Implementation of SSL-Vision Transformer (ViT) for Multi-Lung Disease Classification on X-Ray Images

Rafi Haqul Baasith[1], Theopilus Bayu Sasongko*[2], Arifiyanto Hadinegoro*[3], Uyock Saputro*[4]
** [1,2]Informatic, Computer Science, Universitas Amikom Yogyakarta, Indonesia
rahaba@students.amikom.ac.id [1], theopilus.27@amikom.ac.id [2], arifiyanto@amikom.ac.id [3], uyock@amikom.ac.id [4]

Article Info

Article history:

Received 2025-11-24
Revised 2026-01-05
Accepted 2026-01-17

Keyword:

Self-Supervised Learning,
Vision Transformer,
Multi-label Classification,
CheXpert,
X-ray.

ABSTRACT

Chest X-ray imaging is one of the most widely used modalities for lung disease screening; however, manual interpretation remains challenging due to overlapping pathological patterns and the frequent presence of multiple coexisting abnormalities. In recent years, Vision Transformer (ViT) models have demonstrated strong potential for medical image analysis by capturing global contextual relationships. Nevertheless, their performance is highly dependent on large-scale labeled datasets, which are costly and difficult to obtain in clinical settings. To address this limitation, this study proposes a Self-Supervised Learning Vision Transformer (SSL-ViT) framework for multi-label lung disease classification using the CheXpert-v1.0-small dataset. The proposed approach leverages self-supervised pretraining to learn robust and transferable visual representations from unlabeled chest X-ray images prior to supervised fine-tuning. A total of twelve clinically relevant thoracic disease labels are retained, while non-disease labels are excluded to enhance interpretability and reduce confounding effects. Experimental results demonstrate that SSL-ViT achieves a high recall of 0.73 and a peak AUC of 0.75 on the test set, indicating strong sensitivity in detecting pathological cases. Compared to the baseline ViT model, SSL-ViT exhibits a recall-oriented performance profile that is particularly suitable for screening applications, where minimizing false negatives is critical. Furthermore, Grad-CAM visualizations confirm that the model focuses on anatomically meaningful lung regions, supporting its clinical relevance. These findings suggest that SSL-enhanced Vision Transformers provide a robust and effective solution for multi-label chest X-ray screening tasks.

I. INTRODUCTION

Lung diseases represent one of the leading causes of mortality worldwide, with the prevalence continuing to increase annually[1]. According to the World Health Organization (WHO) and various journal studies, diseases such as pneumonia, tuberculosis, and Chronic Obstructive Pulmonary Disease (COPD) remain a significant global health burden [2] and are projected to continue rising until 2050. Early detection plays a vital role in improving patient prognoses and reducing fatality rates. Chest radiography (X-ray) is one of the primary diagnostic modalities used for lung disease screening[3]. However, the interpretation of chest X-ray images requires considerable clinical expertise and is susceptible to human error, particularly in cases involving multiple concurrent thoracic pathologies (multi-label classification)[4].

In this study, pulmonary X-ray image classification is defined as an automated analysis procedure aimed at identifying one or more lung diseases in chest radiographs using deep learning-based approaches. Such classification belongs to the category of multi-label image classification, where a single radiograph may contain multiple disease labels. Automated decision-support systems developed from this classification approach are intended to assist clinicians in providing faster and more accurate diagnoses, and they have the potential to be integrated into clinical workflows in hospitals and healthcare facilities[5]. Convolutional Neural Networks (CNNs) have traditionally dominated medical image analysis tasks due to their strong inductive biases toward local spatial features.

However, CNN-based approaches often rely on hierarchical receptive fields and local convolution operations, which may limit their ability to capture long-range global dependencies across an entire image. In chest X-ray analysis, pathological patterns frequently span multiple anatomical regions, making global contextual understanding particularly important. To address these limitations, this study adopts the Vision Transformer (ViT) architecture[6] (ViT) a transformer-based model originally developed for natural language processing and later adapted to computer vision. Unlike CNNs, ViT represents an image as a sequence of non-overlapping patches and processes them using self-attention mechanisms. This design enables ViT to model global relationships between distant image regions, which is particularly beneficial for detecting diffuse or overlapping lung abnormalities. Recent studies have demonstrated that ViT achieves competitive or even superior performance compared to CNNs in large-scale image classification challenges such as ImageNet. Chen et al[7] further evaluated fine-tuned ViT models for COVID-19 detection using chest radiographs and benchmarked them against EfficientNet, MViT, and EfficientViT, utilizing a public dataset comprising 3,616 COVID-19 samples, 10,192 normal images, 6,012 lung opacity cases, and 1,345 pneumonia images. Their results indicated that the ViT-based model achieved the highest accuracy of 95.79% in four-class classification and 99.57% in three-class classification, with an AUC of 0.9993 for the COVID-19 category. Ko et al[8]. investigated the impact of six optimization algorithms on three ViT architectures for detecting seven pulmonary diseases from a dataset of 19,003 chest X-ray images. Optimizers based on Adam, particularly RAdam and NAdam, produced the best performance. FastViT with NAdam achieved the highest accuracy of 97.63% under imbalanced conditions, while RAdam performed best on balanced datasets with 95.87% accuracy. Although the models effectively recognized Normal and Tuberculosis classes, they struggled with minority diseases such as MERS and SARS.

This work underscores the importance of selecting suitable optimization strategies and addressing data imbalance to improve ViT-based pulmonary disease classification. Marikkar et al[9]. introduced LT-ViT, a lightweight Vision Transformer architecture enhanced with label tokens for multi-label chest X-ray classification. Unlike prior methods such as C-Tran and Query2Label, LT-ViT enables direct interaction between label tokens and image tokens through cross- and self-attention mechanisms, effectively capturing inter-label relationships and multi-scale visual features. Evaluations on NIH-CXR14 and CheXpert datasets showed improved AUC performance while adding minimal parameters to the ViT-S baseline. The model supports both random and domain-specific pre-training and enables inherent interpretability, as label tokens directly attend to pathological regions. These findings demonstrate LT-ViT as an efficient and accurate solution for multi-label medical image classification using ViTs. Despite the advantages of Vision Transformers, their performance is highly dependent on the availability of large-scale labeled datasets. In medical imaging, acquiring high-quality annotations is expensive, time-consuming, and requires expert radiologists.

