scientific reports

Check for updates

OPEN

Explainable hybrid transformer for multi-classification of lung disease using chest X-rays

Xiaoyang Fu[1,4], Rongbin Lin[1,4], Wei Du[2], Adriano Tavares[3] & Yanchun Liang[1,2]✉

Lung disease is an infection that causes chronic inflammation of the human lung cells, which is one of the major causes of death around the world. Thoracic X-ray medical image is a well-known cheap screening approach used for lung disease detection. Deep learning networks, which are used to identify disease features in X-rays medical images, diagnosing a variety of lung diseases, are playing an increasingly important role in assisting clinical diagnosis. This paper proposes an explainable transformer with a hybrid network structure (LungMaxViT) combining CNN initial stage block with SE block to improve feature recognition for predicting Chest X-ray images for multiple lung disease classification. We contrast four classical pre-training models (ResNet50, MobileNetV2, ViT and MaxViT) through transfer learning based on two public datasets. The LungMaxVit, based on maxvit pre-trained with ImageNet 1K datasets, is a hybrid transformer with fine-tuning hyperparameters on the both X-ray datasets. The LungMaxVit outperforms all the four mentioned models, achieving a classification accuracy of 96.8%, AUC scores of 98.3%, and F1 scores of 96.7% on the COVID-19 dataset, while AUC scores of 93.2% and F1 scores of 70.7% on the Chest X-ray 14 dataset. The LungMaxVit distinguishes by its superior performance in terms of Accuracy, AUC and F1-score compared with other hybrids Networks. Several enhancement techniques, such as CLAHE, flipping and denoising, are employed to improve the classification performance of our study. The Grad-CAM visual technique is leveraged to represent the heat map of disease detection, explaining the consistency among clinical doctors and neural network models in the treatment of lung disease from Chest X-ray. The LungMaxVit shows the robust results and generalization in detecting multiple lung lesions and COVID-19 on Chest X-ray images.

It has been demonstrated that lung disease is one of the major causes of death around the world. According to the WHO, approximately 4 million early deaths occur yearly due to various lung-related illnesses, such as COVID-19[1], asthma and pneumonia[2] of all ages. Lung disease is classified as infectious such as bacteria, viruses, mycoplasmas, chlamydial pneumonia and so on, while noninfectious pneumonia is seen as the body's immune illness caused by chemical, physical or radiation. Chest X-ray scans of humans are an efficient diagnostic method employed by clinical doctors and medical image experts to detect lung lesions[3]. However, even for clinical medical experts, diagnosing various lung diseases is also a challenge because lung disease lesions appearing on X-ray images are ambiguous and without typical symptoms of disorders. On the other hand, Chest X-ray discrepancies have resulted in significant subjective judgements and differences in lung disease diagnosis, and capturing lung lesions from a complicated thoracic image background with human eyes is a time-consuming task. The problem is much more serious due to the shortage of skilled radiologists in developing countries, especially in rural regions. Therefore, a computer-aided diagnostic system (CAD) can be deployed to complete large-scale diagnosis of lung disease by using thoracic X-ray images.

Artificial Intelligence and deep learning algorithms have demonstrated remarkable performance in pathology detection from Chest X-ray (CXR) images in recent years[4,5]. Convolutional Neural Networks (CNNs) have significantly impacted on the field of medical imaging due to their ability to learn complex representations in deep neural networks, which has made an outstanding performance in image processing, voice recognition and pattern recognition[6]. CNNs have also demonstrated prominent improvements in numerous medical imaging modalities, including Radiography, Computed Tomography (CT)[7], Ultrasound Images[8] and Magnetic Resonance Imaging(MRI), to name a few. The operation of CNN is an end-to-end method to make predictions

[1]School of Computer Science, Zhuhai College of Science and Technology, Zhuhai 519040, China. [2]School of Computer Science and Technology, Jilin University, Changchun 130012, China. [3]Department of Industrial Electronics, University of Minho, 4800-058 Guimarães, Portugal. [4]Xiaoyang Fu and Rongbin Lin contributed equally. ✉email: ycliang@jlu.edu.cn

from the extracted valuable and relevant features of the input images. According to the prior study experience of CNN models, leveraging deep learning algorithms to predict lung diseases on Chest X-rays can relieve the load on radiologists. Related works show that ensemble methods of different CNN models with transfer learning have made much progress. Muhammad Mujahid studied an ensemble model structure made by incorporating CNN with Inception-V3, VGG-16, and ResNet50, showing that Inception-V3 with CNN attained the highest accuracy and recall score, respectively in pneumonia classification[9]. Similarly, Mudasir Ali et al. achieved 94.02% accuracy on Guangzhou Women and Children Medical Center X-ray dataset using EfficientNetV2L[10]. Sohaib Bin Khalid Alvi et al. achieved an accuracy of 89.6% on a COVID-19 dataset by applying a federated learning approach. Building on these efforts[11], Syeda Reeha Quasar et al. achieved an accuracy of 98% on a dataset for lung disease classification using CT images[12]. However, applying CNN to practice Chest X-rays classification also faces several substantial challenges. Despite the outstanding performance of CNN models, the problem of over-fitting and spatial information loss induced by normal convolution operation leads to poor result generalization. Recent studies have demonstrated that CNNs exhibit a strong bias toward style rather than content for classification.

Significant research effort has been made by the vision community to integrate the attention mechanisms[13]in CNN-inspired architectures. Alexey Dosovitskiy[14] has shown that these transformer modules can fully replace the standard convolutions in deep neural networks by operating on a sequence of image patches, giving rise to Vision Transformers (ViTs). The transformer is a self-attention-based architecture that emerged as the preferred paradigm in today's visual challenges. The adoption of transformer architecture enabled substantial parallelization and translation quality optimization. However, because the strong model capacity of transformers being imbued with less inductive bias, which leads to over-fitting, ViT does not perform well enough in image recognition without extensive pre-training. The Swin Transformer[15] is one such successful attempt to modify transformers by applying self-attention on shifted non-overlapping windows among these sparse transformer models tailored for vision tasks such as local attention, which leverages hierarchical architectures to compensate for the loss of non-locality. The MaxViT[16] allows global-local spatial interactions on arbitrary input resolutions with only linear complexity and multi-axis attention.

Due to the CNN drawbacks in image recognition and also inspired by the MaxViT, we present a novel deep learning architecture by effectively integrating the multi-axis attention module with convolutional layers based on SENet layers[17]. We propose an end-to-end hierarchical hybrid Chest X-ray image recognition backbone, called LungMaxvit, by blending improved convolutional blocks and MaxViT basic blocks on multiple stages. This paper selected deep learning models firstly by comparing them with the performance of four classical pre-trained models (ResNet50, MobileNetv2, ViT, MaxViT) via a transfer learning approach on the Chest X-ray images. In our experiments, we choose two classical CNN models and two Vit models to study which type model will perform better for predicting Lung diseases from Chest X-ray images. ResNet50 is a classical CNN for vision recognition and MobileNet is a dominant model used lung disease detection from Chest X-ray images. Vit and MaxViT with self-attention mechanism are used to compare with mobilenet in order to screen the best model for lung disease detection. The test results showed that MaxViT model performs the best among the models. The improved architecture based on MaxViT can distinguish the COVID-19 and 14 classes of lung disease automatically among the Chest X-ray images as an auxiliary measure for clinical treatment. The LungMaxVit was pre-trained by transfer learning and fine-tuned by the Chest X-ray datasets, so that the improved model shows the better test results in both Chest X-ray image datasets compared with the other classical models.

The contributions of the paper are summarized as follows:

- The proposed fine-tuned LungMaxViT model blends modified CNN layers with multi-axis transformer block inspired by MaxViT and achieved the highest disease classification accuracy among five pre-trained models.
- A comprehensive multi-class lung disease classification study is conducted using two different public Chest X-ray datasets: COVID-19 data and Chest X-ray14.
- Investigating the well-designed experimental evaluation of the proposed framework with other classical deep learning models, the hybrid framework demonstrated effectiveness for predicting the lung diseases while distinguishing the COVID-19 from the other lung diseases.
- Medically explainable visuals implemented by Grad-CAM that emphasize the crucial regions relevant to the model's prediction of the input image are proposed based on the LungMaxViT model.
- The images are preprocessed by Gaussian Filter[18–20] and CLAHE[21,22] technology to improve the consistency of the data and reduce noise.The remainder of this paper is arranged as follows: section "Methods" focuses on the related works in detail while contributing the study of experiment datasets and the deep learning models, providing the proposed training methods. Section "Results" includes experiments setup, environments and test results and analysis. Section "Discussion" discusses models and their limitation. Section "Conclusion" presents the paper conclusion and future works.

