelectasis, cardiomegaly, pulmonary congestion, pulmonary edema, pleural effusion, emphysema, pulmonary fibrosis, diaphragmatic hernia, infiltration, mass, nodule, pleural thickening, pneumonia, and pneumothorax. The second source of data, COVID-19 Radiography Dataset [1], was used to handle cases of COVID-19. The dataset included 3,616 images, all of which were obtained from Kaggle. The images were then resized from 1024 * 1024 to 456 * 456 to match the proposed models. The images were then split into 80%, 10%, and 10% for training, validation, and testing, respectively. The rationale for selecting this dataset is that it is one of the three largest datasets for chest radiography, containing images of diseases considered most prevalent. The rationale for selecting this dataset is that it is one of the three largest datasets for chest radiography, containing images of diseases considered most prevalent. To test the performance of the model, the model was also tested on another dataset, which was obtained from the VinDr-CXR database [17]. The new dataset included 4,158 images, covering ten different diseases, namely lung collapse, cardiomegaly, pulmonary congestion, pleural effusion, pulmonary fibrosis, infiltration, mass, nodule, pleural thickening, and pneumothorax. Table 2 presents a summary of the dataset employed in this work.

Table 2 : Summary of the dataset.

	Train set (80%))	Validation set (10%)	Test set (10%)
Atelectasis	9099 (10.15%)	1232 (11.61%)	1228 (10.38%)
Cardiomegaly	2170 (2.42%)	265 (2.50%)	341 (2.88%)
Effusion	10515 (11.73%)	1291 (12.16%)	1511 (12.78%)
Infiltration	15930 (17.76%)	1822 (17.16%)	2142 (18.11%)
Mass	4566 (5.09%)	550 (5.18%)	666 (5.63%)
Nodule	5030 (5.61%)	581 (5.47%)	720 (6.09%)
Pneumonia	1134 (1.26%)	145 (1.37%)	152 (1.29%)
Pneumothorax	4272 (4.76%)	386 (3.64%)	644 (5.45%)
Consolidation	3679 (4.10%)	508 (4.79%)	480 (4.08%)
Edema	1857 (2.07%)	200 (1.88%)	246 (2.08%)
Emphysema	1995 (2.22%)	211 (1.99%)	310 (2.62%)
Fibrosis	1354 (1.51%)	168 (1.58%)	164 (1.39%)
Pleural Thickening	2674 (2.98%)	322 (3.03%)	389 (3.29%)
Hernia	176 (0.20%)	20 (0.19%)	31 (0.26%)
COVID-19	2892	361	363

Percentages are calculated independently within each subset (train, validation, and test). As this is a multi-label dataset, percentages may exceed 100% when summed across classes within a subset.

4.2 Data Preprocessing

The data set was then divided into training, validation, and testing data. Images were resized according to each model's input size requirement. ResNet50 and CoAtNet-0-rw require an image size of 224x224, while EfficientNet-B5 requires an image size of 456x456 pixels. Normalization of pixel values was performed to improve stability during training. To overcome class imbalance, data augmentation was performed at different levels of intensity. Mild augmentation was performed for common diseases, while stronger augmentation strategies were used for rare disease classes to improve their representation and aid the model in learning. No augmentation was performed on validation and test data.

4.3 Data Cleaning Strategy

However, because ChestX-ray14 uses automated label extraction, inconsistencies between images and labels can arise. To address this, the disagreement-based filtering method was used.

In Round1, models were trained on the original data set, and prediction probabilities for all 15 classes were obtained. A sample was considered unreliable and removed if there was a significant difference between the model's prediction and actual label (probability greater than or equal to 0.90 for absent diseases and probability less than or equal to 0.10 for present diseases). The removed samples were capped at 2% of the training set.

402 images were removed for ResNet50, 293 images for EfficientNet-B5, and 374 images for CoAtNet-0-rw. After retraining (Round2), AUC improved from 0.82 to 0.83 for ResNet50, from 0.83 to 0.86 for EfficientNet-B5, and from 0.85 to 0.86 for CoAtNet-0-rw.

