SN Computer Science (2023) 4:754
https://doi.org/10.1007/s42979-023-02204-2

ORIGINAL RESEARCH

ViT: Quantifying Chest X-Ray Images Using Vision Transformer & XAI Technique

Yalamanchili Salini[1] · **J. HariKiran**[1]

Received: 16 November 2022 / Accepted: 31 July 2023

Abstract
As the availability of chest X-ray images increases and the demand for automated diagnosis grows, researchers have turned to deep neural networks to achieve accurate classification. While CNN backbones have traditionally been relied upon in computer vision methods, recent studies have demonstrated the superior performance of transformers, commonly used in NLP, in the domain of vision models. However, ensuring transparency in algorithms is essential for clinicians to comprehend the technical processes underlying the provided information and to maintain trust in adopting such systems. Therefore, the integration of Explainable Artificial Intelligence (XAI) techniques becomes imperative, not only for clinician acceptance but also for preserving the doctor–patient relationship. This study focuses on evaluating our proposed model using the "Chest X-ray14" dataset, which includes over 100,000 frontal and back-view images from 30,000 patients with 14 different chest illnesses. Through comprehensive testing across all pathology classes, our model achieves the highest average AUC score and class-wise AUC for each specific disease. To facilitate future research, we present an experimental setup that enables fair benchmarking of existing methods. Additionally, our findings validate the effectiveness of the proposed approach in accurately identifying chest areas associated with various pathologies. This research contributes to the advancement of automated diagnosis in the field of chest X-ray analysis and establishes a foundation for further investigations in this domain.

Keywords Vision transformer · Classification · Object detection · XAI technique

Introduction

In bioinformatics, medical experts have been burdened with repetitive tasks such as medical imaging diagnosis for many years. Additionally, since the global pandemic has emerged, there is an unprecedented increase in interest in reliable, accurate, and fast diagnostic methods for Chest X-rays (CXR). Several studies have shown that Computer Vision algorithms based on Neural Networks are more accurate than human-level algorithms, both in terms of accuracy as well as speed [1–4]. A large amount of labeled/annotated data are required for the training of these networks for CXR diagnosis, which has recently been provided by the National Institutes of Health (NIH). The Chest X-ray data set, "Chest X-ray 14", consists of more than 100 k images of front and back views. The presence of this data set spurred us to explore this problem further.

In several benchmarks, [5] contend that transformers are more efficient than CNNs. In the medical industry, [6] stress that it is not just necessary to have good performance, but also to be transparent and explainable behind intelligent machine decisions. Though medical fields are reluctant to adopt transformers because of their poor performance and difficulty explaining their operation, transformers are an appealing choice in theory for the medical field.

The existing knowledge on transformer models in medical image analysis is lacking due to a lack of research. The gap mentioned above is one of the main motivations for this paper. By classifying medical images using the Chest X-ray 14 data set [7], the present study partially fills this gap. Transformers, in contrast to CNNs, scale well [8] and can learn patterns that are not local [9]. Further, transformers are innately characterized by their attention weights [10]. These weights can be transformed into relevant heat maps for explanation [8, 10] through post hoc methods

✉ Yalamanchili Salini
yalamanchilisalini@gmail.com

J. HariKiran
harikiran.j@vitap.ac.in

[1] School of Computer Science & Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India

Published online: 29 September 2023

[10]. A study of explainability in transformer models is beyond the scope of this paper.

The vision transformer (Vit) [8] was recently developed as transformer-type deep learning architectures [11]; they not only perform well but also address some of the problems associated with CNNs [5, 12].

To achieve the above objective, a transformer architecture based on the Vit approach is compared to a MedMNIST baseline set and ResNet-18/ResNet-50 [13] in multi-class and binary classification. As a result of these insights, a discussion can be had on how transformer-based architectures perform under domain-specific conditions, thereby advancing medical image analysis knowledge. For detecting anomalies in low-level CXR feature corpora, we propose a new ViT architecture that leverages both the data set and network architecture.

Recently, technological advances have led to an increase in the use of AI in the medical field [14]. As a result, black-box models leave clinicians ill-equipped to explain the reasoning behind diagnoses to patients and their families. The traditional doctor–patient relationship must be maintained by introducing techniques that fill the trustworthiness and transparency gap [15].

The field of Explainable Artificial Intelligence (XAI) plays a critical role in upholding transparency in our society. Consider a scenario where a clinician needs to determine whether to perform surgery on a patient with a brain tumor. Given the inherent risks involved in the procedure, it is crucial to avoid unnecessary surgeries if no tumor is present. Conversely, the presence of a tumor significantly increases the patient's chances of survival, making surgical removal necessary. However, if pathologists rely on algorithms they cannot comprehend, it becomes challenging for them to make well-informed decisions.

An algorithm misclassification could have grave consequences, leading to the patient undergoing surgery when no tumor exists, potentially resulting in loss of life on the operating table. Moreover, justifying such a decision to the family of a deceased patient based on a flawed algorithm would be complex for the clinician. However, by utilizing XAI in this scenario, there is a higher likelihood of identifying misclassifications caused by an algorithm's misplaced focus on MRI slices. The incorporation of XAI in critical situations like these has the potential to save lives. Briefly, our contributions can be summarized as follows:

- Based on the representations of the common CXR findings found in the low-level CXR feature corpus, a novel ViT model for Chest X-rays is proposed.
- Our model has been expanded to quantify severity, so clinicians can make treatment decisions based on clinical guidelines instead of just classification.
- It consists of a single multi-task model that integrates both classification and severity quantification models for straightforward application, and both of the tasks are performing better as a result.
- Thus, this research is devoted to exploring and validating effective XAI techniques that foster trustworthiness and transparency in the medical diagnosis domain.
- The proposed model has been evaluated on the data set above (i.e., Chest X-ray14) and yielded the best results. Based on the Chest X-Ray14 data set, we achieve the best average AUC score and class-wise AUC for different classes.

The sections are organized as follows: the section "Background" presents the Background, which encompasses the historical and current applications relevant to this work. The section "Materials" provides details on the Materials used. The section "Methods " elucidates the Methods employed, highlighting the current approaches utilized in this study. Within the section "Methods", the focus will be on discussing the proposed network and its significance. The section "Proposed Work" involves evaluating the performance and comparing it with previous works and existing techniques. The section "Experiments and Analysis" provides an overview of the discussion. Finally, the concluding section, the section "Discussion", summarizes the key findings and offers a conclusion.

Background

In this section, we will discuss some deep learning concepts to provide transparency. The next part of our discussion is transformers and how they can be used for tasks that involve vision, such as ours.

Deep Neural Networks

Deep neural networks have been widely employed for classification tasks in chest X-ray analysis. These networks leverage their ability to automatically learn hierarchical representations from raw image data, enabling them to capture intricate patterns and feature relevant to specific diseases or abnormalities.

Convolutional Neural Networks (CNNs) are a popular choice for chest X-ray classification. CNNs consist of multiple layers, including convolutional layers that extract local features, pooling layers that down sample the features, and fully connected layers that perform the final classification. These networks can effectively learn discriminative features from the image data and achieve high accuracy in disease classification.

Transfer learning is often employed in chest X-ray analysis using deep neural networks. Pre-trained CNN models, such as VGG Net, ResNet, or Inception Net, trained on large-scale data sets like ImageNet, are fine-tuned or used as feature extractors for chest X-ray classification [16]. By leveraging the knowledge gained from pre-training on diverse visual data, transfer learning allows models to generalize well even with limited labeled chest X-ray images.

Ensemble models have also been explored in chest X-ray classification. These models combine the predictions of multiple individual models to improve overall performance. Ensemble methods, such as bagging, boosting, or stacking, help reduce model variance and enhance the robustness of the classification system.

Moreover, advancements in deep learning have led to the exploration of more complex architectures specifically designed for medical image analysis. For example, attention mechanisms, such as self-attention or channel attention, have been integrated into CNNs to improve feature representation and focus on relevant regions in the chest X-ray images.

Overall, deep neural networks, particularly CNNs, have shown great potential in chest X-ray classification tasks. They provide the capability to automatically learn meaningful representations from images, enabling accurate disease identification and assisting healthcare professionals in diagnosing and treating patients.

CNN-Based Models

CNN-based models have been extensively utilized for classification tasks in chest X-ray analysis due to their ability to effectively capture local and global image features. These models have demonstrated significant advancements in automated disease detection and classification. Here are some notable CNN-based models employed in chest X-ray classification:

- *Alex Net*: Introduced in 2012, Alex Net was one of the pioneering CNN architectures that revolutionized image classification. Its deep architecture, consisting of multiple convolutional and fully connected layers, achieved breakthrough performance on the ImageNet dataset and subsequently paved the way for CNN-based medical image analysis, including chest X-rays.
- *VGG Net*: VGG Net, proposed in 2014, is characterized by its uniform architecture with stacked convolutional layers. It offers different variations (e.g., VGG16, VGG19) and has been widely adopted in chest X-ray classification tasks. VGG Net's deep structure enables it to learn complex image representations and exhibit strong performance.
- *ResNet*: ResNet, introduced in 2015, addresses the challenge of training very deep neural networks. It introduces residual connections that enable the network to efficiently learn and propagate gradients, allowing for deeper architectures without suffering from vanishing or exploding gradients. ResNet variants, such as ResNet-50 and ResNet-101, have been successfully employed in chest X-ray classification.
- *Densenet*: Densenet, proposed in 2016, introduces dense connections where each layer receives feature maps from all preceding layers. This architecture promotes feature reuse and facilitates gradient flow, leading to improved model accuracy. Densenet has demonstrated promising results in chest X-ray classification tasks.
- *Inception Net*: Inception Net, also known as Google Net, was introduced in 2014. It employs the concept of "inception modules" with parallel convolutional operations of different filter sizes, allowing the model to capture both local and global image features. Inception Net has been utilized in chest X-ray analysis, achieving competitive performance.