Methods

The proposed framework LungMaxViT in this paper is inspired by MaxViT to combine both deep convolutional neural networks with attention mechanisms for predicting the lung diseases. It is well known that the deep learning CNN models can be used to analyze the spatial correlation among the neighboring pixels in the receptive area determined by the convolutional filter size, ignoring the directional relationships with the distance among these pixels. To solve this, transformers based on deep learning attention mechanism have recently been presented and proven to be more powerful and robust in considering both spatial pixel correlation and their distance relations for visual recognition tasks.

The proposed hybrid deep learning architecture is shown in Fig. 1, where deep learning framework approaches multi-classification of lung diseases prediction from Chest X-ray images. The hybrid framework implements the following processing steps.

- Step 1: Preparing the study datasets of multiple Chest X-ray images from typical benchmark datasets.
- Step 2: Data processing including augmentation, resizing, rotation, normalization and data splitting into training validation and test sets.
- Step 3: Proper deep learning model selection is made by a comprehensive experimental study that compares the performance of the four classical deep learning pre-trained models: ResNet50, MobileNetv2, ViT and MaxViT.
- Step 4: The hybrid deep learning model is proposed by combining CNN initial stage block with SE block and multi-axis transformer inspired by MaxViT back-bone.
- Step 5: The proposed framework is pre-trained by imagenet1K for the initial training parameters and fine-tuned by the Chest X-ray datasets.
- Step 6: The final multiple classification prediction is made by the classification layer. The explainable visualization results were achieved by Grad-CAM using the proposed hybrid model, which determined the disease localization well.

Dataset

The proposed deep learning framework was studied by employing two different Chest X-ray datasets, Chest X-ray14[23] and COVID-QU-Ex (COVID-19)[24,25].

COVID-QU-Ex dataset

The COVID-QU-Ex dataset encompasses three distinct classes in Table 1: COVID-19 positive cases, Non-COVID-19 infections and Normal instances.

Chest X-ray14

The Chest X-ray14 dataset, currently one of the most extensive collections of Chest radiographs available, has evolved significantly since its inception. Initially comprising eight distinct pathology categories, the dataset has expanded to include a total of fifteen disease states, reflecting a broader spectrum of thoracic pathology. As seen in Table 2, these categories now encompass Pneumonia, Cardiomegaly, Edema, Effusion, Consolidation, Mass, Pleural Thickening, Nodule, Emphysema, Hernia, Fibrosis, Pneumothorax, Atelectasis, No Finding and Infiltration.

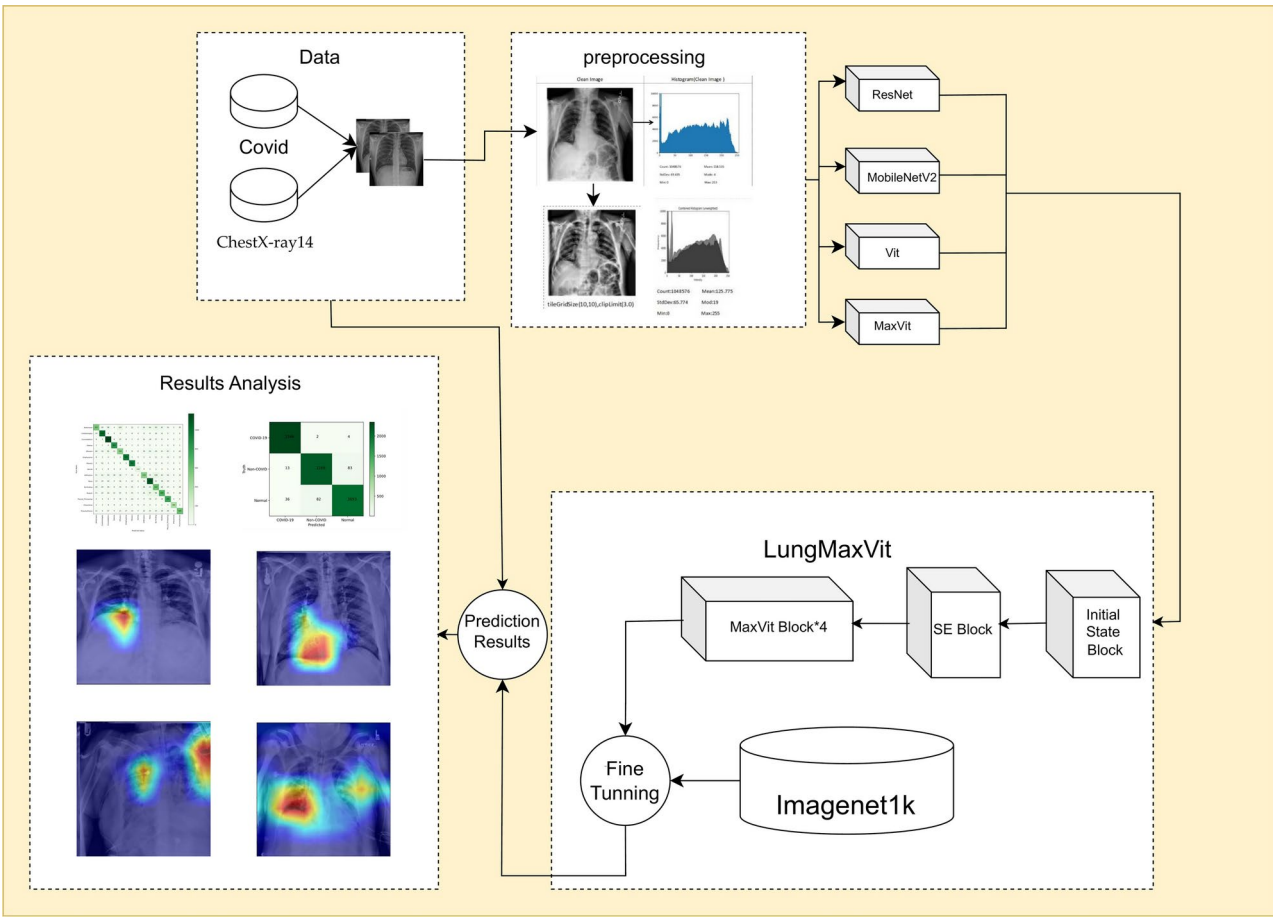

Fig. 1. The overview of deep learning framework approaches for multi classification lung diseases prediction from Chest X-ray images.

Class	Data
Covid 19	11,956
Non COVID infections	11,263
Normal	10,701

Table 1. The distribution quantities of chest X-ray radiographs in the three classifications of the COVID-QU-Ex dataset.

Class	Data
Infiltration	9547
Atelectasis	4215
Effusion	3955
Nodule	2705
Pneumothorax	2194
Mass	2139
Consolidation	1310
Pleural Thickening	1126
Cardiomegaly	1093
Emphysema	892
Fibrosis	727
Edema	628
Pneumonia	322
Hernia	110
No Finding	60,361

Table 2. The distribution quantities of chest X-ray radiographs in the Fifteen classifications of the Chest X-ray14 dataset.

The digital X-ray images included in Chest X-ray14 are standardized at a resolution of 1024×1024 pixels in PNG format, ensuring uniformity for computational analysis and model training. Demographic details provided with the dataset indicate that the ages of the subjects, both male and female, do not exceed 90 years. In conjunction with this dataset expansion, our research endeavors include a comprehensive data analysis to assess the distribution, representation and potential biases inherent within the dataset. This examination aims to identify limitations and propose methodologies for mitigating potential impacts on deep learning model performance and generalization.

For the purpose of enhancing the dataset's variability while improving the robustness of the deep learning models, we implement data augmentation techniques specifically tailored to this domain. These techniques include the application of Gaussian blur, which simulates variations in image focus and can mimic the effect of different imaging equipment, and Contrast Limited Adaptive Histogram Equalization (CLAHE), which is employed to improve the visibility of important features in the images by enhancing their contrast. These augmentation strategies are selectively applied to ensure the preservation of crucial diagnostic features while augmenting the dataset.

CLAHE: The CLAHE enhances the contrast of X-ray images by performing histogram equalization on the local areas of X-ray films. It divides the image into multiple small regions, calculates the histograms and performs equalization within each region, respectively, and then combines the processed regions through methods such as bilinear interpolation to obtain the final results. In this way, it improves the visibility of areas with low contrast in the lung region, and can highlight the details of the lung region, such as lung textures, nodules, shadows, etc. Besides, it makes these features more prominent in the image, which helps the deep learning model to learn and extract key features better, thus improving the recognition accuracy of lung diseases.

Flipping: During the data preprocessing stage, random horizontal or vertical flipping operations are conducted on X-ray films to enhance data diversity. By expanding the training data set, the model can learn lung features from different perspectives. This not only enhances the model's generalization ability but also reduces its sensitivity to image orientation and improves the robustness of lung disease recognition.