Moreover, datasets such as CheXpert[10] contain a significant proportion of uncertain or missing labels, which can negatively impact fully supervised training. To mitigate these challenges, this study incorporates Self-Supervised Learning (SSL)[11] as a pretraining strategy. Self-Supervised Learning (SSL) enables models to learn meaningful visual representations from large amounts of unlabeled data by solving carefully designed pretext tasks, thereby reducing reliance on annotated samples. Beyond improving general feature learning, SSL has been shown to enhance model sensitivity by encouraging the learning of more comprehensive and inclusive feature representations, which is particularly beneficial for detecting subtle or less frequent pathological patterns in medical images. As a result, SSL-pretrained models often demonstrate improved recall, reflecting a reduced rate of false-negative predictions an essential requirement in clinical screening and diagnostic support systems.

When combined with Vision Transformers (ViT), SSL becomes especially effective, as transformer-based architectures strongly benefit from large-scale representation learning and global contextual modeling. SSL pretraining allows ViT encoders to capture long-range dependencies and latent disease-related patterns across chest X-ray images, facilitating improved sensitivity during downstream fine-tuning. This characteristic is particularly advantageous in multi-label medical classification tasks, where multiple coexisting abnormalities may present with varying visual prominence. In this study, the Vision Transformer (ViT) is implemented for multi-lung disease classification using the CheXpert-v1.0-small dataset, which consists of thousands of chest X-ray images annotated with 12 thoracic disease labels. One label represents the normal condition, while a device-related label is excluded from the classification targets. The primary objective of this research is to evaluate the effectiveness of SSL-enhanced ViT models in simultaneously identifying multiple lung diseases, with a particular emphasis on improving recall performance, and to assess their potential as a robust and recall-oriented alternative to conventional CNN-based diagnostic systems.

II. METHODS

This study was conducted through a series of systematic stages, including data collection, exploratory data analysis (EDA), preprocessing, self-supervised learning (SSL) using a Vision Transformer (ViT) model, model evaluation, and deployment. The overall workflow is illustrated in Figure 1.

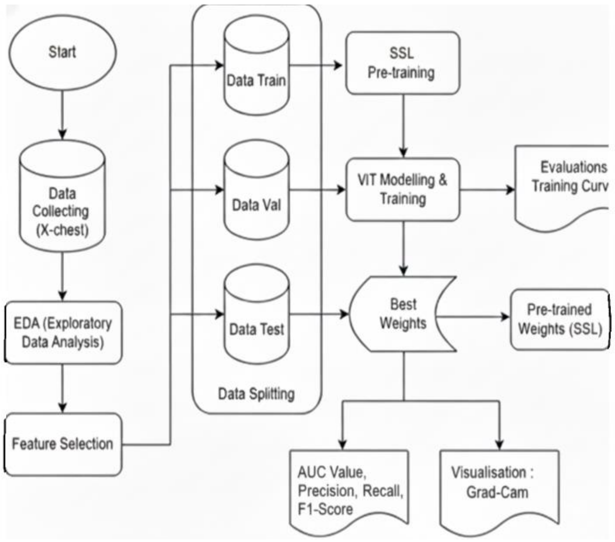
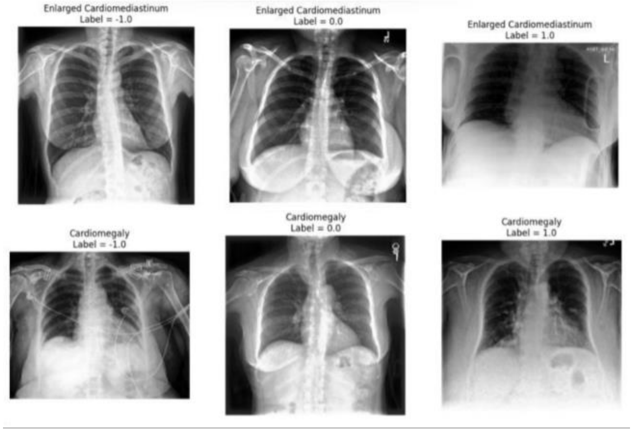
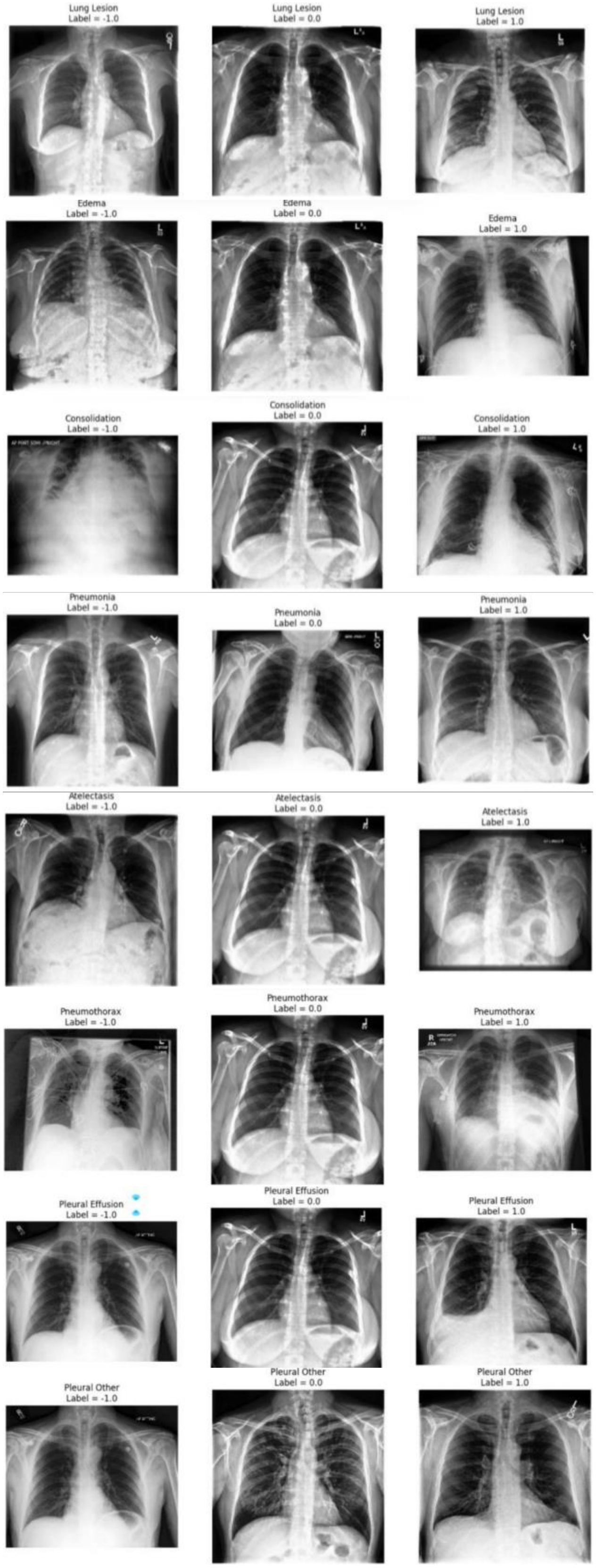