4.4 Networks and Hyperparameters

We chose ResNet-50 as a baseline model based on its popularity as a model and its frequent use as a reference model for the classification of medical images. The model is based on the principle of residual learning, where residual connections are added to the model to ensure that information is propagated to deeper layers without going through all the intermediate linear transformations. This allows the network to be made deeper without compromising its performance, as it is easier to learn residual functions compared to the whole function. This enables the creation of deeper networks [5].

The output layer of the model is modified according to the number of diseases and the multi-label classification problem. The sigmoid function is applied to each category of diseases to predict the presence of multiple diseases in the images. This model is a reference model and is used as a benchmark to compare the improvement achieved by other models with complex architectures.

- EfficientNet b5

Although many models have focused on scaling in one dimension, such as ResNet, which focused on scaling depth and augmenting it, EfficientNet introduced a new concept called Compound Scaling, which scales in a third dimension using certain ratios. This acknowledges that all dimensions are interdependent and cannot function independently. This shows the efficiency of the model, which is used in many architectures.

In addition, EfficientNet is based on a basic network called EfficientNet-B0, which uses Neural Architecture Search (NAS). EfficientNet-B1, EfficientNet-B7, and all other models have a composite parameter called φ, which determines the amount of computational resources used. The scaling of each dimension is given by the following mathematical relationships:

- Depth: $d = a\varphi$;
- Width: $w = b\varphi$;
- Resolution: $r = Y\varphi$.

In order to ensure that there is a balance between the three dimensions, the following constraint must be satisfied:

$$a * b^2 * y^2 \approx 2.$$

In addition, EfficientNet-B7, which is one of the EfficientNet architectures, achieved an accuracy of 84.4% in classifying images using the ImageNet dataset. This shows that it is more efficient than other architectures, such as ResNet-152, which makes it more efficient in medical imaging, which requires high accuracy and low computational costs [6].

The architecture of the model also includes an attention mechanism, referred to as Squeeze-and-Excitation (SE). The attention mechanism improves the representational ability of the network by allowing the model to learn the relative importance of the channel based on the contribution to the feature extraction process, thus improving the quality of the representations. The SE attention mechanism

involves two operations, namely Squeeze and Excitation. In the Squeeze operation, global spatial information is extracted from the feature map through the use of the Global Average Pooling operation, where the spatial information of each channel is compressed into a single descriptor, thus summarizing the channel's response across the entire image. The information is then propagated to the subsequent layers. In the Excitation operation, the extracted feature vector is then fed into a self-gating module consisting of two fully connected layers with ReLU activation followed by a Sigmoid activation function. The self-gating module is then used to learn the channel-wise importance weights, thus allowing channel-wise multiplication with the original feature maps, thus allowing adaptation to the content of the images [18].

CoAtNet- 0- rw. The Convolutional Neural Network (CNN) model has been found to generalize well under data-constrained conditions, whereas the generalization capability of the Transformer model is comparatively low, though it has high modeling capacity, especially in large datasets. To achieve the benefits of both the models, researchers have proposed the hybrid model, which combines the basic benefits of both the models. CNNs have high generalization and convergence capabilities due to their high inductive bias, whereas the modeling capacity of the attention model, which is comparatively low, can be utilized in large datasets. The combination of convolutional and attention layers provides a better balance in generalization and modeling capacity. The proposed model is based on two main concepts, which are discussed in the following sections. The first concept is the combination of depthwise separable convolution and relative attention in a unified framework, which is called Relative Attention. It combines the transitional symmetry of the convolutional model and the adaptive flexibility of relative attention. The second concept is the vertical ordering of convolutional and attention layers. Initially, the convolutional layers are used, followed by the addition of attention layers in the deeper layers. The Squeeze-and-Excitation (SE) mechanism is used to recapture the responses of the channels according to their importance. The spatial information is compressed using the Global Average Pooling technique, and the channel weights are generated by the self-directed mechanism using fully connected layers. The feature maps are then reweighted. In the output stage, the Sigmoid function is used to generate the probabilities of multilabel classification [7].

In all models, the last classification layer was tuned in accordance with the number of disease classes. The sigmoid activation was used in the last layer of the models for the possibility of multi-label prediction. All models were trained in a uniform experimental framework with the AdamW optimizer and an initial learning rate of 1.0e-4, a batch size of 32, and OneCycleLR scheduling. To handle class imbalance, the Asymmetric Loss (ASL) function was used. Early stopping was used with a patience of 3 and a maximum of 30 epochs. The input image size was set at 224x224 for ResNet-50 and CoAtNet-0-rw, and 456x456 for EfficientNet-B5.