Denoising: Denoising methods, such as mean filtering, median filtering, and Gaussian filtering, are operated by applying smoothing techniques to the pixels within the image or leveraging pre-learned noise patterns to eliminate noise, thereby enhancing the image quality. This reduction of noise interference renders the actual anatomical structures and pathological features of the lungs more discernible. It facilitates the model's more precise learning and identification of the characteristics associated with lung diseases, consequently augmenting the recognition accuracy and stability. Nevertheless, in the process of denoising, certain minute details of the image might be forfeited. Particularly for some subtle lesion characteristics, they could potentially be smeared or obliterated during the denoising procedure, which would subsequently undermine the model's recognition

performance. In this study, Gaussian filtering is predominantly employed to attenuate the noise in the image. The noise sources might encompass imaging artifacts, which could be caused by factors like equipment malfunctions or improper calibration, and low-quality scans that occlude crucial features. By diminishing the influence of noise on the model, the classification accuracy can be effectively elevated.

In the domain of medical image classification, the quality of images serves as a pivotal factor influencing the accuracy and reliability of diagnosis. Preprocessing constitutes an essential phase in the data preparation pipeline, substantially contributing to the enhancement of the intrinsic quality of standard medical images. Through the application of targeted preprocessing techniques, we aim to address and rectify specific imperfections and variability inherent in raw medical images. These techniques encompass a range of operations designed to improve various crucial aspects of image quality, such as contrast normalization, noise reduction, and geometric transformations. By optimizing these aspects, the preprocessing phase ensures that the resultant images are more conducive to the accurate classification by deep learning models.

Preprocessing and selection of optimum deep learning model

The Chest X-ray images employed in this study have varied width and height values, thus they were resized to 224×224 pixels before the training process. The reshape size of 224×224 pixels was selected to allow us to do some data augmentation. Each deep learning model could internally resize the input images to fit its structure. Since deep learning models require a massive quantity of data to increase their performance, data augmentation is one solution for dealing with imbalanced data in the training sets. Under our preprocessing phase, the Chest X-ray images in the training set were rescaled (i.e., image magnification or reduction) using the ratio of 5/6 to 1/6, Zoom range of 0.75 to 0.95, rotation range equal to 1 and horizontal flip. The rotation range specifies the span under which the images were spontaneously rotated throughout training.

The first step of this study was to select a neural network model from some classical models to obtain an appropriate model with high classification accuracy on the Chest X-ray dataset by employing transfer learning methods. The four pre-trained classical deep learning models, ResNet50[26], MobileNetV2[27], ViT[14] and MaxViT[16], were analyzed to find the optimum effective deep learning approach for the lung lesion classification task. Then we managed to improve the selected optimum model structure to improve the prediction classification accuracy and find the robustness model for the chest X-ray images.

MobileNetV2[27]

The architecture of MobileNetV2 consists of several key components: convolutional blocks (Conv Blocks), inverted residual blocks and linear bottlenecks. At the core of this design lies depth-wise separable convolution, which comprises two distinct processes: depth-wise convolution and point-wise convolution. In the depth-wise convolution, each channel of the input feature map undergoes independent convolution with different 3×3 filters, enabling the extraction of diverse feature representations within the same receptive field, thereby enriching the model's parameterization. Following the depth-wise convolution, point-wise convolution serves as a feature integration mechanism, employing 1×1 convolutional kernels to merge channel information, facilitating information fusion. This process simplifies the network architecture while preserving the integrity of information flow through residual connections.

MaxVit[16]

The MaxVit architecture can be conceptualized as a comprehensive framework that integrates the advantages of CNNs by accommodating both local and global spatial correlations. It effectively mitigates the computational burden from quadratic to linear complexity through the strategic decomposition of the spatial dimensions. This architecture is structured hierarchically, comprising distinct layers: the convolutional layer, the hybrid transformer Layer, the multi-scale feature fusion layer, and the classification head.

The heart of the MaxVit model lies in the hybrid transformer layer, which is pivotal due to its incorporation of dual attention mechanisms: local attention and global attention. The local attention mechanism is akin to the selective focus observed in human perception, concentrating on particular segments of a comprehensive scene to distill crucial visual details, thereby enhancing differentiation capabilities. This process confines the attention scope to local data segments, allowing the model to prioritize regional attributes more effectively. Conversely, the global attention mechanism transcends spatial constraints, capturing interrelations between disparate locations across the axis, independent of their spatial separation. This facilitates a more nuanced extraction of local details while fostering the synthesis of local and global features, thus augmenting the informational breadth accessible to the model.

Transfer learning

In our experimental setup, we employed a hybrid training methodology, which involves utilizing pre-trained models widely recognized within the medical domain alongside our improved algorithms for comparative analysis. Traditional deep learning paradigms typically need initiating training from an absolute baseline, a process that inherently requires a substantial volume of domain-specific data to achieve model convergence. To circumvent these limitations, our approach leveraged transfer learning techniques which facilitate the development of resilient models even under data scarcity constraints.

Transfer learning not only enables our models to inherit and refine complex features from extensive, preexisting datasets but also enhances the models' adaptability and generalization capabilities across diverse medical scenarios. This strategy is particularly beneficial in the medical imaging sphere, where data privacy concerns and the high costs associated with dataset compilation pose significant challenges. By leveraging transfer learning, our models initially assimilate knowledge from publicly available datasets, thereafter

undergoing specialized refinement to cater to specific medical applications, thereby streamlining the training process and mitigating associated costs.

The proposed model: LungMaxVit

In the traditional domain of computer vision, convolution has become the cornerstone for feature extraction. A variety of convolution layers have emerged, with different convolutional modules leading to diverse outcomes in deep learning. These modules assist models in varying capacities, such as the Inception module, depthwise separable convolution, and dilated convolution. The inception module parallelizes different sizes of convolution kernels and pooling layers to capture multi-scale feature information within the same layer, significantly enhancing model performance. Depthwise separable convolution improves performance by integrating information across channels. Dilated convolution introduces additional "holes" to increase the receptive field, thereby boosting performance. Despite the continuous innovation and enhancement of convolution modules, they possess inherent structural limitations. By utilizing the attention mechanism, we aim to address the shortcomings in global information interaction.

Inspired by the MaxVit framework, we proposed an improved solution by enhancing its convolutional modules and optimizing the receptive field and information extraction capabilities. We noted that while the transformer sufficiently extracts global information in the MaxVit model, there should be have some room for improvement in convolutional information extraction to enhance model performance. We observed that traditional feature extraction in Chest X-ray images may not capture critical features, as they tend to be more concealed compared to features in standard classification tasks. We managed to refine the original model for better adaptation to Chest X-ray lung disease detection by improving initial convolutional layers to enhance feature extraction and representation.

The architecture of the proposed LungMaxVit model is shown in Fig. 2, we introduced an improved backbone model, LungMaxVit, which combines extra initial CNN with SE module and MaxVit backbone to implement better detection of lung disease in the medical field. By extracting deeper features through conv2d, followed by BatchNorm2d and Gelu activation, our network pays closer attention to image details. The inclusion of a squeeze-and-excitation (SE) module for spatial attention helps to capture detailed spatial features of the images, allowing the model to adjust inter-channel relationships adaptively to enhance performance. In the SE module, reducing the number of channels from 64 to 16 amplifies important weights in detail features. By reinforcing feature extraction with subsequent conv2d layers, our model undergoes an artful change in content extraction at the initial stages, which is more suited to medical image classification and requires more detailed feature extraction to adapt to medical classification tasks significantly.

Initial state block

In the overall architecture of our model, we particularly emphasize the initial state block and the SE block to extract the features from Chest X-ray images. The initial state block primarily consists of three convolutional segments where the first convolutional segment can extract features from the Chest X-ray image. We compute the matrix of outputs of one of three CNN module as Eq. (6).

$$F_{out} = (F_{in} - kernel + 2padding)/stride + 1 \tag{1}$$

where F_{out} is the output matrix of a convolution module while F_{in} is its input matrix. The kernel is a neural network filter that moves through a picture, scanning each pixel and turning the data into a smaller or bigger format. Padding is added to the medical image to aid the kernel in processing the image by providing more room for the kernel to cover the image. Stride determines how the filter convolves over the input matrix. In our study, the stride is set to 2.