Start
Data Collecting (X-chest)
EDA (Exploratory Data Analysis)
Feature Selection
Data Train
Data Val
Data Test
Data Splitting
SSL Pre-training
ViT Modelling & Training
Evaluations Training Curv
Best Weights
Pre-trained Weights (SSL)
AUC Value, Precision, Recall, F1-Score
Visualisation : Grad-Cam

Figure 1. Pipeline research

A. Data Collecting

Data collection in this study utilized the publicly available CheXpert-v1.0-small dataset[10], which is accessible through the Kaggle platform. This dataset contains the latest updated chest X-ray images collected from patient samples between 2002 and 2017 has been widely employed in research related to lung disease detection using machine learning and deep learning approaches. The CheXpert-v1.0-small dataset consists of 223,414 data entries representing patients and their radiological examinations in the form of chest X-ray images. Each entry is accompanied by labels representing 14 lung conditions, including No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, and Support Devices. All images in this dataset are provided in JPEG format and include metadata in the form of diagnostic labels. These labels are stored in a .csv file containing diagnosis values for each corresponding image. Figure 2 illustrates the comparison of the dataset across each class, while Table 3 presents the distribution of the data.


Enlarged Cardiomediastinum Label = -1.0
Enlarged Cardiomediastinum Label = 0.0
Enlarged Cardiomediastinum Label = 1.0
Cardiomegaly Label = -1.0
Cardiomegaly Label = 0.0
Cardiomegaly Label = 1.0



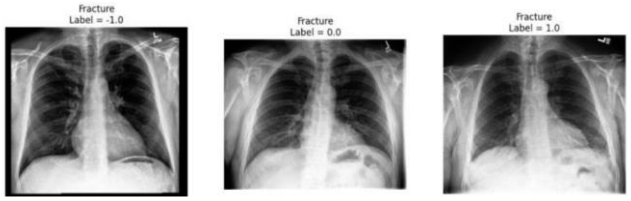


Figure 2. illustrates the comparison of the dataset across each class

TABLE 1.
DISTRIBUTION OF THE DATA ACROSS EACH CLASS

No	Diseases	Positive {1.0}	Negative {0.0}	Uncertain {-1.0}
1	No Finding	22419	196	0
2	Enlarged Cardiomediastinum	10907	21763	12403
3	Cardiomegaly	27068	11282	8087
4	Lung Opacity	105707	6707	5598
5	Lung Lesion	9187	1503	1488
6	Edema	52291	20915	12984
7	Consolidation	14816	28298	27742
8	Pneumonia	6047	3025	18770
9	Atelectasis	33456	1482	33739
10	Pneumothorax	19456	56567	3145
11	Pleural Effusion	86254	35563	11628
12	Pleural Other	3524	549	2653
13	Fracture	9040	2746	642
14	Support Devices	116108	6264	1079

Based on the visual analysis of the comparisons presented in Figure 3 and Table 1, including images with labels –1, 0, and 1 across all disease categories, it is observed that images labeled –1 tend to exhibit higher similarity to those labeled 1 (positive) than to those labeled 0 (negative). In many instances, images assigned the –1 (uncertain) label contain subtle abnormal patterns resembling early pathological indicators. Several diseases, such as Edema, Consolidation, Lung Opacity, and Atelectasis, present overlapping radiological characteristics, making it difficult to distinguish between labels –1 and 1, even for professional radiologists. Similar conditions are found in diseases such as Cardiomegaly and Enlarged Cardiomediastinum, where cardiac enlargement frequently correlates with mediastinal widening, resulting in uncertain labels due to the interrelated nature of the conditions. These findings reinforce the argument that, in clinical practice, the –1 (uncertain) label more closely represents a weakly positive indication rather than a truly neutral or negative state. This study adopts the U-ones approach, in which the –1 label is treated as positive.

This decision is motivated by the visual evidence indicating that images labeled –1 show closer resemblance to positive cases, making it more consistent to handle them as weak positives. While the U-Ones strategy is adopted in this study to handle uncertain labels in the CheXpert dataset, it is important to acknowledge the potential biases associated with this approach. Treating uncertain labels as positive instances may introduce a tendency toward overestimation of disease prevalence, potentially increasing the false positive rate and biasing the model toward higher sensitivity. However, prior studies have shown that, in screening-oriented medical applications, this bias is often acceptable when the primary objective is to minimize false negatives, which pose greater clinical risks than false positive[12].

Alternative strategies such as U-Zeros, which treat uncertain labels as negative, may reduce false positives but risk suppressing subtle pathological patterns, particularly in diseases with ambiguous radiographic manifestations. This can lead to systematic under-detection and degraded recall, especially for conditions with overlapping visual characteristics such as Edema and Consolidation[13]. Another approach, soft labeling, assigns probabilistic values to uncertain labels and has been explored to mitigate hard decision bias; however, it introduces additional complexity in optimization and requires careful calibration to ensure stable training, which may not be feasible in all practical settings[14]. Given these considerations, the U-Ones strategy is selected as a deliberate design choice aligned with the screening-oriented goal of this study. By prioritizing sensitivity and encouraging the model to learn inclusive representations of pathological patterns, U-Ones supports early detection scenarios where uncertain findings should prompt further clinical evaluation rather than dismissal. Nonetheless, future work will investigate comparative analyses across uncertainty-handling strategies, including U-Zeros and soft-labeling, to further assess their impact on model calibration, bias, and generalization performance.