These CNN-based models, along with their variations and modifications, have been extensively explored in chest X-ray classification. They have shown remarkable capabilities in capturing relevant image features, distinguishing between normal and abnormal conditions, and assisting in the diagnosis of various chest diseases and abnormalities.

While CNN-based models and deep neural networks have shown remarkable success in chest X-ray classification, they do have certain drawbacks:

- CNNs and deep neural networks are often referred to as black-box models due to their lack of interpretability.
- Deep neural networks typically require large amounts of labeled data for training.
- Deep neural networks are prone to overfitting, especially when dealing with limited data.

To address these drawbacks and enhance the interpretability of deep neural networks in chest X-ray analysis, Explainable Artificial Intelligence (XAI) techniques and transformers have entered the scene.

XAI techniques, such as Local Interpretable Model-Agnostic Explanations (LIME), Gradient-weighted Class Activation Mapping (Grad-CAM), and attention maps, aim to provide explanations and visualizations for the decisions made by CNN-based models. These techniques help clinicians understand the model's reasoning by highlighting the regions or features contributing to the classification.

Transformers, originally introduced for natural language processing (NLP), have recently gained attention in computer vision tasks. Transformers leverage the self-attention mechanism to capture global relationships and dependencies within an image. Vision Transformers (ViTs) apply

this mechanism to process image patches and have shown promising results in various vision tasks, including chest X-ray analysis.

The introduction of ViTs in quantifying chest X-rays combines the advantages of transformers, such as capturing long-range dependencies and enabling interpretability through attention maps, with the capability to classify chest diseases accurately. This approach allows clinicians to understand and trust the decision-making process of the model while achieving high performance in chest X-ray classification.

By integrating XAI techniques and transformers, researchers aim to enhance the transparency, interpretability, and performance of deep learning models in chest X-ray analysis, leading to more reliable and trustworthy automated diagnosis systems.

Transformers

The Transformer architecture is a type of neural network structure employed for sequence-to sequence assignments like machine translation. The diagram displayed in Fig. 1 illustrates the Transformer architecture, highlighting the functioning of residual connections, which will be elucidated in the following explanation.

- *Input*: The input to the transformer architecture is typically a sequence of tokens or embeddings representing the input data.
- *Patch embeddings*: The input sequence is divided into fixed-size patches, which are then linearly transformed into lower dimensional embeddings. These patch embeddings serve as the initial representations of the input sequence.
- *Multi-head attention*: The patch embeddings are passed through multiple parallel self-attention heads in the multi-head attention layer. Each attention head attends to different parts of the input sequence, capturing different relationships and dependencies. The outputs of the attention heads are concatenated and linearly transformed.
- *Residual add & normalization*: The output of the multi-head attention layer is added element-wise with the initial patch embeddings, creating a residual connection. This helps to preserve the original information and alleviate the vanishing gradient problem. The result is then passed through a layer normalization step, which normalizes the values along each feature dimension.
- *Feed forward network*: The normalized outputs from the previous step are fed into a feed-forward neural network (FFN). The FFN consists of two linear transformations with a non-linear activation function (such as ReLU) in between them. This allows the model to capture more complex and higher-order relationships within the input sequence.
- *Residual add & normalization*: Similar to the previous step, the output of the feed-forward network is added element-wise with the input from the previous step, creating another residual connection. This is followed by another layer normalization step.
- *Output*: The final normalized outputs from the previous step represent the processed information from the transformer architecture. Depending on the task at hand (e.g., machine translation, language modeling), further steps like decoding or classification may be performed on these outputs to generate the final desired output.

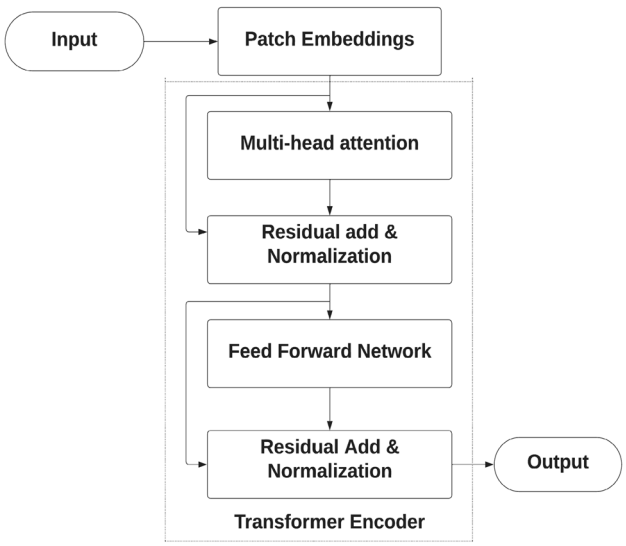


Fig. 1 Transformer architecture

Overall, the transformer architecture leverages multi-head attention to capture global dependencies within the input sequence, while residual connections and layer normalization help preserve and enhance the flow of information throughout the model. The feed-forward network introduces non-linear transformations to capture more complex patterns, leading to more expressive representations.

Vision Transformer

As Transformers became popular in the natural language processing (NLP) community, several attempts have been made to adapt them to computer vision (CV) tasks. Detection Transformer (DETR), Vision Transformer (ViT), Data-efficient image Transformers (DeiT), and Swin Transformer are currently the most representative Transformer-based models in vision.

Vit resembles a conventional Transformer in its architecture. To provide spatial information, the input image is

converted into a series of patches with positional encoding methods that encode each patch's spatial position. The patches are then fed into the Transformer with a class token to calculate the MHSA and yield the learned embeddings. Using the class token as an image representation, the Vit outputs the state of the token. The learned image representation is then classified with the help of a multi-layer perception (MLP) algorithm [17]. A Vit also can take feature maps from CNNs and map them to raw images. In Fig. 2, we see a detailed view of the Vision Transformer. Table 1 shows some of the advantages of ViT over CNN.

Pattern Recognition

One of the main applications of image processing is pattern recognition. In many applications, pattern recognition is used, such as license plate recognition, fingerprint analysis, face recognition, and voice-based authentication. This technique is used in medical imaging, as shown in Fig. 3.

Materials

Data Set

Chest X-ray14 data set is used in this study. This data set released 100,000 frontal/back-view images with 8-bit gray-scale values, released by 30,850 patients with labels from 14 pathology classes. It is possible to have multiple labels on an image. A radiology report is analyzed to collect labels expected to have a 90% accuracy rate. A more accurate

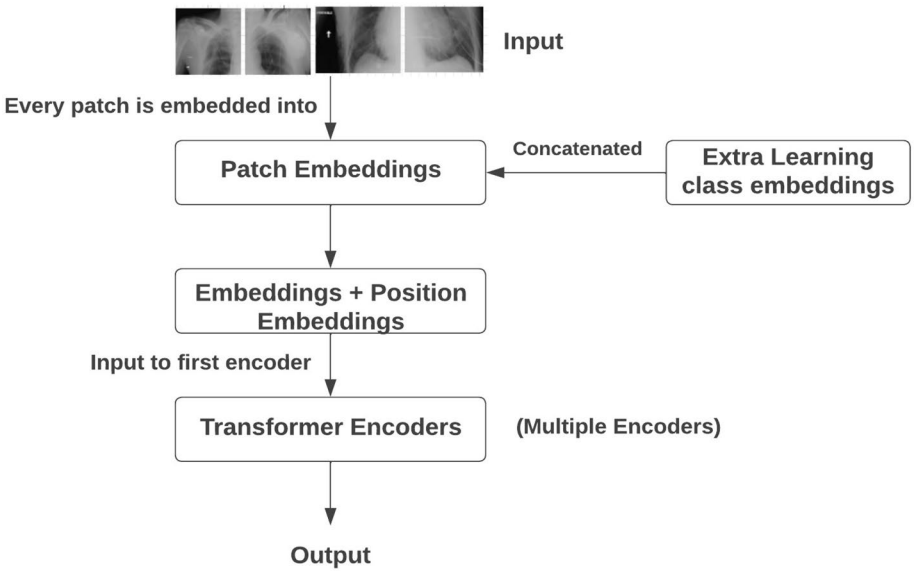


Fig. 2 Vision transformer

Table 1 Differences between ViT and CNN

	ViT	CNN
The similarity between lower and higher levels	Uniform similarity	Grid-like pattern
Attention distances	Lower layers attend globally	Lower layers attend only locally
Skip-connection propagation	Phase transition layer propagation in higher levels	Uniform less propagation in higher levels
Effect of GAP	w/o GAP, clear localization	W GAP, blurry localization
Effect of scale	Larger data sets improve intermediate layer representation	

Fig. 3 Steps involved in pattern recognition

and objective comparison can be made by applying [18] a patient-wise split.