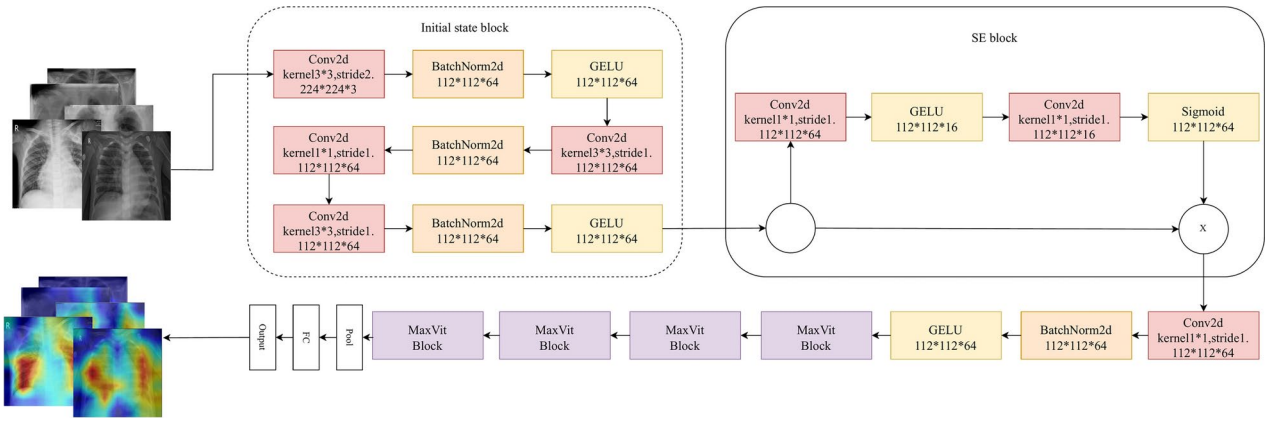

Fig. 2. The architecture of the proposed LungMaxViT model for multi-classification lung disease prediction from Chest X-ray images.

SE block

Following the initial convolutions, we utilize a series of kernels sized 3×3, 1×1, and 3×3 in succession to enhance the model's early receptive field and its extraction of fine details from images. The SE block supplements the prior module by addressing the need for both an expanded receptive field and improved feature extraction. It compacts global information into a single channel through global average pooling. As demonstrated in Eq. (7), z_c refers to the size reduction of the channel u_c, serving as a guide for the feature map u_c, which is derived from a global statistical vector calculated using Eq. (7). This vector consolidates global information over the channels.

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{2}$$

MaxViT block

Self-attention allows for spatial mixing of entire spatial locations while also benefiting from content-dependent weights based on normalized pair-wise similarity. The attention function is the mapping to an output of a set of keys, value pairs and a query. We refer to scaled dot-product Attention shown as Eq. (8). The input consists of queries and keys of dimension d_k, and values of dimension d_v. We compute the dot products of the $\sqrt{d_k}$ query with all keys, divide each by d_k, and apply a softmax function to obtain the weights on the values. The attention matrix contains the set of queries (Q), the keys (K) and values (V), which are used to compute the attention function simultaneously.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V \tag{3}$$

Multi-head attention presented by Eqs. (9) and (10) allows the model to jointly attend to information from different representation subspace at different positions.

$$\text{MultiHead}(Q,K,V) = \text{Concat}\ (\text{head}_1, \ldots, \text{head}_h)\, W^0 \tag{4}$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{5}$$

where the projections are parameter matrices $W_i^Q \in R^{d_{\text{model}} * d_k}$, $W_i^K \in R^{d_{\text{model}} * d_k}$, $W_I^V \in R^{d_{\text{model}} * d_v}$ and $W_i^0 \in R^{hd_v * d_{\text{model}}}$.

However, the original self-attention described above is location-unaware so that relative self-attention has been improved by introducing a relative learned bias added to the attention weights, which has been shown to consistently outperform original attention on many vision tasks. Relative attention has been explored in several previous studies for both NLP and vision. In our study, we used the relative attention mechanism described in Eq. (11).

$$\text{RelAttention}(Q,K,V) = \text{softmax}\left(QK^T/\sqrt{d} + B\right)V \tag{6}$$

where $Q, K, V \in \mathbb{R}^{(H \times W) \times C}$ are the query, key, and value matrices, and d is the hidden dimension. The attention weights are co-decided by a learned static location-aware matrix B and the scaled input-adaptive attention $\frac{QK^T}{\sqrt{d}}$. Considering the differences in 2D coordinates, the relative position bias B is parameterized by a matrix $B \in \mathbb{R}^{(2H-1)\times(2W-1)}$.

MaxViT block unified MBConv and Multi-axis attention layers. MBConv layers prior to the attention block, which can be regarded as conditional position encoding, make a model free of explicit positional encoding layers. The Multi-axis attention we leveraged here can be implemented without modification to the self-attention operation. The Multi-axis attention can be implemented by block and grid operator descried as Eqs. (12) and (13), respectively, to extract Chest X-ray image feature in a spatially-local small window. The unblock ( ) operation is denoted as the reverse of the above block partition procedure. These elements can be easily plugged into many vision architectures, especially on high-resolution tasks that can benefit by global interactions with affordable computation.

$$\text{Block: } (H,W,C) \to \left(\frac{H}{P} \times P, \frac{W}{P} \times P, C\right) \to \left(\frac{HW}{P^2}, P^2, C\right). \tag{7}$$

$$\text{Grid: } (H,W,C) \to \left(G \times \frac{H}{G}, G \times \frac{W}{G}, C\right) \to \underbrace{\left(G^2, \frac{HW}{G^2}, C\right) \to \left(\frac{HW}{G^2}, G^2, C\right)}_{\text{swapaxes(axis1=-2,axis2=-3)}} \tag{8}$$

To this end, we manage to explain hybrid CNN and multi-axis attention architecture as following. Assume x to be the input feature, given an input tensor $x \in \mathbb{R}^{H \times W \times C}$, the whole pipeline of the LungMaxVit model processing can be described from Eqs. (14)–(17), X1 is denoted as matrix output of initial state block and SE block.

$$\mathrm{X1} \leftarrow \mathrm{Proj}(\mathrm{Pool2D}(\mathrm{Initial\ \ state\ \ Block\ (x)})) + \mathrm{Proj}(\mathrm{SE(\ Initial\ \ state\ \ Block\ (x))} \tag{9}$$

SE is the squeeze-excitation layer[17], while Proj is the shrink Conv1x1 to down-project the number of channels. Pool2d that 2D max pooling over an input signal composed of several input planes is used to simplify CNN parameters.

The local Block Attention can be expressed as follows:

$$\mathrm{X2} \leftarrow \mathrm{MBConv(X1)} + \mathrm{Unblock(RelAttention(Block(LN(X1))))} \tag{10}$$

LN denotes the Layer Normalization, where MLP is a standard MLP network consisting of two linear layers.

$$\mathrm{X2} \leftarrow \mathrm{X2} + \mathrm{MLP(LN(X\,2))} \tag{11}$$

The global, dilated Grid Attention module is formulated as:

$$\mathrm{X3} \leftarrow \mathrm{X2} + \mathrm{UnGrid(RelAttention(Grid(LN(X2))))} \tag{12}$$

where we apply the RelAttention operation in Eqs. (15) and (17) for simplicity instead of Eq. (11).

$$\mathrm{X}output \leftarrow \mathrm{Unblock(X3 + MLP(LN(X\,3)))} \tag{13}$$

LN denotes the Layer Normalization[2], where MLP is a standard MLP network consisting of two linear layers: x ? W2GELU(W1x). Xoutput is output of lung disease classification prediction by LungMaxVit.

Results

In our comprehensive experiment, we utilized two datasets across four classic models alongside our improved model. We conducted transfer learning for each model, utilizing pre-trained weights provided by the official PyTorch website. These models were then applied to two specific datasets: ChestX-ray14 and COVID-QU-Ex (COVID-19) for ablation studies. We assessed the performance using key metrics prevalent in deep learning to determine their effectiveness. Additionally, we employed heat maps to present the explainable model, focusing on the lung lesion area.

Experimental environment

All experiments were performed on a Nvidia A100 server with the following hardware specifications: Intel Xeon Gold 5218 CPU @ 2.3 GHz (base frequency) with a maximum turbo frequency of 3.9 GHz, comprising a total of 64 cores (4 CPUs, each with 16 cores and 32 threads), 512.0 GB RAM, and supporting 5 NVIDIA A100 GPUs with 40 GB memory each.

Evaluation methods

To evaluate the transfer learning models, various evaluation metrics have been employed to evaluate the performance of the proposed hybrid deep learning framework, including Accuracy (ACC), Specificity, Sensitivity/Recall, Precision and F1-score. The parameters are calculated by applying a confusion matrix generated for each individual model. Accuracy is calculated to determine the percentage of correct predictions, while precision is calculated to determine the probability of positive classifications. Specificity determines the percentage of correctly predicted negative classifications from all performance parameters. Contrasting specificity, the recall is employed to determine the percentage of correctly predicted positive classes. The F1 score is used to determine the balance between specificity and recall. The performance parameters are expressed in the following: Eqs. (1)–(5).

$$\mathrm{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\mathrm{Sensitivity\ /Recall} = \frac{TP}{TP + FN} \tag{16}$$

$$\mathrm{Specificity} = \frac{TN}{TN + FP} \tag{17}$$

$$F1\text{-Score} = 2\left(\frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}\right) \tag{18}$$

Accuracy: The accuracy intuitively reflects the overall correct detection level of the model and enables us to quickly understand the general performance of the model in the detection of lung diseases. A high accuracy rate means that the model can accurately determine whether patients have lung disease in most cases, and whether patients can be diagnosed as sick or healthy. The model can provide relatively reliable references for doctors and patients and reduce the possibility of misdiagnosis and missed diagnosis.