B. Feature Selection

The initial step involves selecting only the disease-related labels that are directly relevant to the objective of this study, which focuses on multi-label lung disease detection from chest X-ray images. Non-disease labels such as Sex, Age, Frontal/Lateral, AP/PA, No Finding, and Support Devices are deliberately excluded from the learning process. This exclusion is based on the methodological consideration that such labels either represent demographic or acquisition metadata, or correspond to non-pathological conditions, and therefore do not contribute to the extraction of radiological features associated with pulmonary abnormalities. From a clinical and radiological perspective, lung disease manifestations in chest X-ray images are determined by pathological changes in pulmonary structures rather than by patients' personal information or imaging acquisition parameters. Several prior studies on the CheXpert dataset have adopted a similar label selection strategy, focusing exclusively on disease-related findings to improve model interpretability and clinical relevance while reducing potential confounding factors. By excluding non-disease labels, the model is encouraged to learn visual representations that are directly attributable to pathological patterns rather than spurious correlations. Furthermore, the exclusion of the No Finding label is consistent with established practices in multi-label chest X-ray classification, as this label represents the absence of disease rather than a specific pathological category and may introduce ambiguity in multi-label learning settings. Similarly, labels related to imaging devices and acquisition views are removed to ensure that the learned features are not biased toward non-anatomical artifacts, which has been shown to negatively affect generalization performance in medical imaging models[15]. As a result of this label selection process, a total of 12 disease labels are retained for model training and evaluation: Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, and Fracture. This selection aligns with the primary goal of capturing clinically meaningful thoracic abnormalities while maintaining transparency and reproducibility in the experimental design. By clearly defining the criteria for disease label inclusion, the proposed methodology ensures that the resulting model performance can be reliably interpreted and fairly compared with prior CheXpert-based studies.

C. Splitting Dataset

After the data pre-processing stage is completed, the next step involves data splitting to ensure that the model is trained, validated, and tested effectively. In this study, the dataset is not only divided into training and testing subsets as in conventional approaches but also includes a validation subset as an essential component in the deep learning training process. The dataset is partitioned into 60% for the training set, 20% for the validation set, and 20% for the testing set[16]. The training data are utilized to adjust the model weights in order to learn the patterns from the X-ray images, while the validation data are employed to monitor model performance during training and to prevent overfitting. The testing data are used exclusively to evaluate the model’s generalization capability on unseen samples. With this well-controlled data separation, the training process of the Vision Transformer can be objectively assessed, ensuring reliable and trustworthy performance results. The dataset was divided into three subsets, resulting in 134,048 data instances in the training set and 44,683 data instances each in the validation and test sets.

D. Image Transformations & Data Augmentation

The image transformation process is conducted to meet the input requirements of the Vision Transformer (ViT) model, which operates with a fixed image resolution. Initially, a cropping operation is applied to enlarge the lung region, ensuring that the model focuses on the most relevant area. The images are then resized to 224×224 pixels, normalized, and converted into tensor format to comply with the ViT input specifications. Furthermore, each data frame undergoes a specific transformation procedure. To enhance the model’s robustness against data variability and to reduce the risk of overfitting, data augmentation is applied to the training set. Several augmentation techniques are listed in Table 2.

TABLE 2.
DATA AUGMENTATION TECHNIQUES

No	Data Augmentation	Values
1	Random Resized Crop	0.85-1.0
2	Random Horizontal Flip	True
3	Random Rotation	10
4	Color Jitter	Brightness = 0.15, contrast = 0.15

To improve the model’s generalization capability and increase robustness against variations in image data, several augmentation techniques are applied during training. Random Resized Crop is utilized by randomly cropping a portion of the image with a scale range of 0.85–1.0. A Random Horizontal Flip[17] is performed enhancing sensitivity to left-right orientation differences. Random Rotation up to ±10 degrees is applied to account for slight rotational changes that may occur during image acquisition. Additionally, Color Jitter[18] is incorporated by adjusting the image brightness and contrast by a factor of 0.15, enabling the model to remain robust under varying lighting conditions.

E. Self-Supervised Learning (SSL)

To enhance feature representations learning and reduce dependency on large amounts of labelled data, a Self-Supervised Learning (SSL)[19] stage was incorporated prior to the supervised training phase. SSL enables the model to learn meaningful and transferable representations by exploiting intrinsic structures within unlabelled data through carefully designed pretext tasks. This approach is particularly beneficial in medical imaging domains, where labelled data are often scarce, expensive, and time-consuming to obtain. In this study, SSL is applied during a pretraining phase using only the training images without label information. Multiple stochastic data augmentations are performed to generate different views of the same input image. These augmented views are then used to define a pretext task that encourages the model to learn invariant and discriminative representations. By learning from unlabelled data, the model captures low-level and high-level visual patterns that are robust to variations in illumination, orientation, and noise.

The Vision Transformer (ViT) encoder is employed as the backbone network during the SSL pretraining stage. Depending on the SSL paradigm, such as contrastive learning or masked image modelling, the encoder is optimized to either maximize agreement between different augmented views of the same image or reconstruct missing image patches from partial observations. This process allows the ViT encoder to learn contextual relationships and global dependencies within medical images more effectively than purely supervised learning approaches. After the SSL pretraining phase, the learned encoder weights are transferred to the supervised learning stage. The pretrained ViT encoder is then fine-tuned using labelled data for the target classification task. This transfer learning strategy significantly improves convergence speed, generalization performance, and robustness, particularly when the labelled dataset is limited. The integration of SSL thus strengthens the overall learning framework by providing a well-initialized representation space that enhances downstream classification accuracy and reduces overfitting.

F. ViT Modelling

The model training process in this study utilizes the Vision Transformer (ViT) architecture, which has demonstrated strong effectiveness in image processing tasks. ViT operates by dividing an input image into small patches and converting them into vector representations (embeddings) that are processed through a self-attention mechanism. This mechanism enables the model to capture global relationships among different regions of an image, allowing it to learn complex patterns in chest X-ray data. ViT model can illustrate show in Figure 3.