4.5 Training Configuration

All models were trained with the same optimization settings for comparison. A 15-node sigmoid classifier was used for multi-label prediction, and training was performed with an adaptive optimizer and scheduled learning rate adjustment. See Table 3 for the Hyperparameters used in the experiment.

Table 3: Training configuration.

Parameter	Value
Optimizer	AdamW
Initial Learning Rate	0.0001
Batch Size	32
Scheduler	OneCycleLR
Loss Function	ASL
Early Stopping	Enabled (patience = 3)
Maximum Epochs	30
Image Size	224/456

4.6 Loss Functions

Asymmetric Loss (ASL). However, the problem of data imbalance is identified as a significant challenge in the training of multilabel classification models, owing to the imbalanced availability of disease-related data. Hence, the use of a standard loss function like BCE may not result in a balanced performance, as the BCE loss is affected by negative classes and can reduce the effectiveness of the model in detecting rare diseases. Therefore, to reduce the effects of negative classes in the BCE loss, the Asymmetric Loss (ASL) function was used in the current study. The ASL function is a variant of the Focal Loss function that minimizes the effects of easily misclassified negative classes. The ASL function uses different parameters to reduce the effects of negative classes and improve the training of the model. This characteristic of the ASL function makes it more appropriate for handling sensitive and imbalanced medical data [19].

4.7 Other Training Tricks

We also used several other training strategies to improve the robustness and accuracy of the model. The images were standardized to 456×456 pixels, with ImageNet statistics (mean, standard deviation). Mild augmentations were done for all images, while strong augmentations were done for six images. To handle the imbalance problem, WeightedRandomSampler was used, where the probability of selecting images from the minority class was higher. Multiple models, namely EfficientNet-B5, CoAtNet 0rw, ResNet-50, were trained for comparison. The models were trained with ImageNet pre-trained weights, where the Asymmetric Loss (ASL) function was used. The model was then fine-tuned with AdamW, where differential learning rates were used for the column, spine, and classification head. The model was also fine-tuned with OneCycleLR for 30 epochs. Mixed Precision Training (AMP) was also used, where Exponential Moving Average (EMA) was implemented. The threshold was fixed at 0.5, while early stopping was implemented with patience equal to three, where two macro-scale AUROC, F1-macro, and F1-macro are used.

4.8 Performance Evaluation Metrics

A set of measures for evaluating the performance of the proposed models has been utilized in a manner appropriate to the nature of the study. The measures include multi-label classification, area under the receiver operating characteristic curve (AUROC), receiver operating characteristics, accuracy, and macro-averaged F1 score (Macro F1-Score). Since the classes of diseases are imbalanced, it is not appropriate to use a single measure to evaluate the performance of the models in discriminating between different disease conditions.

4.9 Hybrid X CNN–SVM Model

The method of using the combination of convolutional neural networks and a support vector machine classifier was adopted. This combination increases the performance of the model. After the CNN models had been trained, the weights of the last layer of the models were adopted as a feature extractor. The global mean pooling operation was performed on the feature vector of each image. The features were reduced to 256 dimensions using principal component analysis. A linear SVM classifier was then trained using the One-vs-Rest method and a fixed threshold of 0.5. This was similar to the CNN models. The performance metrics adopted for the CNN models were adopted for this model as well.

The motivation for using this model was that CNNs are best for probabilistic outputs, while SVMs are best for maximizing the margin between classes. The margin-based classifier could be more useful in imbalanced multi-label medical datasets because they could provide better decision stability and robustness in separating minority classes. The performance of the hybrid model was evaluated in order to determine whether margin-based classification could be useful in improving the performance of the model over the CNN model.

5 RESULTS

The Tables 4, 5, 6 show the performance of the proposed models in tests using Accuracy, AUROC, Macro-F1, Macro-Precision, and Macro-Recall metrics. In the test, the CNN model recorded an overall AUROC of 0.862 and a micro accuracy of 0.950. In terms of Macro-F1, the CNN recorded 0.278, with Macro-Precision at 0.530 and Macro-Recall at 0.242 when using a fixed threshold of 0.5. After the application of the SVM hybridization process, the micro accuracy was slightly better at 0.952, while the AUROC fell to 0.836 and Macro-F1 fell to 0.224, with a low Macro-Recall at 0.178.