AUC: The AUC (Area Under the Curve) comprehensively takes into account the true positive rate and the false positive rate of the model under different thresholds, and can fully evaluate the model's discriminative ability for lung diseases. In the detection of lung diseases, as the distribution of diseased and non-diseased samples may overlap, the model needs to accurately identify the diseased samples while minimizing misjudgments of the non-diseased samples as much as possible. The higher the AUC value is, the stronger the model's ability to distinguish between diseased and non-diseased samples will be. It can maintain better detection performance under different disease risk thresholds and provide doctors with more accurate diagnostic bases.

F1-Score: In the detection of lung diseases, both precision and recall are of great importance. High precision indicates that when the model judges that a patient has lung disease, the result has a relatively high credibility, which reduces unnecessary further examinations and the psychological burden on patients. High recall ensures that most patients with diseases can be detected by the model, reducing the risk of missed diagnoses. The F1-score combines the advantages of precision and recall and can evaluate the performance of the model in lung disease detection more comprehensively. When the F1-score is high, it means that the model can accurately detect patients with diseases while also ensuring the reliability of the detection results. In the classification of lung diseases, problems caused by under-reporting (false negatives) and false positives can be avoided.

In unbalanced data sets, traditional metrics like accuracy might not be the best indicators. The confusion matrix allows for a more nuanced model evaluation and helps improve understanding and performance in underrepresented classes, especially crucial in medical domains where precise categorization is vital.

In the study of the model on the 14 chest radiograph dataset, shown in Fig. 3, even though LungMaxVit exhibited superior performance compared to other models while identifying a more significant number of diseases with substantial precision, the model also encountered challenges due to an imbalanced data set with fifteen categories, resulting in subpar performance in those classes with underrepresented features such as 'No Finding' and 'Effusion.' The analysis reveals that the LungMaxVit model has learned a good representation of the majority of lung diseases from Chest X-ray images, where feature extraction and disease identification are inherently more challenging in multi-classification.

As seen in Fig. 4, the confusion matrix reveals significant accuracy across the entire test dataset: 2346 out of 2352 cases were correctly identified for COVID-19, only identified wrong with 2 cases for Non-COVID and 4 cases for Normal category, 2169 out of 2265 for Non-COVID, and 2053 out of 2171 for Normal category, highlighting the model's effectiveness in accurate disease classification, especially for COVID-19 prediction. Normal classification performance is not as good enough as that of the other cases due to similar feature representation among the mild lung disease cases and normal cases.

Each disease exhibits distinct common pathological locations, and since the scope involves pulmonary diseases, pathological manifestations can occur anywhere within the lungs, significantly complicating model classification tasks. Despite these challenges, LungMaxVit stands out across various metrics in ablation studies, providing it to be the most effective model among the study models.

Comparative analysis of LungMaxVit and other pre-trained models

Accuracy (ACC), Area Under the Curve (AUC) and F1-score were employed to evaluate the performance of the proposed LungMaxViT model and other pre-trained models. As seen in Fig. 5, LungMaxVit surpassed all the other models in all metrics on the COVID-QU-Ex dataset, achieving 96.8% accuracy which is higher than others by margins up, AUC of 98.3%, which surpassed other models by up and F1-score of 96.7%, which exceed all other competitors. Overall, LungMaxVit demonstrated superior performance across all key metrics among the five neural network models.

In our comprehensive disease comparison, particularly for multi-class disease classification, we managed to find a solution for lung disease feature extraction in imbalanced Chest X-ray datasets by contrasting the performance metrics of various models for each disease category. For the COVID-19 category in our COVID-QU-Ex test set, we evaluated the models using precision, recall, and F1-score. Precision assesses the ratio of true positives to the total predicted positives while recall complements it by measuring the proportion of actual positives correctly identified. LungMaxVit outperformed the performance of the other models in the COVID-19 category evaluation in precision, recall and F1-score. From Table 3, the precision of the three categories evaluated by LungMaxVit are 1.00, 0.96, 0.95, outperforming one percent of the other models by COVID-19 category and Normal category, while one percent less than MaxVit by Non-COVID. The F1 scores evaluated by LungMaxVit all exceed or are all equal the other models by the three categories. The recalls evaluated by the LungMaxVit are 0.98, 0.96 and 0.96, exceeding five percent over the ResNet model by Normal, while one percent less than Maxvit on COVID-19 and Normal. In summary, the LungMaxvit shows prominent robustness and feature presentation over the other models.

In the Chest X-ray14 dataset, we evaluated our model's performance by using metrics such as AUC and F1 scores. The AUC and F1 score of the LungMaxVit are 0.932 and 0.707, respectively, which outperforms all other models, especially exceeding ResNet model by by 5.9 percent using AUC, as shown in Fig. 6.

In Table 4, we compare the AUC and F1-score for all 14 lung diseases and normal (no Finding case) with 5 models. The AUC of LungMaxVit exceeds or equals all the other models for the 15 classification, especially over 10 percent Nodule. In the context of multi-class classification, it is crucial to examine the metrics for

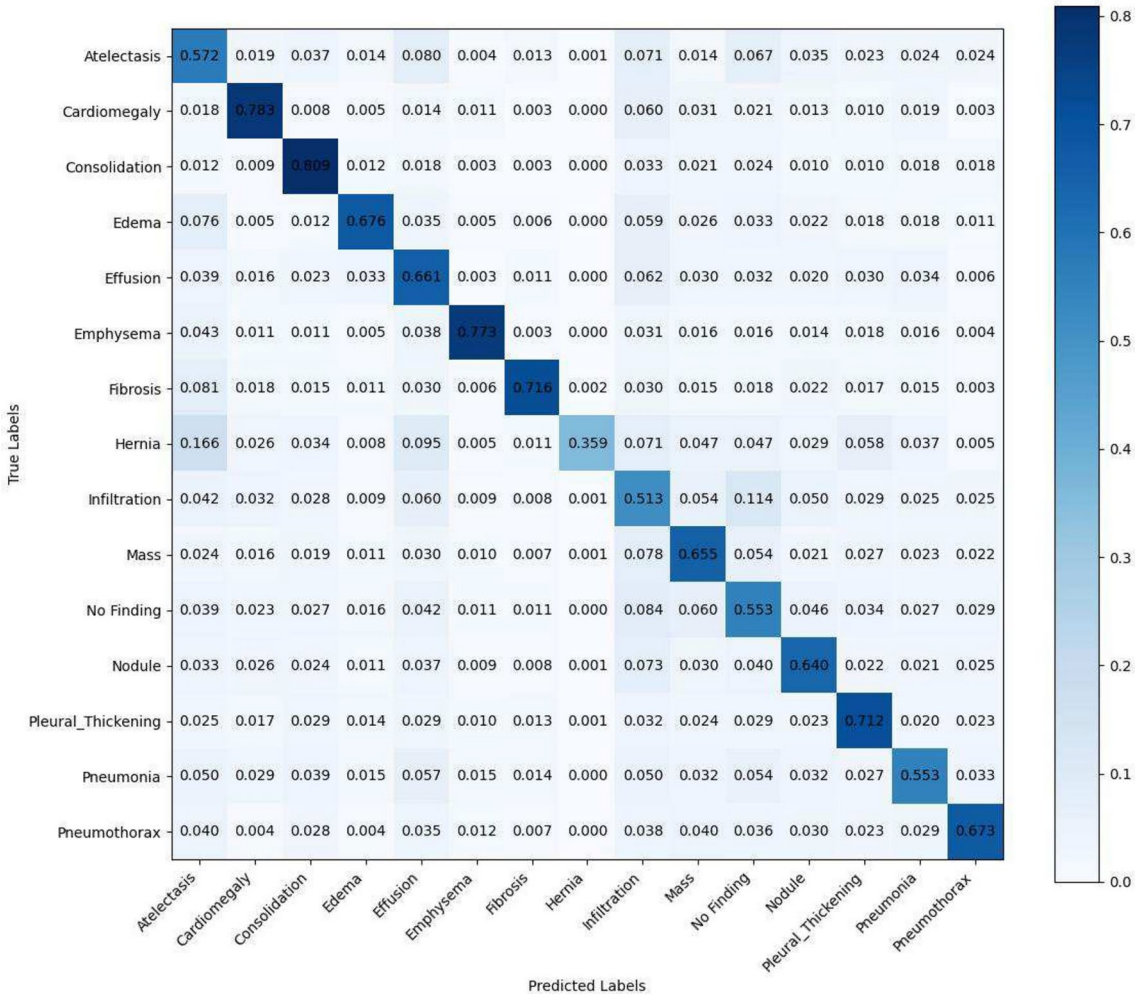

Fig. 3. Confusion matrix of LungMaxViT model for Chest X-ray14 datasets.