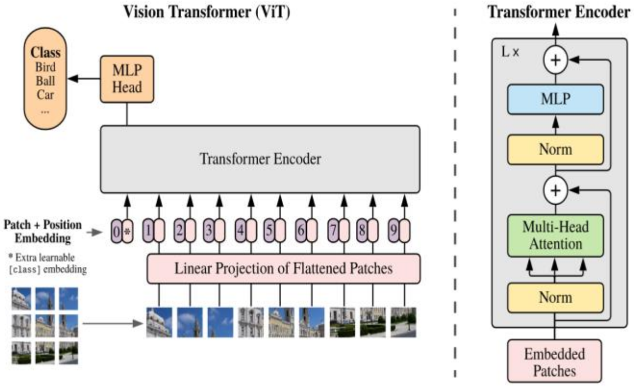


Figure 3. illustrates the model ViT

Each image patch is transformed into a vector through an embedding process and augmented with positional encoding to preserve spatial ordering. The resulting sequence of patches is then processed by a transformer encoder composed of multi-head self-attention and feed-forward network layers. The final output token serves as a global image representation for classification purposes. Through this approach, the model is expected to provide a more comprehensive understanding of pulmonary disease indicators compared with conventional CNN-based methods. A fine-tuning stage is then performed by adapting the pre-trained model weights using the pre-processed research dataset following a data-splitting procedure. In this stage, all ViT layers are re-optimized to better align with the characteristics of medical imaging data. The fine-tuning configuration is designed to support multi-label classification, given that a single chest X-ray may indicate more than one type of lung disease. A sigmoid activation function is applied in the output layer to allow the model to generate independent probability scores for each disease class. The training process employs the vit-small-patch16-224 model from HuggingFace, which was previously pre-trained by Google on the ImageNet-21k dataset. The output layer is modified to detect twelve pulmonary disease labels included in the CheXpert dataset.

G. Evaluations Metrics

Evaluation is a critical stage in the development and validation of machine learning and deep learning models, particularly in the context of multi-label medical image classification. The primary objective of evaluation is to assess how well the model performs in solving the assigned task based on validation and testing datasets that were not observed during the training process. The evaluation results are used to determine whether the model is reliable for real world applications and to compare the performance of different approaches or algorithms. In multi-label classification tasks, conventional accuracy metrics are insufficient because each sample may contain more than one correct label. Therefore, specific evaluation metrics such as precision, recall, F1-score, and Area Under the Curve (AUC) are utilized to comprehensively measure model performance[20]. Precision indicates the correctness of positive predictions made by the model, where a high precision score implies a low false-positive rate, which is essential in medical applications requiring high accuracy:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

where TP represents True Positive and FP represents False Positive. Recall (sensitivity) measures the model's ability to correctly identify all positive samples, which is crucial for minimizing false-negative predictions in disease detection:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

where FN denotes False Negative. The F1-score is the harmonic mean between precision and recall, and it is employed to establish a balance between detecting all positive cases and maintaining prediction accuracy:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

AUC describes the ability of the model to discriminate between positive and negative classes across various decision thresholds, representing multi-label classification performance more effectively. In multi-label medical classification research, such as the CheXpert benchmark, AUC is considered more representative than accuracy, F1-score, precision, and recall for several reasons. First, AUC is more robust to class imbalance, which frequently occurs in medical datasets where disease label distributions are highly uneven. While accuracy may produce misleading results by favoring majority classes, AUC evaluates class discrimination independently of the distribution of positive and negative labels[21].

Second, AUC assesses model performance across multiple decision thresholds, providing a more comprehensive evaluation compared to metrics that operate only at a single threshold, such as precision, recall, and F1-score[22]. Finally, in multi-label scenarios, each instance includes multiple binary decisions, and both macro and micro-averaged AUC offer consistent evaluation across labels despite significant disparities in positive and negative sample counts. Comprehensive evaluation provides a realistic understanding of the strengths and limitations of a model, as well as its readiness for deployment in actual medical environments. Therefore, this stage must not be overlooked and should be carefully designed to align with the specific objectives and characteristics of the addressed problem.

III. Results And Discussion

The object of this study is chest X-ray images contained in the CheXpert-v1.0-small dataset, a large-scale medical image collection widely used for pulmonary disease detection through radiological imaging. The primary objective of this research is to develop a multi-label classification framework based on the Vision Transformer (ViT) that is capable of identifying one or more lung conditions simultaneously within a single image. To enhance feature representation learning and reduce reliance on labeled data, a Self-Supervised Learning (SSL) stage is incorporated prior to the supervised fine-tuning process.In the proposed framework, the ViT encoder is first pretrained using an SSL paradigm on the training images without utilizing label information. This pretraining stage enables the model to learn robust and generalizable visual representations from chest X-ray images by exploiting intrinsic image structures through a pretext task. The SSL-pretrained encoder weights are subsequently transferred to the supervised learning stage and fine-tuned for multi-label lung disease classification. The supervised fine-tuning process employs the vit-small-patch16-224 architecture[23], which is initialized using both ImageNet-21k pretrained weights provided by Google and the SSL-pretrained representations learned from the CheXpert training data. The training configuration is implemented using the TrainingArguments framework with a learning rate of 3×10^{-5}, a cosine decay learning rate scheduler, and a warm-up ratio of 0.1. The AdamW optimizer is utilized to update model parameters. A batch size of 32 is applied consistently for both training and evaluation. The model is trained for a maximum of 10 epochs, with early stopping activated if no performance improvement is observed over three consecutive evaluation cycles. Model evaluation is conducted at the end of each epoch, and the best-performing model checkpoint is automatically restored based on the highest validation accuracy achieved during training. The fine-tuning process is carried out using the HuggingFace Trainer framework, where the model is trained on the training set and validated on the validation set, with evaluation metrics computed at each evaluation step. Upon completion of training, both the final model and the associated image processor are saved in the chexpert-vit-model directory (the directory name is configurable) to support future deployment and inference. Table 2 presents the training performance tracked across checkpoints, indicating that the entire training process 20 epochs required approximately 2 hours and 18 minutes.

TABLE 3.
TRAINING PERFORMANCE PROCESS

Epoch	Train Loss	Val Loss	Precision	Recall	F1-Score	AUC
1	1.0915	1.0739	0.2286	0.7181	0.2815	0.7165
2	1.0703	1.0729	0.2468	0.5916	0.3022	0.7181
3	1.0470	1.0370	0.2296	0.7264	0.3038	0.7317
4	1.0295	1.0349	0.2340	0.6717	0.3117	0.7334
5	1.0197	1.0352	0.2309	0.7091	0.3143	0.7395
6	1.0066	1.0328	0.2475	0.6290	0.3190	0.7417
7	0.9961	1.0309	0.2433	0.6493	0.3224	0.7432
8	1.0279	1.0991	0.2076	**0.7919**	0.2822	0.7099
9	1.0214	1.0272	0.2272	0.7316	0.3067	0.7419
10	0.9852	1.0312	0.2435	0.6648	0.3167	0.7437
11	0.9677	1.0297	0.2517	0.6592	0.3227	0.7465
12	0.9490	1.0237	0.2388	0.7121	0.3164	0.7477
13	0.9321	1.0475	0.2451	0.6732	0.3218	0.7478
14	0.9118	1.0446	0.2467	0.6877	0.3224	**0.7510**
15	0.8920	1.0564	0.2441	0.6872	0.3227	0.7509
16	0.8710	1.1194	0.2479	0.6794	0.3252	0.7496
17	0.8539	1.1616	0.2498	0.6631	0.3259	0.7493
18	0.8365	1.2135	0.2514	0.6524	0.3284	0.7492
19	0.8270	1.2717	0.2515	0.6377	0.3304	0.7491
20	0.8187	1.2772	0.2521	0.6436	0.3299	0.7484

Table 3 presents the training performance of the SSL-ViT model across 20 epochs. The results demonstrate stable convergence, high recall performance, and a peak AUC of 0.7510, highlighting the effectiveness of SSL pretraining in enhancing model sensitivity for multi-label chest X-ray classification.