- A text-mined label identifies 14 different pathological conditions in each image in the data set.
- Eight different diseases can be diagnosed using these tests.
- Each 14 labeled pathology is classified using these data in a multi-class classification model due to its accuracy and reliability.
- As a result, it will predict 'positive' or 'negative' outcomes for each disease.

For 5 out of 14 pathologies, this data set has been annotated by consensus among four different radiologists below:

- Consolidation.
- Edema.
- Effusion.
- Cardiomegaly.
- Atelectasis.

Addressing Class Imbalance

In the realm of chest X-ray classification, the issue of class imbalance frequently arises due to the varying prevalence of diseases. As certain diseases are more prevalent than others, it creates an imbalance in the distribution of classes within the data set. Addressing class imbalance can be a complex task, since it can result in models that exhibit a bias toward the majority class. Consequently, these models tend to predict the majority class more frequently, even when the minority class should be the correct prediction. Figure 4 illustrates the frequency distribution of each label in the data set.

The plotted data reveal substantial variations in the prevalence of positive cases across different pathologies, and these trends align with the patterns observed in the complete data set. Notably:

- The Hernia pathology exhibits the most significant imbalance, with the proportion of positive training cases being approximately 0.2%.
- Even in the case of the Infiltration pathology, which has the least imbalance, only 17.5% of the training cases are labeled as positive.
- Ideally, for training our model, we would prefer a balanced data set where positive and negative training cases contribute equally to the loss.
- In the presence of a highly unbalanced data set, as demonstrated here, utilizing a standard cross-entropy loss function would result in the algorithm prioritizing the majority class (negative cases in this scenario) due to its greater impact on the loss function.

Dealing with Class Imbalance in Machine Learning

A loss function measures your prediction model's accuracy in predicting the expected outcome. Increasing the predicted probability over the actual label increases cross-entropy loss [24]. Cross entropy loss is a commonly used loss function in classification tasks that measures the dissimilarity between predicted probabilities and the true class labels. As a result of cross-entropy loss, Eq. 1 says: in this model, here, x_i and y_i are the input features and the label, and $f(x_i)$ represents the probability that the output is optimistic based on the input features

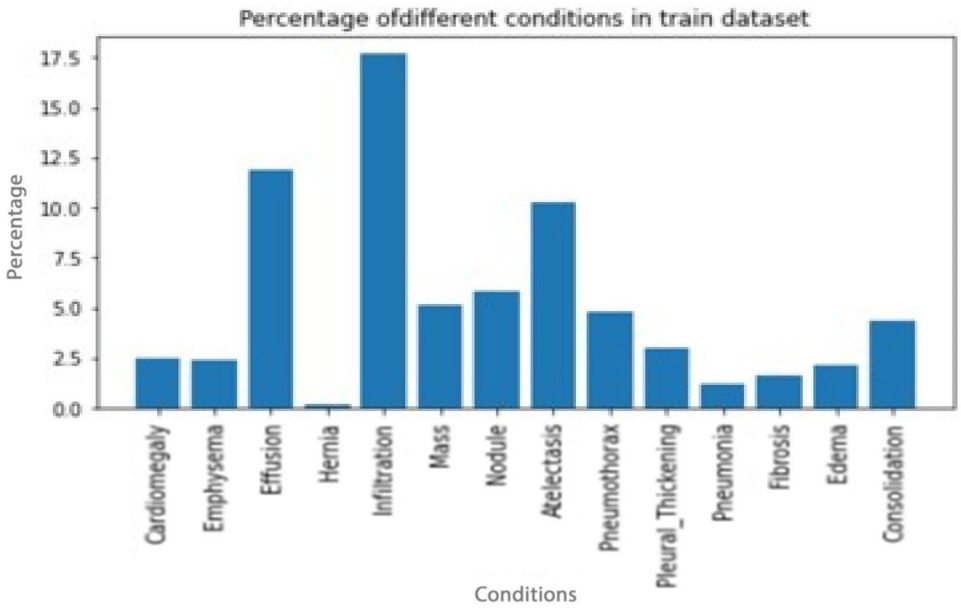


Fig. 4 Pre-processed data of diseases

and label. In the diagram Fig. 5, a loss function measures the quality of the output of the network

$$L_c\text{ross} - \text{entropy}(x_i) = -(y_i \log(f(x_i)) + (i - x_i)\log(1 - f(x_i))) \tag{1}$$

$$L_c\text{ross} - \text{entropy}(D) = \left(\frac{1}{N}\right)\left(\sum_{\text{positive examples}} \log\left(f(x_i)\right)\right) + \left(\sum_{\text{positive examples}} \log\left(1 - f(x_i)\right)\right). \tag{2}$$

Due to the formulation described in Eq. 2, an issue may arise if there is a substantial class imbalance within the data set, particularly when there are only a few positive training cases. In such cases, the negative class tends to dominate the loss calculation. This dominance of the negative class can lead the model to prioritize predicting the negative class even when the positive class is the correct classification, as demonstrated in Eqs. 3 and 4.

To address this concern, we employ the formulation outlined in Eqs. 3 and 4, where the contributions of each class (i.e., pathological condition) are summed over all training cases. Through this training process, we aim to achieve balanced classes, as illustrated in Fig. 7

$$\text{Count}_p = \frac{\text{No of positive classes}}{N} \tag{3}$$

$$\text{Count}_n = \frac{\text{No of negative classes}}{N}. \tag{4}$$

Computing and Visualizing Class Imbalance

Analyzing class imbalance entails examining how instances are distributed across various classes within a dataset to identify potential imbalances. To assess this, we consider the entire dataset and determine the frequency of different conditions detected among the X-rays. Subsequently, we visualize the relative percentages depicting the presence of various conditions, as depicted in Fig. 6. As depicted in the preceding plot, the contributions of positive cases are notably lower compared to the contributions of negative cases. Nonetheless, our objective is to equalize these contributions. One approach to achieving this is by assigning class-specific weight factors, denoted as w_p and w_n, to each example from their respective classes. This ensures that the overall contribution from each class remains consistent. To fulfill this objective, we aim to satisfy Eq. 5

$$w_p \times \text{freq}_p = w_n \times \text{freq}_n, \tag{5}$$

which we can do simply by taking

$$w_p = \text{freq}_n w_n = \text{freq}_p.$$

This way, we will be balancing the contribution of positive and negative labels. Now, the balanced classes obtained as shown in Fig. 7 are used for visualization.