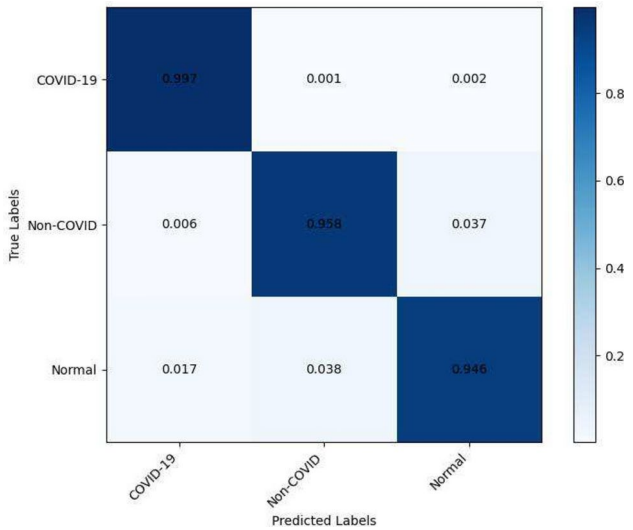

Fig. 4. Confusion matrix of LungMaxViT model for COVID-19 datasets.

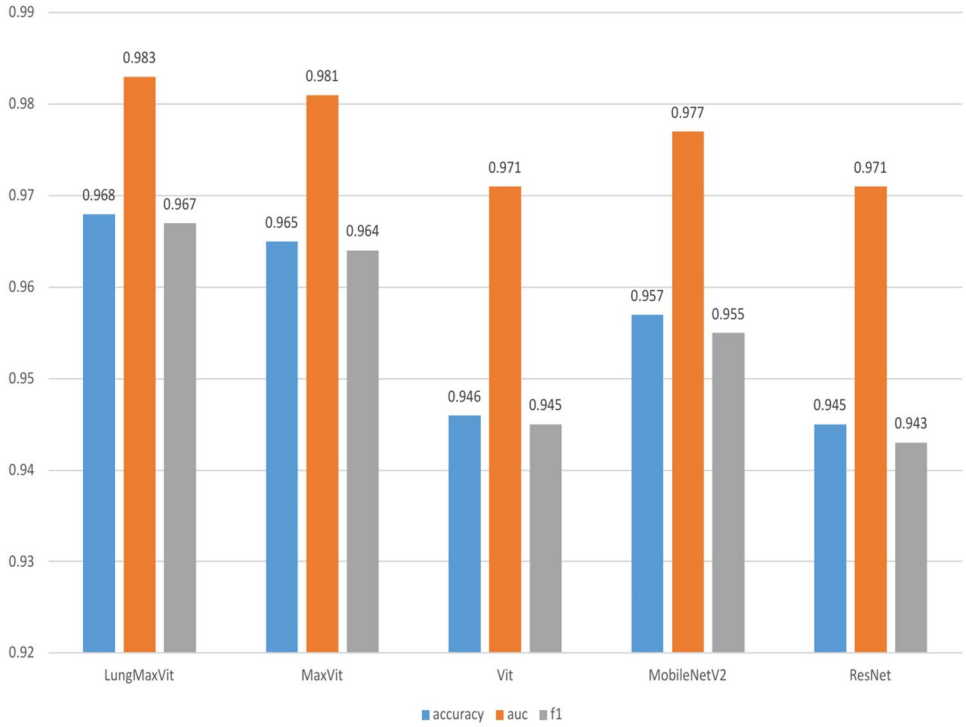

Fig. 5. Comparative analysis of all models for COVID-19 dataset.

Class	Evaluation metrics	Transfer learning models				Our method
		MaxVit	Vit	MobileNetV2	ResNet	LungMaxVit
COVID-19	Precision	0.99	0.99	0.99	0.99	1.00
	Recall	0.99	0.97	0.98	0.97	0.98
	F1-score	0.99	0.98	0.98	0.98	0.99
Non-COVID	Precision	0.97	0.92	0.96	0.91	0.96
	Recall	0.94	0.94	0.93	0.95	0.96
	F1-score	0.96	0.93	0.95	0.93	0.96
Normal	Precision	0.94	0.93	0.92	0.93	0.95
	Recall	0.97	0.92	0.96	0.91	0.96
	F1-score	0.95	0.93	0.94	0.92	0.95

Table 3. The comparative evaluation results of the five deep learning models on COVID-19 datasets.

each category to accurately gauge the model's classification capabilities. Notably, the model showed subpar performance in the 'Infiltration' and 'No Finding' and 'Atelectasis' categories. However, achieving exemplary classification across all categories in a multi-class context faces several challenges, especially when dealing with issues such as imbalanced data distribution and limited dataset size. Despite these challenges, LingMaxVit showed the superior performance on this dataset among the five study models.

Discussion

In the comparative experiments, all models were trained using SGD with a learning rate of 0.001 and momentum of 0.9, throughout the training and validation process. LungMaxVit leverages advanced convolutional networks for feature extraction using initial state block and integrating transformer mechanisms to avoid saturation with the two datasets, demonstrating robust performance compared to conventional convolutional networks and purely ViT.

As seen in Figs. 7 and 8, the training loss of LungMaxVit on the COVID-19 dataset is the lowest among all five models as well as the accuracy is much higher than that of Vit, Resnet and Mobilenet models despite much less distinction with MaxVit. On the other hand, the AUCs of LungMaxVit on the Chest X-ray 14 dataset are the highest level among all the five pretraining models, as seen in Fig. 9. Although the loss of the LungMaxVit is not the lowest among all the models, the loss dropped dramatically among the five models, as seen in Fig. 10.

Table 5 shows the comparison between our LungMaxVit model and the existing works in the literature on the COVID dataset. The study paper reference, datasets, study models and reported accuracy are presented

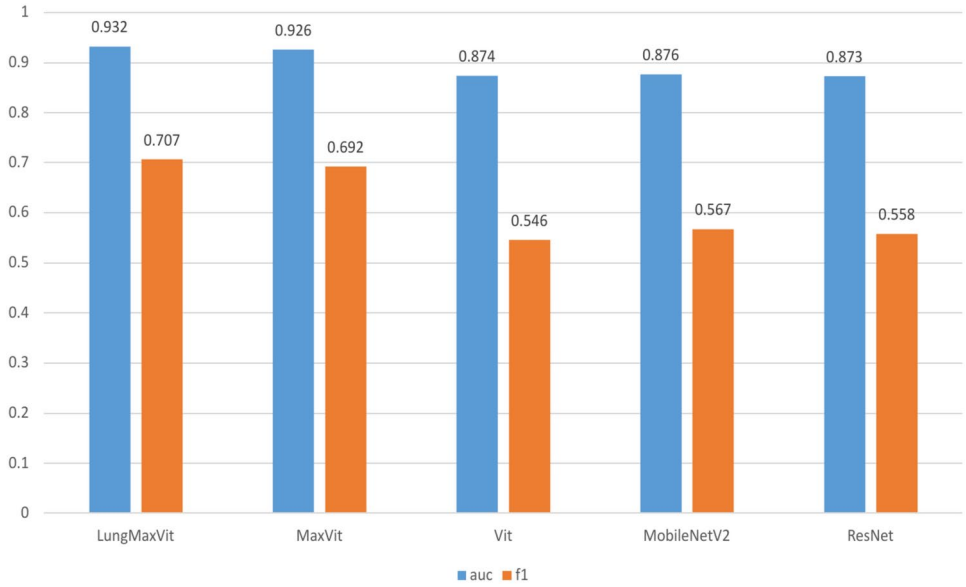

Fig. 6. Comparative analysis of all models for Chest X-rays 14 dataset.