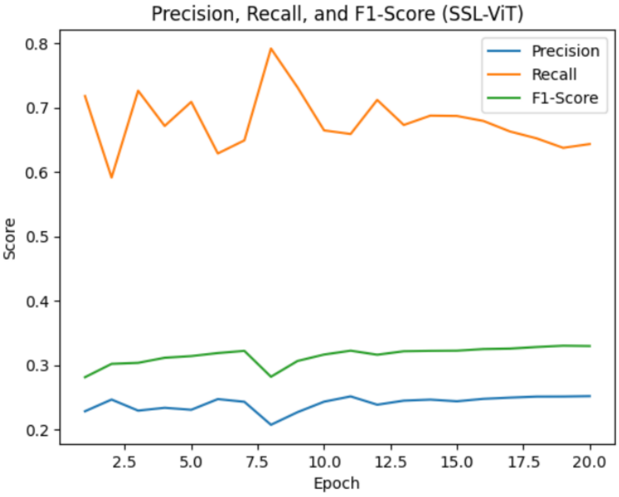


Figure 4. Precision, Recall, and F1-Score (SSL-ViT)

Figure 4 The experimental results reveal a consistent recall-oriented behavior throughout the training process. Specifically, the recall values remain relatively high, ranging approximately from 0.63 to 0.79 across epochs, indicating the

model's strong sensitivity in identifying positive lung disease cases. In contrast, precision remains comparatively low but stable, fluctuating within the range of 0.22 to 0.25, while the F1-score demonstrates a gradual and stable improvement as training progresses. This performance pattern is highly consistent with the theoretical characteristics of Self-Supervised Learning (SSL), which encourages the learning of broad and inclusive feature representations rather than highly restrictive decision boundaries. Such recall-dominant performance is particularly relevant in chest X-ray screening applications, where false-negative predictions may lead to missed diagnoses and delayed clinical intervention. From a clinical perspective, false negatives are considerably more critical than false positives, as undetected pathological conditions pose greater risks to patient outcomes. Therefore, although the SSL-ViT model sacrifices precision to some extent, its ability to consistently achieve high recall underscores its suitability for medical screening and decision-support systems, where maximizing sensitivity is a primary objective. These findings further confirm the recall-oriented nature of the SSL-ViT framework and highlight its practical relevance in safety-critical medical imaging tasks. The best-performing trained model was assessed using Area Under the Curve (AUC) metrics, as illustrated in Figure 5.

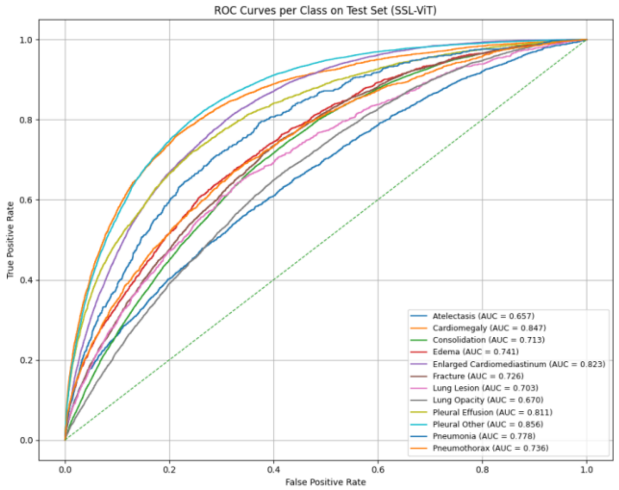


Figure 5. ROC Curves all label in test set

Figure 5 presents the Receiver Operating Characteristic (ROC) curves for each thoracic disease class evaluated on the test set using the SSL-ViT model. Overall, the ROC curves demonstrate that the proposed SSL-ViT framework achieves robust discriminative performance across multiple disease categories, with Area Under the Curve (AUC) values consistently exceeding random chance for all classes. Several disease categories exhibit strong classification capability, notably *Pleural Other* (AUC = 0.856), *Cardiomegaly* (AUC = 0.847), *Enlarged Cardiomediastinum* (AUC = 0.823), and *Pleural Effusion* (AUC = 0.811). These conditions typically present distinctive global structural or intensity patterns in chest X-ray images, which are effectively captured by the transformer-based representations learned through self-supervised pretraining. Moderate performance is observed for classes such as *Pneumonia* (AUC = 0.778), *Edema* (AUC = 0.741), *Pneumothorax* (AUC = 0.736), *Fracture* (AUC = 0.726), and *Consolidation* (AUC = 0.713). The ROC curves of these classes indicate a favorable trade-off between sensitivity and specificity, suggesting that SSL-ViT can learn clinically relevant features despite inter-class visual overlap and label ambiguity commonly found in chest X-ray datasets. Lower AUC values are observed for *Atelectasis* (AUC = 0.657), *Lung Opacity* (AUC = 0.670), and *Lung Lesion* (AUC = 0.703). These findings are consistent with prior studies, as these conditions often exhibit subtle or diffuse radiographic patterns that are challenging to distinguish even for human experts. Nevertheless, the ROC curves for these classes remain substantially above the diagonal baseline, confirming that the model retains meaningful discriminative power. Importantly, the ROC characteristics align with the study's primary objective of maximizing recall for screening-oriented applications. The SSL-ViT model demonstrates strong true positive rates at relatively low false positive rates across most classes, which is particularly desirable in clinical screening scenarios where false negatives are more critical than false positives. This behavior supports the suitability of self-supervised pretraining for enhancing sensitivity in multi-label chest X-ray classification tasks. The evaluation results obtained using the classification report on the test set demonstrate consistent misclassification patterns across several labels, as illustrated in Table. 4.