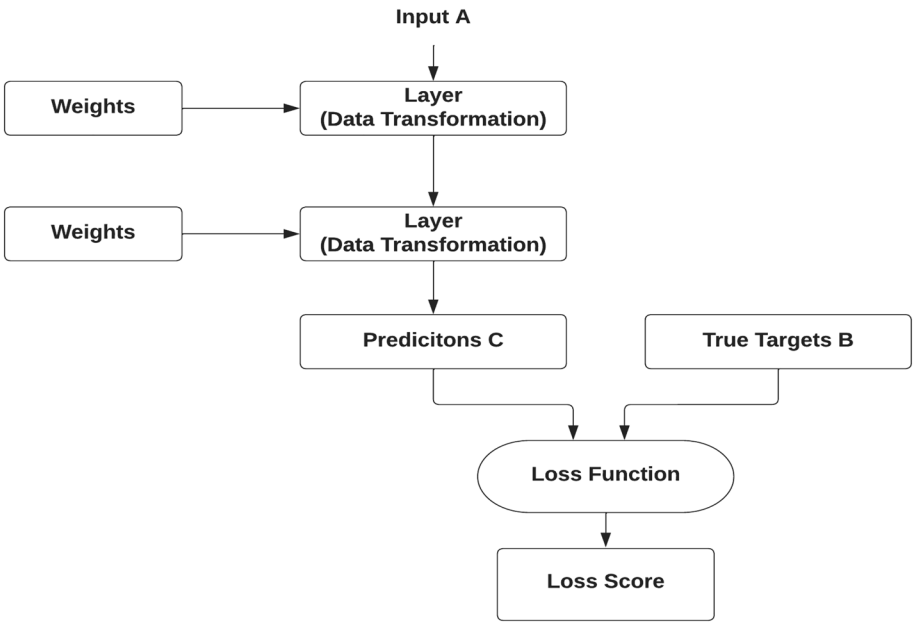


Fig. 5 Cross entropy loss

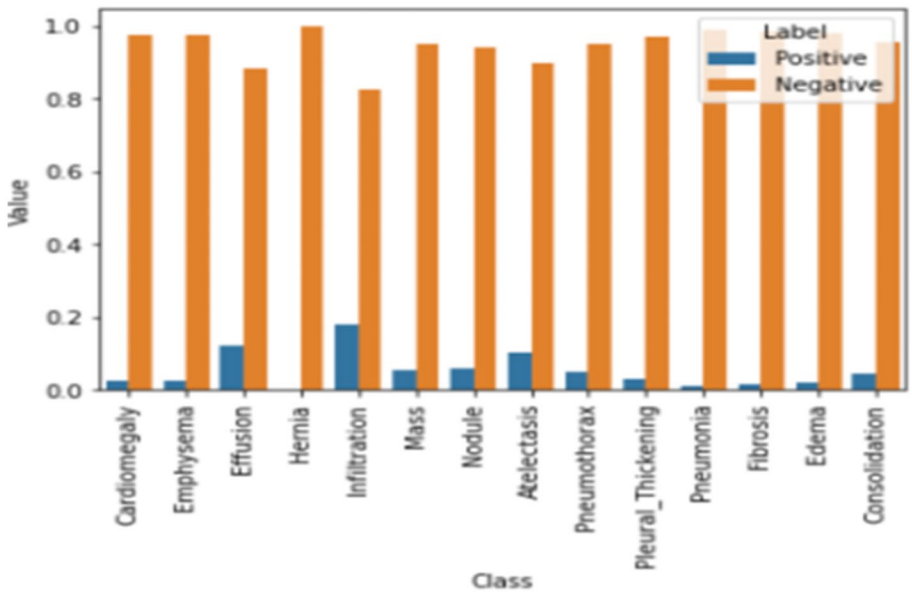


Fig. 6 Class imbalance was accounted on different conditions

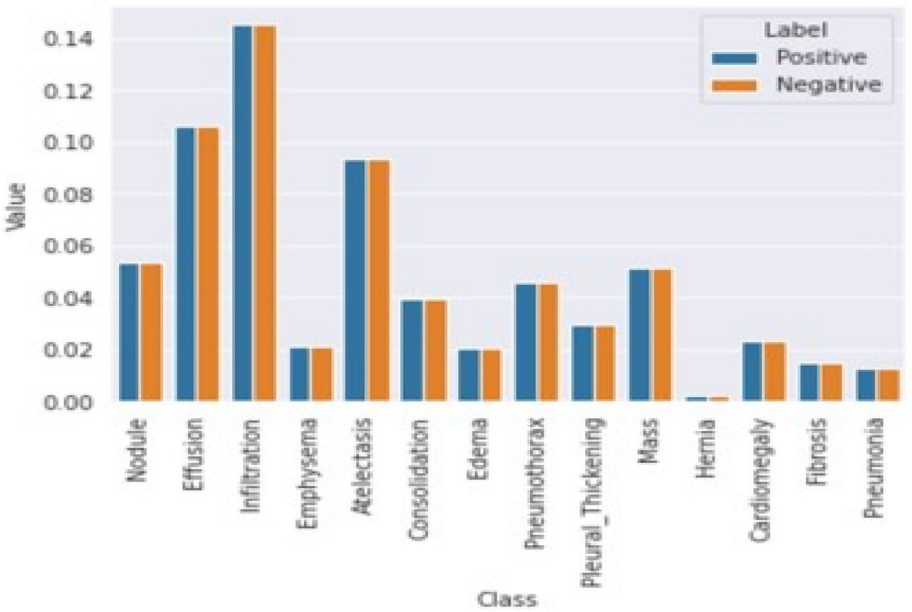


Fig. 7 Balanced classes was accounted on different conditions

Calculating the Weighted Loss

By referring to the provided Fig. 7 depicting balanced class data, it can be observed that applying these weightings ensures that both positive and negative labels within each class make an equal overall contribution to the loss function. Next, we will move forward with the implementation of this loss function. Once the weights have been computed, the resulting weighted loss for each training case can be represented using Eq. 6

$$L_{\text{cross-entropy}}(w, x) = -\left(w_p y \log(f(x)) + w_n(1 - y)\log(1 - f(x))\right). \quad (6)$$

Methods

Image Preparation

With the Keras framework, we can utilize the Image Data Generator class to construct a "generator" that generates images using a data frame as a basis.

- A primary data augmentation method, such as flipping images horizontally, is also supported by this class.
- The generator also applies transformations to the values of each batch, ensuring that their mean becomes 0 and

the standard deviation becomes 1. This standardization of input distributions through the generator aids in facilitating model training.

- As well as converting gray-scale images to three-channel formats, the generator repeats each channel's value across the image. Our trained model needs three-channel inputs so we need this.

Object Detection

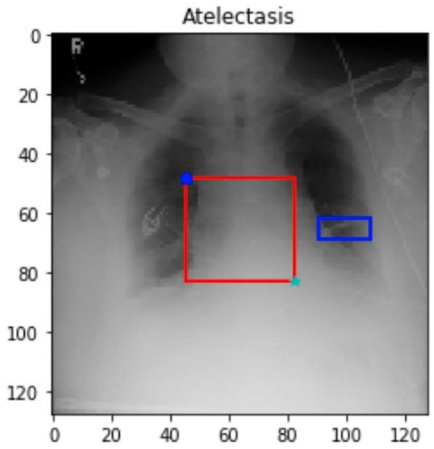
Object detection is the task of finding and localizing objects in an image. In the context of chest X-ray images, this could involve finding and locating nodules, tumors, or other abnormalities. It ensures that each disease is appropriately enclosed within a bounding box, even if multiple diseases are present in a single image (see Fig. 8). It includes both the detection of these diseases and the provision of their respective class labels and accurate spatial coordinates.

Bounding Box: Region of Interest (ROI)

To improve the accuracy of object recognition, we need to come up with a new way to represent the location of objects in images. Bounding boxes are often used to represent the location of objects that have been detected by an object detection algorithm.

Figure 9 shows a more accurate way to represent the location of objects in images. This method uses four coordinates to specify the location of an object: the top-left corner, the bottom-right corner, the width, and the height. This method is more accurate, because it takes into account the size and shape of the object.

The new method for representing the location of objects in images can be used to improve the accuracy of object recognition. This can be done using the coordinates of the bounding boxes to extract features from the image that are specific to the object. These features can then be used to train a machine-learning model to recognize the object.




```
x0    90.779661
y0    61.868928
h0     6.915254
w0    17.627119
Name: 3, dtype: float64
[45.373013 48.026855 34.715107 37.187298]
```

Fig. 9 Obtaining bounding box with coordinates to locate the pathological class

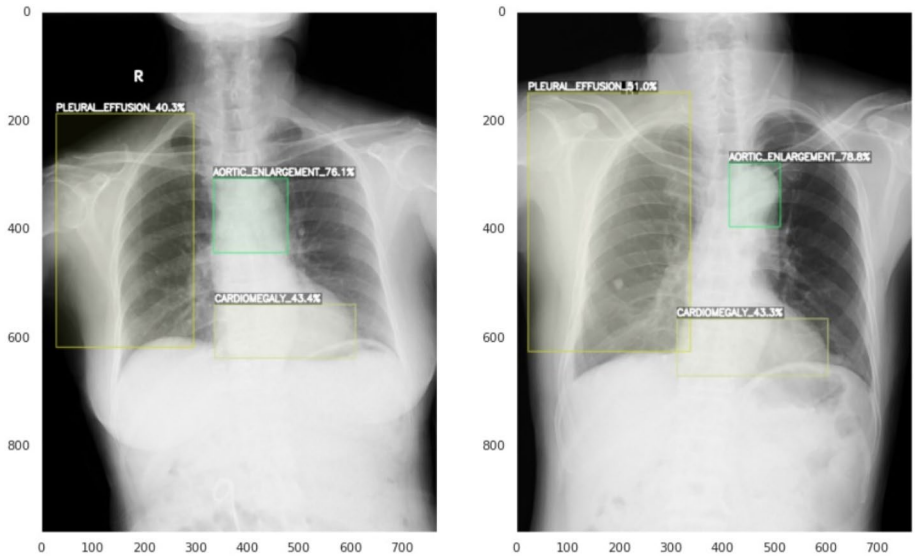


Fig. 8 Disease classification using object detection

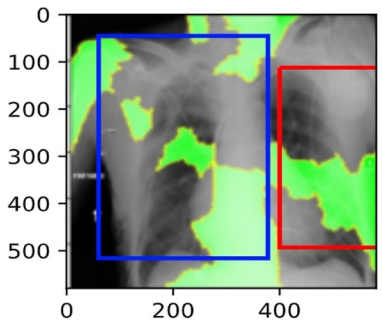


Fig. 10 Regions of interest are marked with a bounding box that helps a clinical expert identify and locate disease in an image

The Role of Bounding Boxes in Object Detection

There are a few reasons why bounding boxes are used instead of simply determining the location of an object with object detection. First, bounding boxes provide a way to represent the location of objects in an image in a standardized way. This makes it easier to compare the results of different object detection algorithms. Second, bounding boxes can be used to extract features from images that are specific to objects. This can be helpful for tasks, such as classification and diagnosis.

For example, an object detection algorithm might be used to detect nodules in chest X-ray images. The algorithm would first identify the potential locations of nodules in the image. Then, it would use bounding boxes to define the location of each nodule. Finally, the algorithm would extract features from the image that are specific to each nodule, such as the size, shape, and texture of the nodule.

The use of bounding boxes in chest X-ray images can be helpful for a number of reasons. First, bounding boxes provide a way to represent the location of objects in an image in a standardized way. This makes it easier to compare the results of different object detection algorithms. Second, bounding boxes can be used to extract features from images that are specific to objects. This can be helpful for tasks, such as classification and diagnosis, as shown in Fig. 10.

Here are some of the benefits of using bounding boxes in chest X-ray images:

- *Standardized representation*: Bounding boxes provide a standardized way to represent the location of objects in an image. This makes it easier to compare the results of different object detection algorithms.
- *Feature extraction*: Bounding boxes can be used to extract features from images that are specific to objects. This can be helpful for tasks such as classification and diagnosis.
- *Localization*: Bounding boxes can be used to localize objects in an image. This can be helpful for tasks such as surgical planning.

XAI Technique: LIME

The Local Interpretable Model-Agnostic Explanation (LIME) is a widely recognized technique in the field of Explainable Artificial Intelligence (XAI). Its "local" nature signifies that LIME is employed to elucidate individual observations or records rather than providing explanations for the entire data set. By focusing on a single observation, LIME generates human-understandable explanations that shed light on the model's decision-making process.