Class/models	Evaluation metrics	Transfer learning models				Proposed
		MaxVit	Vit	MobileNetV2	ResNet	LungMaxVit
Atelectasis	auc	0.86	0.82	0.82	0.82	0.86
	F1-score	0.53	0.42	0.45	0.44	0.54
Cardiomegaly	auc	0.98	0.96	0.96	0.96	0.99
	F1-score	0.84	0.74	0.74	0.72	0.86
Consolidation	auc	0.95	0.89	0.89	0.89	0.95
	F1-score	0.77	0.62	0.63	0.63	0.80
Edema	auc	0.99	0.96	0.96	0.96	0.99
	F1-score	0.88	0.75	0.77	0.77	0.88
Effusion	auc	0.89	0.87	0.87	0.88	0.90
	F1-score	0.56	0.50	0.51	0.50	0.55
Emphysema	auc	0.98	0.94	0.95	0.94	0.99
	F1-score	0.83	0.70	0.73	0.72	0.86
Fibrosis	auc	0.98	0.93	0.94	0.93	0.99
	F1-score	0.84	0.66	0.69	0.67	0.84
Hernia	auc	0.99	0.97	0.96	0.96	0.99
	F1-score	0.89	0.71	0.73	0.70	0.88
Infiltration	auc	0.81	0.75	0.73	0.74	0.81
	F1-score	0.44	0.32	0.32	0.33	0.45
Mass	auc	0.92	0.85	0.85	0.86	0.92
	F1-score	0.67	0.54	0.54	0.56	0.71
No finding	auc	0.81	0.73	0.73	0.73	0.82
	F1-score	0.44	0.32	0.34	0.33	0.46
Nodule	auc	0.87	0.78	0.80	0.78	0.88
	F1-score	0.55	0.38	0.41	0.40	0.56
Pleural thickening	auc	0.93	0.86	0.86	0.85	0.94
	F1-score	0.71	0.50	0.54	0.51	0.71
Pneumonia	auc	0.98	0.91	0.90	0.88	0.99
	F1-score	0.81	0.59	0.62	0.60	0.84
Pneumothorax	auc	0.90	0.83	0.85	0.85	0.91
	F1-score	0.64	0.46	0.48	0.50	0.66

Table 4. Graphs illustrating the quantities of various diseases.

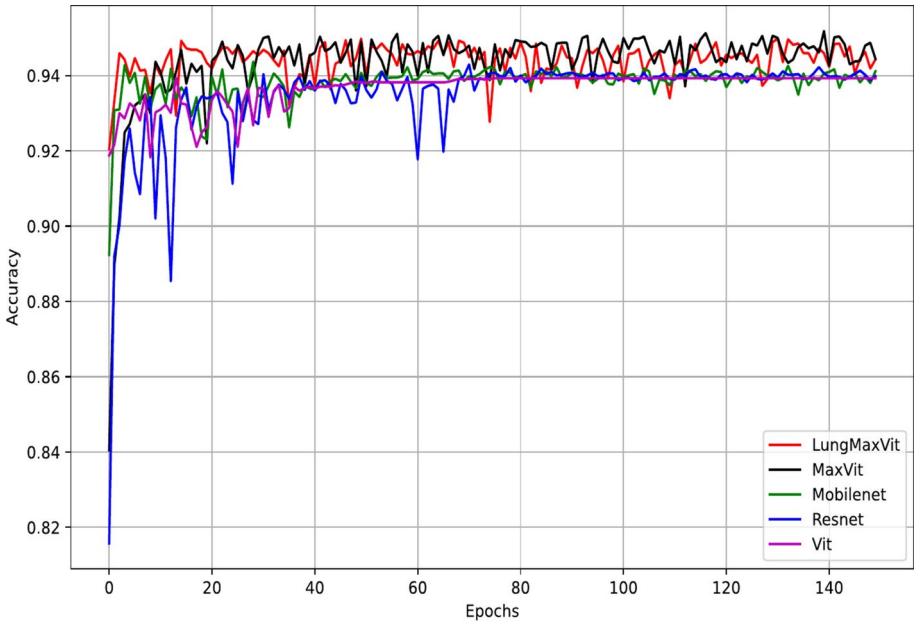

Fig. 7. The comparative analysis among the LungMaxViT and the other pre-training models of validation accuracy of 150 epochs with COVID-19 dataset.

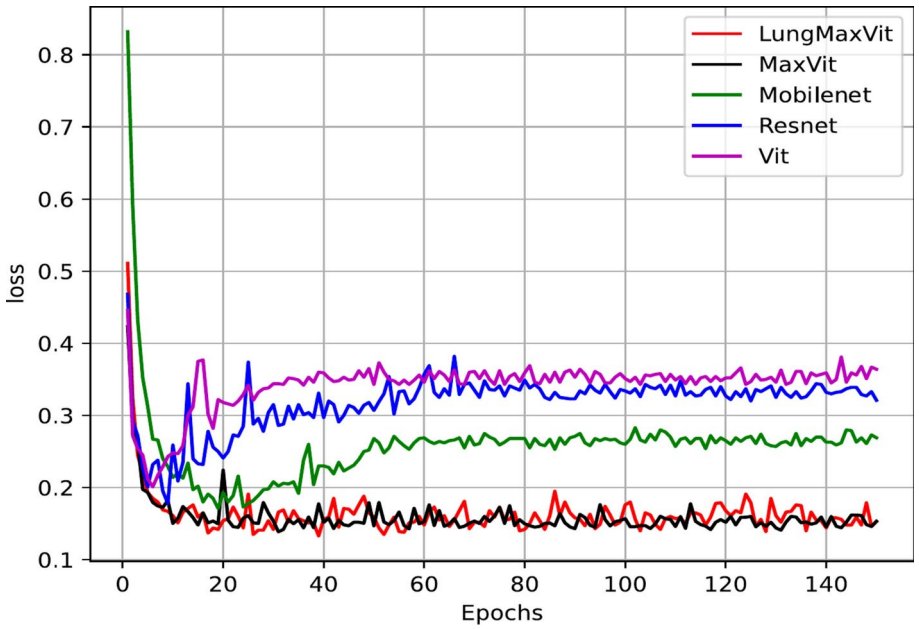

Fig. 8. The comparative analysis among the LungMaxViT and the other pre-training models of validation loss of 150 epochs with COVID-19 dataset.

in the table. Seen from Datasets, we compared with are three classes, including Normal, Pneumonia and COVID-19. In this study, we achieved 96.8% accuracy with our proposed model across the three classes. The accuracy of our model is higher than Bunyodbek Ibrokhimov with the same datasets. Even though the accuracy of Apostolopoulos is 98.0%, which is the highest of all relative models, they used small samples of test images to evaluate their model performance on COVID-19 datasets, having only have 224 samples in the three classes. Even though our model did not achieve the highest performance in terms of classification accuracy, we believe the proposed methodology in this study has substantially better accuracy. This is because we used the largest datasets with over 33,900 X-ray images. By the way, we find that the increase in the normal categories of data in the dataset will introduce the interference factor in the ability to distinguish between these three categories of data.

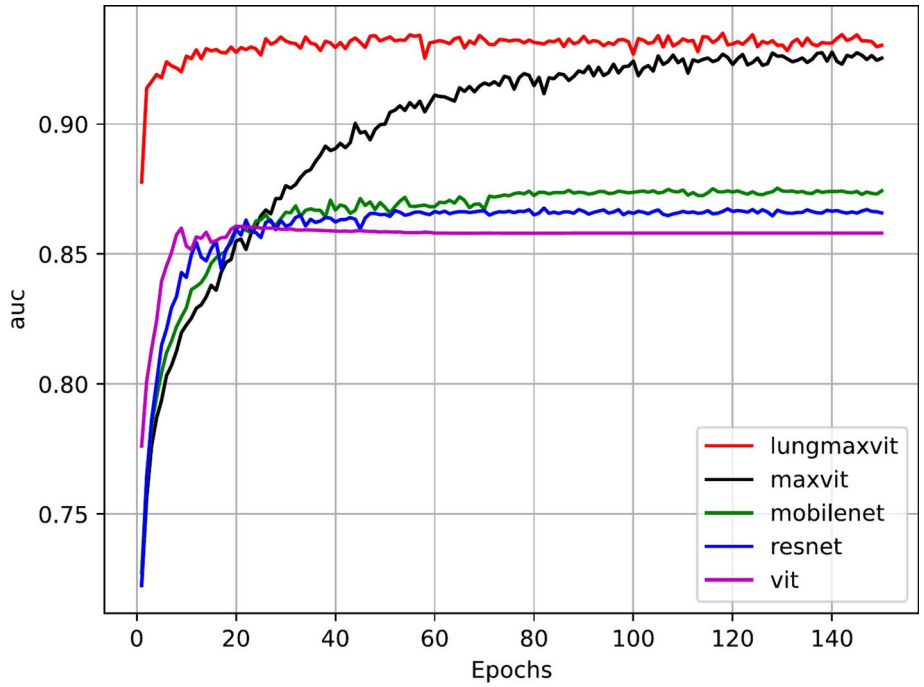

Fig. 9. The comparative analysis among the LungMaxViT and the other pretraining models of validation AUC of 150 epochs with Chest X-ray 14 dataset.