TABLE 4.
CLASSIFICATION REPORT OF SSL-VIT

Disease Class	**Precision**	**Recall**	**F1-score**	**Support**
Enlarged Cardiomediastinum	0.07	0.58	0.13	2,061
Cardiomegaly	0.36	0.70	0.47	5,367
Lung Opacity	0.62	0.70	0.66	21,063
Lung Lesion	0.08	0.71	0.14	1,843
Edema	0.46	0.76	0.57	10,505
Consolidation	0.12	0.69	0.20	2,969
Pneumonia	0.05	0.59	0.09	1,146
Atelectasis	0.21	0.75	0.32	6,689
Pneumothorax	0.21	0.74	0.33	3,919
Pleural Effusion	**0.68**	**0.79**	**0.73**	17,303
Pleural Other	0.04	0.64	0.08	710
Fracture	0.08	0.63	0.14	1,769
Micro Average	0.28	0.73	0.40	75,344
Macro Average	0.25	0.69	0.32	75,344
Weighted Average	0.46	0.73	0.53	75,344

The class-wise classification report indicates that the proposed SSL-ViT model consistently achieves high recall across most thoracic disease categories, with recall values ranging from 0.58 to 0.79. Notably, *Pleural Effusion* demonstrates the strongest overall performance, achieving the highest precision (0.68), recall (0.79), and F1-score (0.73), suggesting that the model effectively captures its distinctive radiographic patterns. Diseases such as *Edema*, *Lung Opacity*, *Atelectasis*, and *Pneumothorax* also exhibit strong recall performance (≥0.70), confirming the model's sensitivity in identifying clinically relevant abnormalities. Conversely, lower precision is observed for classes with subtle or overlapping visual characteristics, including *Pleural Other*, *Pneumonia*, and *Lung Lesion*. This trade-off reflects the model's emphasis on sensitivity, which is desirable in screening-oriented applications. Overall, the micro-averaged recall of 0.73 further highlights the effectiveness of self-supervised pretraining in enhancing detection sensitivity across multiple disease categories. The observed performance profile high recall with moderate precision is well aligned with clinical screening requirements, where false negatives pose a greater risk than false positives. The Grad-CAM heatmap demonstrates that the SSL-ViT model primarily focuses on the bilateral lower lung regions in Figure 6, with pronounced activation observed along the right lung field. These regions are commonly associated with radiographic manifestations of lung opacity, including alveolar infiltration and increased parenchymal density. Importantly, the model's attention is largely confined within anatomically relevant pulmonary areas, rather than being distracted by non-diagnostic regions such as the background or image borders.

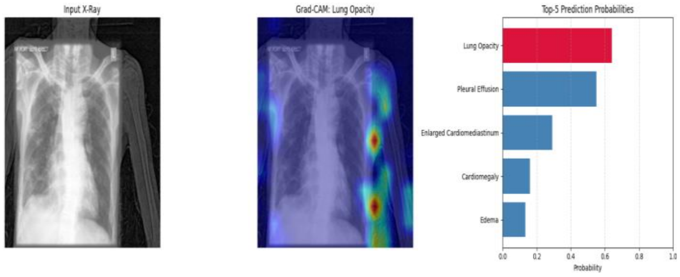


Figure 6. Grad-Cam Result of SSL-VIT

The observed Grad-CAM patterns reflect one of the key advantages of self-supervised learning in medical imaging: the ability to learn robust and transferable visual representations without relying solely on labeled data. By pretraining on unlabeled chest X-ray images, the SSL-ViT model develops a global understanding of lung anatomy, which translates into focused and clinically interpretable attention maps during downstream classification. From a screening perspective, this behavior is particularly desirable. The model's emphasis on lung parenchymal regions supports its strong recall performance reported in earlier experiments, reinforcing its suitability for early detection scenarios where minimizing false negatives is critical. The comparison results with the baseline ViT model without self-supervised learning are presented in Table 5.

TABLE 5.
COMPARISON VIT VS SSL-VIT

Model	**Precision**	**Recall**	**F1-Score**	**AUC**
ViT-Baseline	0.66	0.49	0.56	0.77
SSL-ViT (proposed)	0.28	**0.73**	0.40	0.75

Table 5 presents a comparative evaluation between the baseline Vision Transformer (ViT) model and the proposed Self-Supervised Learning Vision Transformer (SSL-ViT) model. The results highlight a clear trade-off between precision and recall across the two approaches. The baseline ViT model achieves a higher precision (0.66) and moderate recall (0.49), resulting in an F1-score of 0.56 and an AUC of 0.77. This indicates that the baseline model is more conservative in predicting positive cases, producing fewer false positives but at the cost of missing a substantial portion of true pathological cases. Such behavior is more aligned with confirmatory or diagnostic settings, where higher precision is prioritized.

In contrast, the proposed SSL-ViT model demonstrates a substantially higher recall (0.73), indicating improved sensitivity in detecting pathological findings. This improvement comes with a decrease in precision (0.28), reflecting an increased number of false-positive predictions. Consequently, the F1-score is lower (0.40), while the AUC remains comparable (0.75), suggesting that the overall discriminative ability of the model is preserved despite the shift in prediction behavior. Importantly, the performance profile of SSL-ViT is particularly suitable for screening-oriented clinical applications, such as chest X-ray analysis, where minimizing false negatives is critical. The higher recall achieved by SSL-ViT ensures that fewer diseased cases are overlooked, which is a key requirement in early detection and triage systems. The comparable AUC values further indicate that self-supervised pretraining enhances sensitivity without significantly compromising the model's overall ranking capability. Overall, these results demonstrate that incorporating self-supervised learning into the Vision Transformer framework effectively shifts the model toward a high-sensitivity regime, making SSL-ViT a robust alternative for large-scale chest X-ray screening, while the baseline ViT may be better suited for scenarios requiring higher precision.

IV. CONCLUSION

This study presents an SSL-ViT framework for multi-label lung disease classification on chest X-ray images using the CheXpert dataset. By incorporating a self-supervised learning stage prior to supervised fine-tuning, the proposed approach effectively addresses the limitations of labeled data scarcity in medical imaging. The experimental results demonstrate that SSL pretraining significantly enhances the model's sensitivity, as reflected by consistently high recall values across most disease categories and a peak AUC of 0.7510. Although the precision of the SSL-ViT model is lower than that of the baseline ViT, the recall-oriented behavior aligns well with the primary objective of clinical screening, where false-negative predictions pose a greater risk than false positives.