LIME achieves this by investigating how perturbing a specific input, such as manipulating its features, influences the model's predictions. Both Fig. 11 and Eq. 7 illustrate the utilization of a local interpretable model to explain the predictions made by a classifier or regressor in a faithful manner.

The primary objective of LIME is to bridge the gap between AI systems and human understanding. It adopts a people-centric approach and concentrates on two key aspects:

- Trusting the model.
- Trusting the prediction

$$\text{Bounding box} = \frac{\text{Total LIME pixels inside Bounding Box}}{\text{Total LIME pixels}}. \tag{7}$$

Object detection and bounding boxes are techniques for identifying and localizing objects in images. They can be used to explain the predictions of models that have been trained on chest X-rays, but they have some limitations. For example, object detection and bounding boxes can be difficult to interpret, especially for complex objects. They can also be biased by the features that are used to identify the objects.

LIME addresses these limitations by creating a simplified version of the model that is interpretable. This simplified model can then be used to explain the predictions of the original model in a way that is easy to understand. In addition, LIME can be used to explain the predictions of models that have been trained on different types of data, including chest X-rays. This makes it a more versatile tool than object detection and bounding boxes. Here is a Table 2 that summarizes the benefits of using LIME over object detection and bounding boxes:

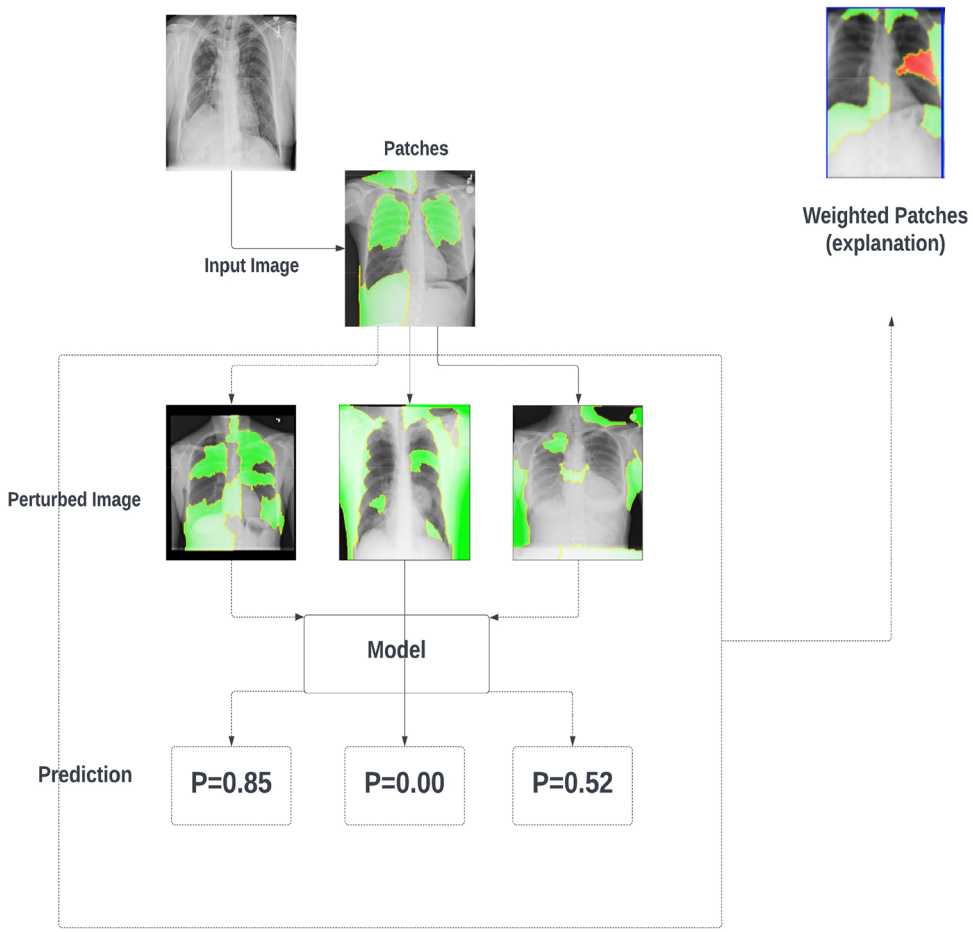


Fig. 11 LIME technique for feature mapping in X-ray images

Table 2 Comparative analysis on LIME over object detection and bounding boxes

Benefit	LIME	Object detection and bounding boxes
Explainability	Easy to understand	Can be difficult to interpret
Accuracy	Accurate in explaining the predictions of machine-learning models	Can be biased by the features that are used to identify the objects
Flexibility	Can be used to explain the predictions of models that have been trained on different types of data	Can only be used to explain the predictions of models that have been trained to identify objects

Proposed Work

The paper demonstrates that the Transformer model can be optimized by utilizing the low-level CXR corpus and backbone network to produce common CXR findings. Despite having fewer labeled cases, subsequent models built on this backbone network are less likely to overfit, because it is trained with many more data points. As a result, the network can generalize more effectively.

To improve the applicability and performance of individual tasks, we integrated the chest X-ray classification and severity quantification models that can accomplish two tasks simultaneously. As a result, the model offered better applicability and improved performance.

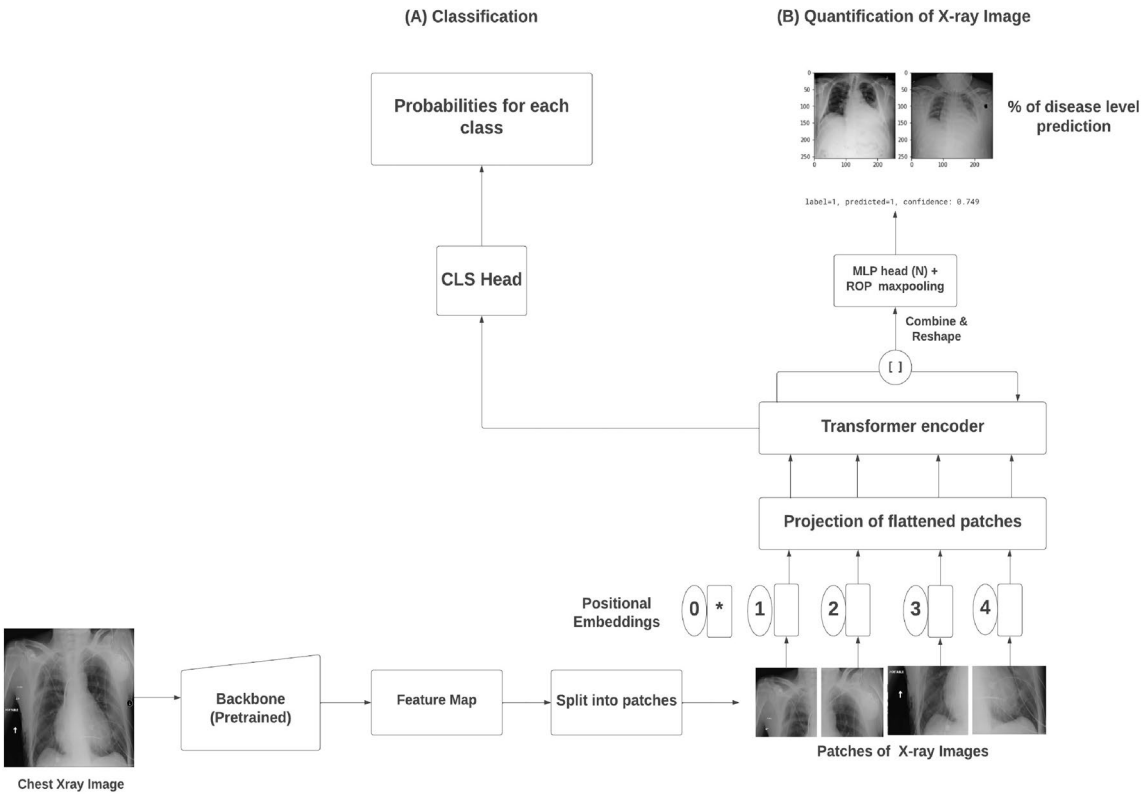


Fig. 12 A multi-task vision transformer model is proposed for diagnosing and quantifying chest X-ray severity, with **A** a shared backbone and Transformer and **B** task-specific heads

Vision Transformer for Chest X-Rays: Classification

To make the classification model more interpretable, we adapted [19] a saliency map that computes relevance for the Transformer network. The methods in Chefer et al., Smilkov et al., Srinivas & Fleuret differs from gradient propagation methods [20–22] or attribution propagation methods, which use heuristic propagation in conjunction with attention graphs or attention maps that are obtained. In (2020), the local significance is calculated through deep Taylor decomposition and propagated throughout all layers. For Transformer architecture models, this relevance propagation method is beneficial, since it bypasses continuous operations and overcomes the self-attention problem.

A linear classifier can be added to the [class] token to obtain the diagnosis result y of the input CXR image x (see Fig. 12A).

- *Input representation*: Let X be the input Chest X-ray image from Eq. 8

$$X \in \mathrm{RH} \times W \times C \tag{8}$$

where H represents the image height, W represents the image width, and C represents the number of channels (e.g., 1 for grayscale or 3 for RGB images).