6 DISCUSSION

In addition to the findings reported in Tables 5, 7, and 8, macro-level metrics offer further understanding of model performance. Although the proposed CNN model demonstrated good micro-accuracy and AUROC performance, as reported in Table 5, its lower macro-recall value indicates that it is less sensitive to minority disease classes. Such findings are consistent with those of other multi-label chest X-ray datasets that suffer from severe class imbalance and variable disease class prevalence.

According to the per-class performance reported in Table 5, it is found that misclassification occurred between diseases that are radiographically similar, such as infiltration and pneumonia. These two diseases commonly display overlapping air-space opacities and patchy consolidation. Such differentiation is challenging even for clinicians. In addition, poor quality of labels from the NIH

ChestXray14 dataset may be another reason for the observed overlapping performance.

From the comparison reported in Tables 7 and 8, it is found that although the proposed hybrid CNN-SVM model demonstrated improved micro-accuracy, its lower macro-F1 and AUROC values indicate that it is less sensitive to minority disease classes under a fixed decision threshold of 0.5. As presented in Table 9, the models performed exceedingly well in the classification process in the presence of external test data.

Table 4: Comparisons of Accuracy among different models.

	EfficientNet-b5	CoAtNet-0-rw	ResNet50
Atelectasis	0.900	0.901	0.900
Cardiomegaly	0.972	0.972	0.972
Effusion	0.891	0.897	0.890
Infiltration	0.821	0.827	0.824
Mass	0.948	0.947	0.947
Nodule	0.938	0.940	0.941
Pneumonia	0.988	0.988	0.988
Pneumothorax	0.950	0.950	0.947
Consolidation	0.959	0.961	0.960
Edema	0.980	0.980	0.980
Emphysema	0.976	0.975	0.976
Fibrosis	0.987	0.987	0.987
Pleural Thickening	0.968	0.968	0.968
Hernia	0.998	0. 998	0. 998
COVID-19	1.000	1.000	1.000
Overall	0.951	0.952	0.951

Table 5: Comparisons of AUROCs among different models.

	EfficientNet-b5	CoAtNet-0-rw	ResNet50
Atelectasis	0.837	0.831	0.895
Cardiomegaly	0.904	0.918	0.882
Effusion	0. 889	0.890	0.878
Infiltration	0.755	0.755	0.734
Mass	0.891	0.887	0.847
Nodule	0.825	0.806	0.754
Pneumonia	0.772	0.773	0.748
Pneumothorax	0.902	0.907	0.872
Consolidation	0.810	0.814	0.794
Edema	0.903	0.905	0.892
Emphysema	0.955	0.950	0.925
Fibrosis	0.813	0.826	0.805
Pleural Thickening	0.814	0.803	0.770
Hernia	0.869	0.869	0.828
COVID-19	1.000	1.000	1.000
Overall	0.862	0.836	0.835

Table 6: All metrics.

All Metrics	CNN (CoAtNet)	Hybrid (CoAtNet + svm)
Accμ	0.950	0.952
Macro F1	0.2779	0.2244
Macro Precision	0.5296	0.4365
Macro Recall	0.2421	0.1784
AUROC (Macro)	0.862	0.836
AUROC (Micro)	0.9067	0.8906

Table 7: Comparisons of accuracy among different models of (CNN vs HYPRID).

	EfficientNet-b5	CoAtNet-0-rw	ResNet50
	Hybrid Cnn	Hybrid Cnn	Hybrid Cnn
Atelectasis	0.900 0.888	0.901 0.894	0.900 0.895
Cardiomegaly	0.972 0.973	0.972 0.973	0.972 0.973
Effusion	0.891 0.877	0.897 0.887	0.890 0.886
Infiltration	0.821 0.805	0.827 0.816	0.824 0.807
Mass	0.948 0.944	0.947 0.945	0.947 0.946
Nodule	0.938 0.932	0.940 0.935	0.941 0.939
Pneumonia	0.988 0.988	0.988 0.988	0.988 0.988
Pneumothorax	0.950 0.948	0.950 0.947	0.947 0.943
Consolidation	0.959 0.957	0.961 0.957	0.960 0.959
Edema	0.980 0.980	0.980 0.980	0.980 0.979
Emphysema	0.976 0.978	0.975 0.977	0.976 0.977
Fibrosis	0.987 0.987	0.987 0.987	0.987 0.987
Pleural Thickening	0.968 0.968	0.968 0.968	0.968 0.968
Hernia	0.998 0.998	0.998 0.998	0.998 0.998
COVID-19	1.000 1.000	1.000 1.000	1.000 1.000
Overall	0.951 0.948	0.952 0.950	0.9518 0.949

Table 8: Comparisons of AUROCs among different models of (CNN vs HYPRID).