The class-wise evaluation reveals that the model performs particularly well on diseases with distinct radiographic patterns, such as Pleural Effusion and Cardiomegaly, while still maintaining meaningful discriminative capability for conditions with subtle or overlapping features. Grad-CAM analysis further confirms that the model attends to clinically relevant lung regions, supporting the interpretability and reliability of the proposed approach. Comparative analysis with a baseline ViT model highlights a clear trade-off between precision and recall, demonstrating that SSL-ViT is more suitable for early detection and triage scenarios, whereas the baseline model may be preferable in confirmatory diagnostic settings. Overall, this research confirms that integrating self-supervised learning with Vision Transformers provides a powerful and practical solution for multi-label chest X-ray classification. Future work will explore alternative uncertainty-handling strategies, advanced SSL paradigms, and external dataset validation to further improve model robustness, calibration, and clinical applicability.

REFERENCES

[1] J. Zhou, Y. Xu, J. Liu, L. Feng, J. Yu, and D. Chen, “Global burden of lung cancer in 2022 and projections to 2050: Incidence and mortality estimates from GLOBOCAN,” *Cancer Epidemiol*, vol. 93, p. 102693, Dec. 2024, doi: 10.1016/j.canep.2024.102693.

[2] Z. Wang *et al.*, “Global, regional, and national burden of chronic obstructive pulmonary disease and its attributable risk factors from 1990 to 2021: an analysis for the Global Burden of Disease Study 2021,” *Respir Res*, vol. 26, no. 1, p. 2, Jan. 2025, doi: 10.1186/s12931-024-03051-2.

[3] K. E. S. Wijaya, G. A. Pradipta, and D. Hermawan, “Optimisasi Parameter VGGNet melalui Bayesian Optimization untuk Klasifikasi Nodul Paru,” *Seminar Hasil Penelitian Informatika dan Komputer (SPINTER)| Institut Teknologi dan Bisnis STIKOM Bali*, pp. 882–887, 2024.

[4] Z. Ge, D. Mahapatra, S. Sedai, R. Garnavi, and R. Chakravorty, “Chest X-rays Classification: A Multi-Label and Fine-Grained Problem,” Jul. 2018, doi: 10.48550/arXiv.1807.07247.

[5] N. I. Khani and S. Rakasiwi, “Penerapan Convolutional Neural Network dengan ResNet-50 untuk Klasifikasi Penyakit Kulit Wajah Efektif,” *Edumatic: Jurnal Pendidikan Informatika*, vol. 9, no. 1, pp. 217–225, Apr. 2025, doi: 10.29408/edumatic.v9i1.29572.

[6] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *ICLR 2021 Conference Paper*, Oct. 2021.

[7] T. Chen *et al.*, “A vision transformer machine learning model for COVID-19 diagnosis using chest X-ray images,” *Healthcare Analytics*, vol. 5, p. 100332, Jun. 2024, doi: 10.1016/j.health.2024.100332.

[8] J. Ko, S. Park, and H. G. Woo, “Optimization of vision transformer-based detection of lung diseases from chest X-ray images,” *BMC Med Inform Decis Mak*, vol. 24, no. 1, p. 191, Jul. 2024, doi: 10.1186/s12911-024-02591-3.

[9] U. Marikkar, S. Atito, M. Awais, and A. Mahdi, “LT-ViT: A Vision Transformer for Multi-Label Chest X-Ray Classification,” in *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, Oct. 2023, pp. 2565–2569. doi: 10.1109/ICIP49359.2023.10222175.

[10] J. Irvin *et al.*, “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison,” Jan. 2019, doi: 10.48550/arXiv.1901.07031.

[11] X. Liu *et al.*, “Self-supervised Learning: Generative or Contrastive,” *IEEE Trans Knowl Data Eng*, pp. 1–1, 2021, doi: 10.1109/TKDE.2021.3090866.

[12] L. Oakden-Rayner, G. Carneiro, T. Bessen, J. C. Nascimento, A. P. Bradley, and L. J. Palmer, “Precision Radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework,” *Sci Rep*, vol. 7, no. 1, p. 1648, May 2017, doi: 10.1038/s41598-017-01931-w.

[13] P. Rajpurkar *et al.*, “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” *stanfordmlgroup*, Dec. 2017, [Online]. Available: http://arxiv.org/abs/1711.05225

[14] F. C. Ghesu *et al.*, “Marginal Space Deep Learning: Efficient Architecture for Volumetric Image Parsing,” *IEEE Trans Med Imaging*, vol. 35, no. 5, pp. 1217–1228, May 2016, doi: 10.1109/TMI.2016.2538802.

[15] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study,” *PLoS Med*, vol. 15, no. 11, p. e1002683, Nov. 2018, doi: 10.1371/journal.pmed.1002683.

[16] V. R. Joseph, “Optimal ratio for data splitting,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/sam.11583.

[17] Apache MXNet, “Random Horizontal Flip,” Papers With Code.

[18] S. Zini, A. Gomez-Villa, M. Buzzelli, B. Twardowski, A. D. Bagdanov, and J. van de Weijer, “Planckian Jitter: countering the color-crippling effects of color jitter on self-supervised training,” *ArXiv*, Feb. 2023.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning*, in ICML’20. JMLR.org, 2020.

[20] R. Yulvina *et al.*, “Hybrid Vision Transformer and Convolutional Neural Network for Multi-Class and Multi-Label Classification of Tuberculosis Anomalies on Chest X-Ray,” *Computers*, vol. 13, no. 12, p. 343, Dec. 2024, doi: 10.3390/computers13120343.

[21] J. Li, “Area under the ROC Curve has the most consistent evaluation for binary classification,” *PLoS One*, vol. 19, no. 12, p. e0316019, Dec. 2024, doi: 10.1371/journal.pone.0316019.

[22] M. B. A. McDermott, H. Zhang, L. H. Hansen, G. Angelotti, and J. Gallifant, “A Closer Look at AUROC and AUPRC under Class Imbalance,” Jan. 2025, doi: 10.48550/arXiv.2401.06091.

[23] B. Wu *et al.*, “Visual Transformers: Token-based Image Representation and Processing for Computer Vision,” Nov. 2020, doi: https://doi.org/10.48550/arXiv.2006.03677.