- *Patch extraction*: The input image X is divided into a set of non-overlapping patches, represented as Eq. 9

$$P \in \mathrm{RN} \times P \times P \times C \tag{9}$$

where N is the total number of patches, and P is the patch size (e.g., 16×16). Each patch is flattened into a vector, resulting in a reshaped patch matrix with the below Eq. 10.

$$P\mathrm{reshaped} \in \mathrm{RN} \times (P \times P \times C) \tag{10}$$

- *Embedding*: The reshaped patch matrix Preshaped is linearly projected to a lower-dimensional representation by multiplying it with a learnable weight matrix from Eq. 11

$$E \in R(P \times P \times C) \times d \tag{11}$$

where d is the embedding dimension. The resulting embedded patches matrix represents the embedded features of the patches by Eq. 12.

$$P\text{embedded} \in \text{RN} \times d \tag{12}$$

- *Positional encoding*: Positional encoding is added to the embedded patches to provide spatial information. Let Eq. 13 be the positional encoding matrix.

$$\text{PE} \in R^{N \times d} \tag{13}$$

- *Transformer encoder*: The embedded patches P embedded with positional encoding PE are inputted into a stack of Transformer Encoder layers. Each Transformer Encoder layer consists of a multi-head self-attention mechanism followed by feed-forward neural networks. The output of the Transformer Encoder stack is denoted as:

$$O \in R^{N \times d} \tag{14}$$

where Eq. 14 represents the encoded features of the patches.

- *Classification head*: The encoded features O are fed into a classification head, which typically consists of one or more fully connected layers. The final output of the classification head is the predicted probability distribution over the classes. Let Eq. 15 represent the predicted probabilities, where C′ is the number of classes.

$$Y\text{pred} \in RC' \tag{15}$$

The overall forward pass of the vision transformer can be summarized as Eq. 16

$$Y_{\text{pred}} = \text{Classification Head (Transformer Encoder (Positional Encoding (Embedding (Patch Extraction (X))))).} \tag{16}$$

Vision Transformer for Chest X-Rays: Severity Quantification

Figure 12 shows how this works. Identifying diseases severity map is produced by combining reshaped output features except for [class] tokens using an additional lightweight network. As shown in Fig. 12B, we first subtract the [class] token position from the Transformer output. This vector N is used as the input in the map head network.

Additionally, Transformer is trained to perform downstream diagnoses using its self-attention mechanism. As a result, our method mimics that of clinical experts in determining the final diagnosis of X-ray images by examining all 14 classes with their probabilities, as shown in Fig. 13. Moreover, our ViT framework facilitates sequential follow-up of severity and helps clinicians make treatment decisions by checking severity quantification and localization which can be evaluated from Eq. 17.

Severity calculation: The encoded feature O can also be used to calculate the severity score for the detected abnormality. Let this

$$\text{Spred} \in R^{C''} \tag{17}$$

represents the predicted severity scores, where C″ is the number of severity levels. The severity calculation can be performed using fully connected layers or other appropriate mechanisms [23].

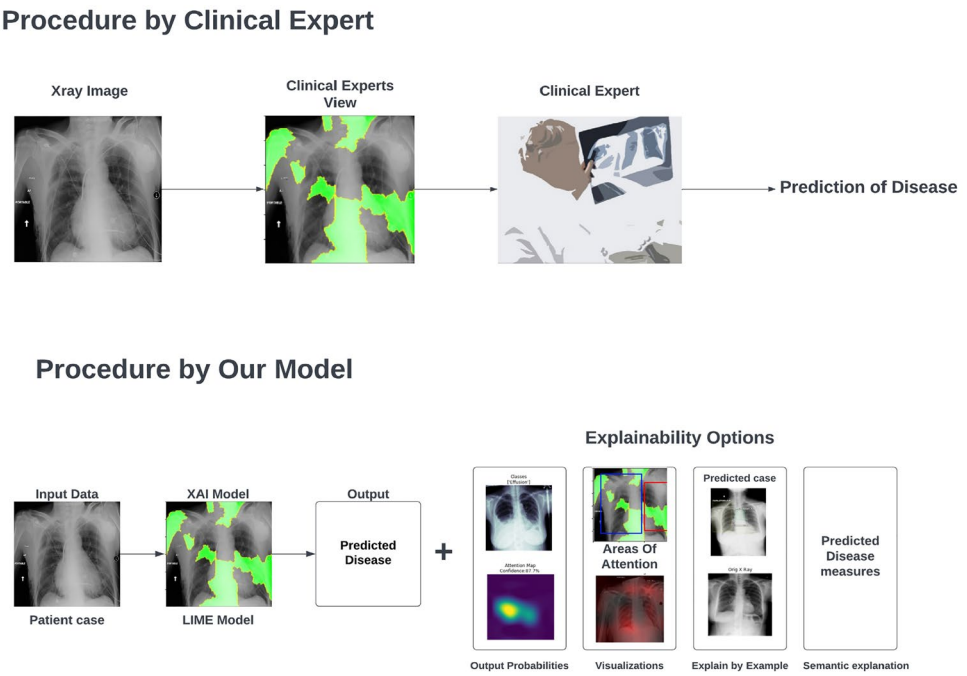


Fig. 13 Analogy between clinical expertise and our method of diagnosis

Experiments and Analysis

We run several experiments on the Chest X-ray 14 data set to evaluate the performance of various transformers. Furthermore, we compare our results against state-of-the-art research in the field. Optimizing can be sped up, and local minimums can be found more quickly if a good starting point is chosen. These experiments are described and discussed in the following sections.

Localization

To enable the model to learn distinct and intricate functions for each pathology, an MLP head structure is connected to the output of the VIT transformer. This approach is adopted to compensate for the lack of standardized parameters and to facilitate the learning of more intricate functions tailored to individual diseases. The investigation then proceeds to explore various head structures to assess their influence on the accuracy of predictions. By examining different head structures, the researchers aim to evaluate how they impact the overall predictive performance of the model.

Figure 14 displays LIME heat maps for samples encompassing cardiomegaly, effusion, nodule, and consolidation. These heat maps offer insights into the specific regions within the lungs that our model focuses on when making diagnoses related to lung diseases. By visualizing these mapped areas, domain experts can discern potentially significant regions within the input image that are associated with each specific disease. This analysis aids in the identification of crucial areas of attention for accurate disease diagnosis.

In Fig. 14a, an instance of cardiomegaly, characterized by an enlarged heart, is presented. The location of the heart, particularly its edge portions, plays a significant role in determining this condition. Figure 14b illustrates the lower portions of the two lungs and their relative positions, which are crucial for our model's diagnosis of cardiomegaly. By comparing the relative positions of the two lungs, as shown in Fig. 14c, our model is able to make accurate diagnoses. The LIME heat mapping employed by our model effectively detects this condition.

Consolidation, a lung condition that manifests as white or opaque areas in Chest X-ray images can be clearly observed. In Fig. 14d, our model successfully identifies these consolidated lung regions and provides a positive diagnosis.

Results

To determine the optimal performing model, we conducted a series of experiments. We trained several models using a combination of dense layers at the lower part of the network to predict our binarized classes. From these experiments, we identified the top-performing networks and leveraged their predicted classes to create an ensemble.

Ensembling involves combining multiple algorithms within the ensemble to address biases and errors inherent in individual models, ultimately leading to improved performance in the analysis of chest X-rays. Various ensemble methodologies were explored during the research process. The following operations can be performed:

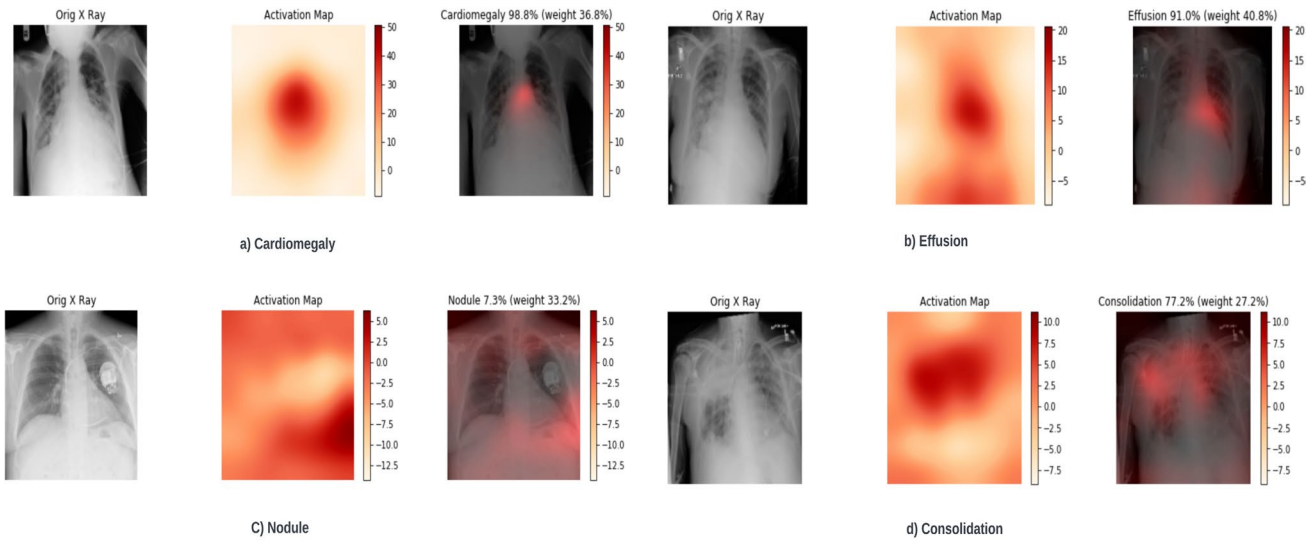


Fig. 14 Validating important regions for positive samples containing cardiomegaly, effusion, nodule, and consolidation. For each sample, the original image is shown left-side and the heat maps are shown on the right