	EfficientNet-b5	CoAtNet-0-rw	ResNet50
	Cnn hybird	Cnn hybrid	Cnn hybrid
Atelectasis	0.837 0.815	0.831 0.815	0.895 0.790
Cardiomegaly	0.904 0.898	0.918 0.888	0.882 0.850
Effusion	0.889 0.878	0.890 0.877	0.878 0.868
Infiltration	0.755 0.740	0.755 0.750	0.734 0.727
Mass	0.891 0.870	0.887 0.867	0.847 0.834
Nodule	0.825 0.817	0.806 0.800	0.754 0.749
Pneumonia	0.772 0.762	0.773 0.705	0.748 0.727
Pneumothorax	0.902 0.882	0.907 0.887	0.872 0.853
Consolidation	0.810 0.762	0.814 0.779	0.794 0.768
Edema	0.903 0.869	0.905 0.866	0.892 0.842
Emphysema	0.955 0.948	0.950 0.933	0.925 0.899
Fibrosis	0.813 0.805	0.826 0.772	0.805 0.754
Pleural Thickening	0.814 0.787	0.803 0.742	0.770 0.714
Hernia	0.869 0.847	0.869 0.836	0.828 0.807
-COVID19	1.000 1.000	1.000 1.000	1.000 1.000
Overall	0.862 0.847	0.862 0.836	0.835 0.812

Table 9: AUROC metric on VinDr-CXR data.

	EfficientNet-b5	CoAtNet-0-rw	ResNet50
Atelectasis	0.750	0.764	0.722
Cardiomegaly	0.791	0.865	0.783
Effusion	0.828	0.900	0.868
Infiltration	0.697	0.742	0.699
Mass	0.703	0.832	0.779
Nodule	0.729	0.832	0.734
Pneumothorax	0.790	0.868	0.784
Consolidation	0.846	0.919	0.926
Fibrosis	0.670	0.822	0.667
Pleural Thickening	0.665	0.826	0.730
Overall	0.74	0.832	0.769

7 CONCLUSIONS

In this study, improved deep learning models have been developed to address the problem of classifying chest diseases with respect to chest X-ray images. The focus of this investigation is to address two primary challenges that are most common in medical image processing. These challenges arise in the form of heterogeneous presentation of various diseases and class imbalance in disease classes. Both these challenges affect the overall accuracy and reliability of automated diagnostic systems, making it difficult to apply these systems in real-world clinical settings.

The results obtained in this experiment show that the proposed approach is capable of delivering consistent results in terms of multiple evaluation criteria. For example, very high values have been recorded in terms of classification accuracy and AUROC, indicating that the model is capable of distinguishing between pathological and non-pathological classes. Such results suggest that the model is capable of learning relevant features related to chest diseases, ensuring consistent predictions in this context.

The primary contribution of this study is to introduce a diagnostic framework that can be used to assist medical professionals in delivering accurate results. Such a diagnostic framework is likely to be beneficial in settings where access to skilled radiologists is not feasible. Despite the fact that very promising results have been reported in this study, it is important to consider that this investigation is restricted to image-based data without considering additional clinical data. Therefore, future studies in this context should consider incorporating additional data with respect to patients to enhance the reliability and generalization capacity of this diagnostic system.

REFERENCES

[1] D. Huang, X. Lian, F. Song, H. Ma, Z. Lian, Y. Liang, T. Qin, W. Chen, and S. Wang, “Clinical features of severe patients infected with 2019 novel coronavirus: a systematic review and meta-analysis,” Annals of Translational Medicine, vol. 8, no. 9, Art. no. 576, May 2020, [Online]. Available: https://doi.org/10.21037/atm-20-2124.