Table 3 Comparison with different models on Chest X-Ray 14

	Class	Mobile net	ResNet	VGG	AVG ensemble	Wght AVG ensemble	Hard Vote ensemble	Max vote ensemble	ViT
0	Atelectasis	0.757	0.762	0.538	0.773	0.773	0.619	0.769	**0.781**
1	Cardiomegaly	0.904	0.885	0.560	0.911	**0.911**	0.670	0.908	0.891
2	Consolidation	0.702	0.687	0.588	0.704	0.704	0.500	0.701	**0.748**
3	Edema	0.846	0.842	0.700	0.853	**0.853**	0.507	0.850	0.848
4	Effusion	0.830	0.827	0.586	0.838	0.838	0.739	**0.839**	0.824
5	Emphysema	0.860	0.864	0.495	0.877	**0.877**	0.655	0/873	0.914
6	Fibrosis	0.769	0.755	0.360	0.772	0.772	0.500	0.767	**0.826**
7	Infiltration	0.680	0.682	0.551	0.689	0.689	0.641	0.686	**0.701**
8	Mass	0.782	0.769	0.452	0.799	0.799	0.578	0.796	**0.822**
9	Nodule	0.695	0.702	0.421	0.708	0.708	0.509	0.706	**0.78**
10	Pleural-thickening	0.736	0.7171	0.437	0.737	0.737	0.500	0.730	**0.778**
11	Pneumonia	0.677	0.678	0.560	0.693	0.693	0.500	0.684	**0.713**
12	Pneumothorax	0.841	0.839	0.491	0.852	0.852	0.700	0.847	**0.871**

Bold is to emphasize specific data points that exhibit superior accuracy compared to others

The last column shows our proposed approach, i.e., vision transformer (ViT)

- *Simple averaging of predictions*: The predictions from each model within the ensemble are averaged together to obtain the final prediction.
- *Weighted average of predictions*: Similar to simple averaging, but each model's prediction is assigned a weight based on its performance or confidence level.
- *Hard vote*: In this approach, the predicted values are coerced into binary format, and a voting mechanism is applied to determine the final prediction. The class with the majority of votes is selected.

These ensemble techniques provide an opportunity to improve the overall performance and reliability of the models in chest X-ray analysis. Table 3 displays the performance metrics obtained from the suggested method along with various pre-trained models. Our proposed method exhibits higher accuracy in detecting segmented images on the Chest X-ray 14 data set compared to the other pre-trained models, achieving an accuracy of 80.7%. Additionally, Fig. 15 visually presents the ROC curves and plots depicting the performance metrics achieved through our proposed method.

The hard vote ensemble is an interesting approach, because it sacrifices overall accuracy for a lower false-negative rate. This means that the model is less likely to miss a positive case, even if it does result in more false positives. This could be a useful approach in situations where it is more important to avoid missing a positive case than to avoid false positives.

The ViT transformer output is connected to an MLP head structure, so the model can learn different and more complex functions based on pathology. The model can therefore learn more complex functions specific to each pathology, because the head structures do not share parameters. Through this approach, all pathologies can learn low-level image features shared between them, while their head sections can learn high-level features unique to them.

As you can see in Table 4, ViT achieves the highest AUC score for all 14 diseases. This means that it is able to correctly classify the presence or absence of these diseases with a higher accuracy than other methods.

The improvement in performance is likely due to the fact that ViT uses a transformer-based architecture, which is better at capturing long-range dependencies in images. This is important for chest X-rays, as many diseases can only be detected by looking at multiple parts of the image. Other methods, such as DenseNet-121 and DenseNet-10, also achieve good performance on the Chest X-ray 14 dataset. However, ViT is able to achieve a slightly higher accuracy.

Comparative Analysis with Different Pre-trained Models with Best Results Highlighted

The Weight AVG Ensemble model was able to achieve an AUC (Area Under the Curve) of 0.785, which is considered a representative performance benchmark for pre-trained models in multi-label classification on Chest X-ray images. However, our ViT model outperforms this benchmark with

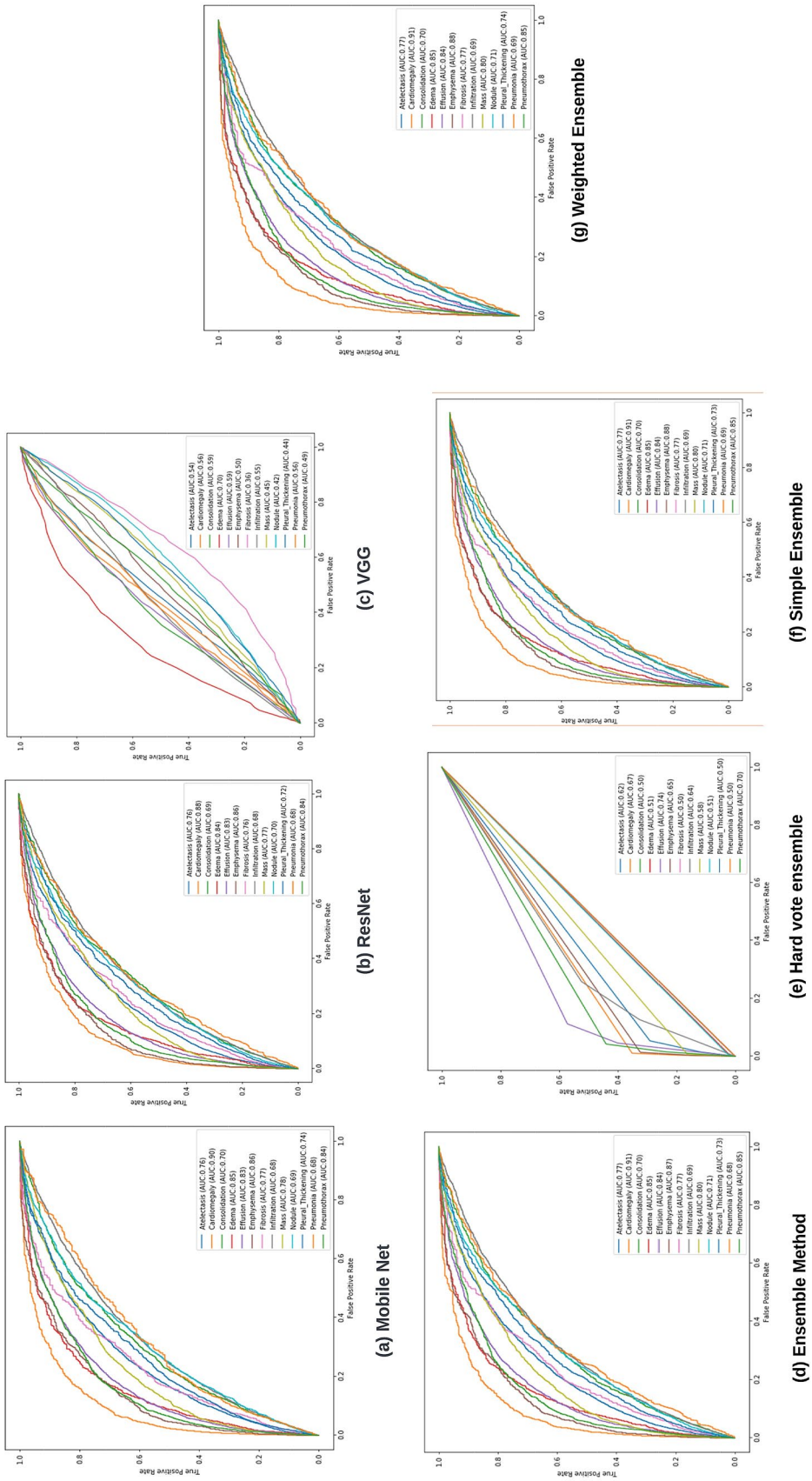


Fig. 15 AUC curves

Table 4 Comparison of 14 pathological diseases with State-of-the-Art Chest X-ray 14 dataset

	Class	DenseNet-121 [24]	DenseNet-10 [25]	ResNet-38 [25]	ViT
0	Atelectasis	0.747	0.738	0.762	**0.781**
1	Cardiomegaly	0.783	0.775	0.792	**0.891**
2	Consolidation	0.792	0.784	**0.804**	0.748
3	Edema	0.742	0.733	0.757	**0.848**
4	Effusion	0.751	0.742	0.763	**0.824**
5	Emphysema	0.728	0.720	0.749	**0.914**
6	Fibrosis	0.720	0.711	0.742	**0.826**
7	Infiltration	0.784	0.775	**0.793**	0.701
8	Mass	0.788	0.780	0.798	**0.822**
9	Nodule	0.794	0.786	**0.806**	0.78
10	Pleural- Thickening	0.768	0.759	**0.780**	0.778
11	Pneumonia	0.804	0.796	**0.812**	0.713
12	Pneumothorax	0.789	0.781	0.799	**0.871**

Bold is to emphasize specific data points that exhibit superior accuracy compared to others

Table 5 Analyzing the AUCs for multi-label classification with different methods and the proposed method

Methods	Mean for all of the pathologies
Mobile net	0.776
Resnet	0.769
VGG	0.518
AVG ensemble	0.773
Wght AVG Ensemble	0.785
Hard vote ensemble	0.586
Max vote ensemble	0.781
ViT	**0.807**

Bold is to emphasize specific data points that exhibit superior accuracy compared to others

an AUC score of 0.807. Furthermore, our model exhibits dominance over Wght AVG Ensemble in terms of diseases such as Cardiomegaly, Edema, and Emphysema. Additionally, our model holds an advantage over Wght AVG Ensemble by eliminating the need for adaptive learning rates. The utilization of adaptive learning rates introduces additional hyperparameters, determining when and how much to decrease the learning rate. These hyperparameters require significant time and effort to tune, potentially leading to overfitting the validation set. In contrast, our model offers a more straightforward approach without these complexities. The results presented in Table 4 demonstrate that our proposed evaluation method yields an average AUC that is 2.2% higher than that of the Weight AVG Ensemble model for all pathology's, further affirming the superior performance of our ViT transformer.