[2] Y. Chen, Y. Wan, and F. Pan, “Enhancing multi-disease diagnosis of chest X-rays with advanced deep-learning networks in real-world data,” Journal of Digital Imaging, Mar. 2023, [Online]. Available: https://doi.org/10.1007/s10278-023-00801-4.

[3] F. Li et al., "Lesion-aware convolutional neural network for chest radiograph classification," Clinical Radiology, vol. 76, pp. 155.e1-155.e14, 2021, [Online]. Available: https://doi.org/10.1016/j.crad.2020.08.027.

[4] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.

[6] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, 2019, pp. 6105-6114.

[7] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in Proc. 35th Int. Conf. Neural Information Processing Systems (NeurIPS), 2021.

[8] S. Banerjee, R. Misra, M. Prasad, E. Elmroth, and M. H. Bhuyan, "Multi-diseases classification from chest X-ray: A federated deep learning approach," in Proc. Int. Conf. Artificial Intelligence (AI 2020), Lecture Notes in Computer Science, vol. 12576, Springer, Dec. 2020, pp. 3-15, [Online]. Available: https://doi.org/10.1007/978-3-030-64984-5_1.

[9] F. S. Yimer, A. W. Tessema, and G. L. Simegn, "Multiple lung diseases classification from chest X-ray images using deep learning approach," International Journal of Advanced Trends in Computer Science and Engineering, vol. 10, no. 5, pp. 2936-2946, Sep.-Oct. 2021, [Online]. Available: https://doi.org/10.30534/ijatcse/2021/021052021.

[10] Kabiraj, T. Meena, P. B. Reddy, and S. Roy, "Detection and classification of lung disease using deep learning architecture from X-ray images," in Proc. Int. Symp. Vis. Comput. (ISVC 2022), Lecture Notes in Computer Science, vol. 13598, Springer, Dec. 2022, pp. 444-455, [Online]. Available: https://doi.org/10.1007/978-3-031-20713-6_34.

[11] Md. Nahiduzzaman, Md. R. Islam, and R. Hassan, "ChestXray6: Prediction of multiple diseases including COVID-19 from chest X-ray images using convolutional neural network," Expert Systems with Applications, vol. 211, Art. no. 118576, 2023, [Online]. Available: https://doi.org/10.1016/j.eswa.2022.118576.

[12] Y. Gu, R. Shi, S. Yang, L. Yang, B. Zhang, J. Wang, X. Lu, J. Li, X. Liu, Y. Zhao, D. Yu, S. Tang, and Q. He, "Multi-disease X-ray image classification of the chest based on global and local fusion adaptive networks," Current Medical Imaging, vol. 20, Art. no. e15734056291283, 2024, [Online]. Available: https://doi.org/10.2174/0115734056291283240808045952.

[13] M. Ucan, B. Kaya, O. Aygun, M. Kaya, and R. Alhajj, "Comparison of EfficientNet CNN models for multi-label chest X-ray disease diagnosis," PeerJ Computer Science, vol. 11, Art. no. e2968, 2025, [Online]. Available: https://doi.org/10.7717/peerj-cs.2968.

[14] K. Divya Reddy and A. Patil, "CXR-MultiTaskNet: A unified deep learning framework for joint disease localization and classification in chest radiographs," Scientific Reports, vol. 15, Art. no. 32022, 2025, [Online]. Available: https://doi.org/10.1038/s41598-025-16669-z.

[15] D. Penman, S. H. Ralston, M. W. J. Strachan, and R. P. Hobson, Davidson's Principles and Practice of Medicine, 24th ed. London, UK: Elsevier, 2023.

[16] J. S. Klein and M. L. Rosado-de-Christenson, "A Systematic Approach to Chest Radiographic Analysis," in Diseases of the Chest, Breast, Heart and Vessels 2019-2022, J. Holder et al., Eds., IDKD Springer Series, Springer, 2019, [Online]. Available: https://doi.org/10.1007/978-3-030-11149-6_1.

[17] H. Q. Nguyen, K. Lam, L. T. Le et al., "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," Scientific Data, vol. 9, no. 1, Art. no. 429, Jul. 2022, [Online]. Available: https://doi.org/10.1038/s41597-022-01498-w.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132-7141.

[19] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 82-91.