Discussion

In this research, the Chest X-ray14 data set, comprising 14 different chest pathologies, was subjected to classification using various models based on vision transformers, as outlined in this paper by Table 5.

Additionally, a comparative analysis was conducted between the results obtained from the ViT method and those achieved by the previous models. To ensure a fair and unbiased comparison, the previous state-of-the-art models were evaluated using the authors' own evaluation method, enabling the selection and assessment of the model based on this approach.

As you can see in Table 6, ViT achieves the highest average AUC score on the Chest X-ray 14 dataset. This means that it is able to correctly classify the presence or absence of 14 different diseases in chest X-rays with the highest accuracy.

Other state-of-the-art approaches, such as ResNet-38 and Guendel et al. [26], also achieve good performance

Table 6 Comparison of the proposed model with other state-of-the-art approaches using the Chest X-ray 14 dataset

Methods	Mean for all of the pathologies
ResNet-38-large-meta	0.805
Guendel et al. [26]	0.806
DenseNet-121 [27]	0.799
DenseNet-10 [25]	0.789
ViT (our method)	**0.807**

Bold is to emphasize specific data points that exhibit superior accuracy compared to others

on the Chest X-ray 14 dataset. However, ViT is able to achieve a slightly higher accuracy.

To showcase the effectiveness of vision transformers, a multi-label classification task utilizing chest X-ray images was employed. The primary focus of the testing and experiments was placed on the ViT transformer. However, it is worth noting that future investigations can explore and incorporate alternative vision transformers, broadening the scope of the study.

Furthermore, the Chest X-ray14 data set presents challenges, such as images with scattered lung locations and variations in contrast. To address these issues effectively, the researchers developed a pre-processing technique. This technique aimed to efficiently eliminate the aforementioned challenges from the images before training the models. By implementing this pre-processing technique, the researchers sought to enhance the quality and consistency of the data, thereby improving the overall training process.

Conclusion

The objective of this research was to develop an advanced deep learning model capable of efficiently diagnosing infection levels in real-world healthcare settings, encompassing 14 distinct variants. To achieve this, an extensive collection of over 100,000 frontal and back-view images was assembled by combining various open-source data sets.

Among the numerous model architectures available, the Vision Transformer (ViT) model architecture was specifically chosen for this study. The ViT model demonstrated remarkable performance, achieving an impressive Area Under the Curve (AUC) score of 80.7%. The AUC score is a widely utilized metric for evaluating classification models, and this high score indicates the model's proficiency in accurately diagnosing infection levels.

To enhance the model's interpretability and instil trust in its predictions, an XAI-based (Explainable Artificial Intelligence) LIME model visualization pipeline was implemented. Local Interpretable Model-agnostic Explanations (LIME) enabled the visualization of the model's decision-making process by highlighting the specific regions within the X-ray images that contributed significantly to the final diagnosis. This interpretable inference pipeline provided valuable insights into how the model arrived at its predictions, increasing confidence in its reliability.

The proposed approach holds the potential to revolutionize the use of chest X-ray images as a simple and cost-effective diagnostic tool for all 14 pathological diseases. Particularly in situations where rapid testing methods are unavailable, the utilization of chest X-ray images can be of immense value. By leveraging this approach, medical professionals can efficiently assess infection levels, aiding in prompt and accurate diagnoses.

As the availability of more extensive data sets continues to grow, future research endeavors will focus on proposing an alternative variant of the Vision Transformer model. This variant aims to further enhance the model's performance in diagnosing infection levels across the 14 variants. The utilization of larger data sets is anticipated to provide valuable insights and improvements, contributing to more accurate and reliable diagnostic capabilities.

In summary, this research successfully developed a sophisticated deep learning model specifically tailored for efficient infection-level diagnosis in real-world healthcare scenarios. The study incorporated the Vision Transformer model architecture, leveraged an XAI-based LIME model visualization pipeline, and emphasized the potential of chest X-ray images as a cost-effective diagnostic tool.

Funding No funding.

Code Availability A software application or custom code will be made available based on a reasonable request.

Data Availability Data sets will be made available based on reasonable request.

Declarations

Conflict of Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Goodfellow I, Warde-Farley D, Mirza M, Courville A, Bengio Y. Maxout networks. In: Dasgupta, S., McAllester, D. (eds) Proceedings of the 30th International Conference on Machine Learning, vol. 28 of Proceedings of Machine Learning Research, p.1319–1327 (PMLR, Atlanta, Georgia, USA, 2013).
2. Greff K, Srivastava RK, Schmidhuber J. Brainstorm: fast, flexible and fun neural networks, Version 0.5 (2015).
3. Sabour S, Frosst N, Hinton GE, et al. Dynamic routing between capsules. In: Guyon I, et al., editors. Advances in Neural

Information Processing Systems, vol. 30. Curran Associates Inc: Red Hook; 2017.
4. Mazzia V, Salvetti F, Chiaberge M. Efficient-CapsNet: capsule network with self-attention routing. Sci Rep. 2021;11:14634. https://doi.org/10.1038/s41598-021-93977-0.
5. Khan, Salman, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Khan, and Mubarak Shah. 2021. Transformers in vision: A survey.
6. Tjoa E, Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Trans Neural Networks Learn Syst. 2020;32(11):4793–813.
7. Yang J, Shi R, Ni B. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE; 2021. p. 191–5.
8. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint. 2020. arXiv:2010.11929.
9. Valanarasu JMJ, Oza P, Hacihaliloglu I, Patel VM. Medical transformer: Gated axial-attention for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer International Publishing; 2021. p. 36–46.
10. Abnar S, Zuidema W. Quantifying attention flow in transformers. arXiv preprint. 2020. arXiv:2005.00928.
11. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. PMLR; 2021. p. 10347–57.
12. Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021. p. 782–91.
13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–8.
14. Zhang Y, Weng Y, Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. Diagnostics. 2022. https://doi.org/10.3390/diagnostics12020237.
15. Lotsch J, Kringel D, Ultsch A. Explainable artificial intelligence (XAI) in biomedicine: making AI decisions trustworthy for physicians and patients. BioMedInformatics. 2021;22(1):1–17.
16. Taslimi S et al. SwinCheX: multi-label classification on chest X-ray images with transformers. arXiv preprint. 2022 arXiv:2206.04246.
17. Fürst J. Validating XAI techniques in medical image diagnosis: A venture towards algorithm transparency in a socio-technical system. Bachelor's thesis, University of Twente. 2022.
18. Wang X, et al. Chestx-ray: hospital-scale chest x-ray database and benchmarks on weaklysupervised classification and localization of common thorax diseases. IEEE Conf on Comput Vis Pattern Recognit (CVPR). 2017. https://doi.org/10.1109/cvpr.2017.369.
19. Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization. arXiv preprint. 2020 arXiv:2012.09838.
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. pp. 618–626.
21. Smilkov D, Thorat N, Kim B, Vi'egas F, Wattenberg M. SmoothGrad: removing noise by adding noise. arXiv preprint. 2017 arXiv:1706.03825.
22. Srinivas S, Fleuret F. Full-gradient representation for neural network visualization. arXiv preprint. 2019 arXiv:1905.00780.
23. Park S, et al. Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification. Med Image Anal. 2022;75:102299.
24. Chhabra M, Kumar R. A smart healthcare system based on classifier DenseNet 121 model to detect multiple diseases. In Mobile Radio Communications and 5G Networks: Proceedings of Second MRCN 2021. Singapore: Springer Nature Singapore; 2022. p. 297–31.
25. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest X-ray classification. Sci Rep. 2019;9(1):6381.
26. Guendel S, Grbic S, Georgescu B, Liu S, Maier A, Comaniciu D. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 23rd Iberoamerican Congress, CIARP 2018, Madrid, Spain, November 19–22, 2018, Proceedings 23. Springer International Publishing; 2019. pp. 757–765.
27. Antin B, Kravitz J, Martayan E. (2017). Detecting pneumonia in chest X-Rays with supervised learning. 2017. https://www.semanticscholar.org/paper/Detecting-Pneumonia-in-Chest-X-Rays-with-Supervised-Antin-Kravitz/bbc749a5c9139dc642a78647c1dfed1df71bba07.