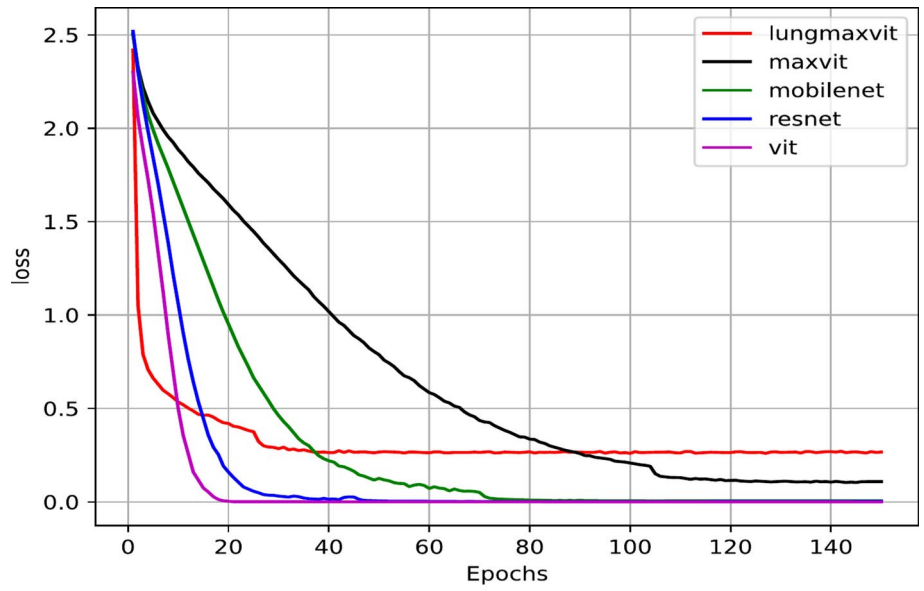

Fig. 10. The comparative analysis among the LungMaxViT and the other pre-training models of validation loss of 150 epochs with Chest X-ray 14 dataset.

Table 6 contains the related AUC values. The quantitative analysis presents that LungMaxVit achieved the best results compared to other models. The Mean AUC of our model is 93.2% while the Z-Net Mean AUC is 85.8% as the second good results. The AUC value of every disease of our model exceeds that of all other relative models. The "Edema" (AUC = 0.994), "Emphysema" (AUC = 0.989), "Cardiomegaly" (AUC = 0.990), "Hernia" (AUC = 0.997) and "Pneumonia" (AUC = 0.988) are perfectly classified by the proposed approach when compared to other diseases. Especially, "Hernia" has a high performance (AUC = 0.997 as compared to others) due to data pre-processing through data augmentations.

The successful neural networks and artificial intelligence models are usually applied in a black box manner where no information is provided about what exactly makes them arrive at their predictions. Since lack of

Study paper	Datasets	Study models	Accuracy
Ozturk et al.[28]	1000 Normal		
	500 Pneumonia	DarkCovidNet	0.870
	125 COVID-19		
Wang et al.[29]	8066 Normal	COVIDNet	
	5538 Pneumonia	VGG19	0.933
	358 COVID-19	ResNet50	
Apostolopoulos et al.[30]	504 Normal		
	714 Pneumonia	VGG19, Inception,	0.980
	224 COVID-19	Xception, MobileNet	
Bunyodbek Ibrokhimov et al.[31]	10,701 Normal		
	11,263 Pneumonia	VGG19, ResNet50	0.966
	11,956 COVID-19		
Proposed LungMaxVit	10,701 Normal		
	11,263 Pneumonia	LungMaxVit	0.968
	11,956 COVID-19		

Table 5. Comparison of the improved methodology with the related methods on the COVID dataset.

Pathology	Wang et al.[32]	CheXNet[33]	Thorax-Net[34]	Guan et al.[35]	Z-Net[36]	Proposed LungMaxVit
Atelectasis	0.706	0.779	0.750	0.785	0.821	0.866
Consolidation	0.708	0.754	0.741	0.763	0.746	0.958
Infiltration	0.613	0.689	0.681	0.699	0.722	0.816
Pneumothorax	0.789	0.851	0.825	0.871	0.898	0.910
Edema	0.835	0.849	0.835	0.850	0.864	0.994
Emphysema	0.815	0.924	0.842	0.924	0.923	0.989
Fibrosis	0.769	0.821	0.804	0.831	0.764	0.987
Effusion	0.736	0.826	0.818	0.835	0.889	0.904
Pneumonia	0.633	0.735	0.693	0.738	0.755	0.988
Pleural Thickening	0.708	0.792	0.776	0.746	0.784	0.945
Cardiomegaly	0.814	0.881	0.871	0.899	0.872	0.990
Nodule	0.716	0.781	0.714	0.775	0.744	0.881
Mass	0.560	0.830	0.799	0.838	0.840	0.928
Hernia	0.767	0.932	0.902	0.922	0.768	0.997
Mean AUC	0.813	0.818	0.787	0.822	0.858	0.932

Table 6. Comparison of AUC scores of LungMaxVit with latest research on the ChestX-ray14 dataset.

transparency can be a major drawback in medical applications, the development of methods for visualizing, explaining and interpreting deep learning models has recently attracted increasing attention[37].

LIME, a model-agnostic interpretation approach, hinges on performing linear approximation on a model within local areas to elucidate its prediction outcomes. Due to the necessity of generating a substantial number of perturbed samples and conducting model predictions, LIME can be rather time-consuming when handling complex models and large-scale datasets.

SHAPE is predominantly applicable to medical data with time-series characteristics, such as physiological signals of patients under dynamic monitoring (e.g., electrocardiogram, electroencephalogram, respiratory signals, etc.). For non-time-series data types, like static medical images (e.g., X-ray films, MRI, etc.) or medical records in textual format, this method has relatively limited applicability.

Class Activation Mapping (CAM) is primarily based on the feature maps output by the last convolutional layer of a convolutional neural network. CAM calculates the weights of each feature map channel and generates a heatmap via weighted averaging to display the regions where the model focuses on during classification. Moreover, it relies on the fully connected layers immediately following the last convolutional layer. Gradient-weighted Class Activation Mapping (Grad-CAM)[38] is an extension of CAM which generates class activation maps (CAM) based on the gradient information of the target category with respect to the last convolutional layer. It can effectively leverage the structural characteristics of CNNs, particularly the information of convolutional layers. In the study such as X-ray film disease diagnosis, CNN is a mainly employed model architecture. Grad-CAM can efficiently mine the image regions related to disease diagnosis in the CNN model. Through the generated class activation maps, one can directly identify on the image which regions play a crucial role in the model's specific category

prediction (e.g., pneumonia diagnosis). These regions are typically highlighted in the maps, enabling doctors or researchers to intuitively comprehend the decision-making basis of the model.

In the field of medical imaging for image recognition, the application of Grad-CAM is particularly significant. It can be employed in the analysis of X-rays, CT scans, and other medical images to assist in disease diagnosis, while also highlighting the affected areas. This aids medical professionals in making educated decisions by providing insights into the visual features associated with diseases. Furthermore, it enables a deeper understanding of whether deep learning models are correctly identifying relevant features, thereby informing algorithm design and performance enhancement.

The Grad-CAM model as explanation method used in this study are not much different from previous studies. As a popular deep learning CNN model visual explanation method, it demonstrates that our proposed model architecture with attention mechanism focus on the chest lesions without identify some wrong feature patterns through Figs. 11, 12 and 13.

Figure 11 displays the heatmap of COVID-19. Notably, the dataset did not furnish any information regarding the lesion sites. Nevertheless, it is evident that the model's focus predominantly lies within the lung region. It can be reasonably inferred that, for a lung disease model, when its attention is chiefly centered on the lung area, it can provide valuable assistance to doctors in the pursuit of locating lesion sites.

Turning to Fig. 12, this showcases the heatmap of an 8-class subset derived from Chest X-rays14, and it incorporates detailed information about the lesion sites corresponding to eight distinct disease categories. As can be clearly seen, the model's attention is almost entirely concentrated within the areas of the lesion sites. This convincingly validates that the classification rationale underpinning our trained model, LungMaxViT, is accurate.

Finally, Fig. 13 represents a heatmap covering various classes of Chest X-rays14. Similar to Fig. 11, its primary purpose is to offer auxiliary support to doctors during the process of identifying and localizing lesion sites.

In summary, Grad-CAM is an important diagnostic tool in the realms of deep learning and medical imaging. It can also be applied to evaluate the effectiveness of models, such as the two discussed previously. By focusing on the lung areas, Grad-CAM can discriminate between classes more accurately, better expose the trustworthiness of our model and help identify biases in datasets, providing valuable insights into their operational efficacy and areas of focus.

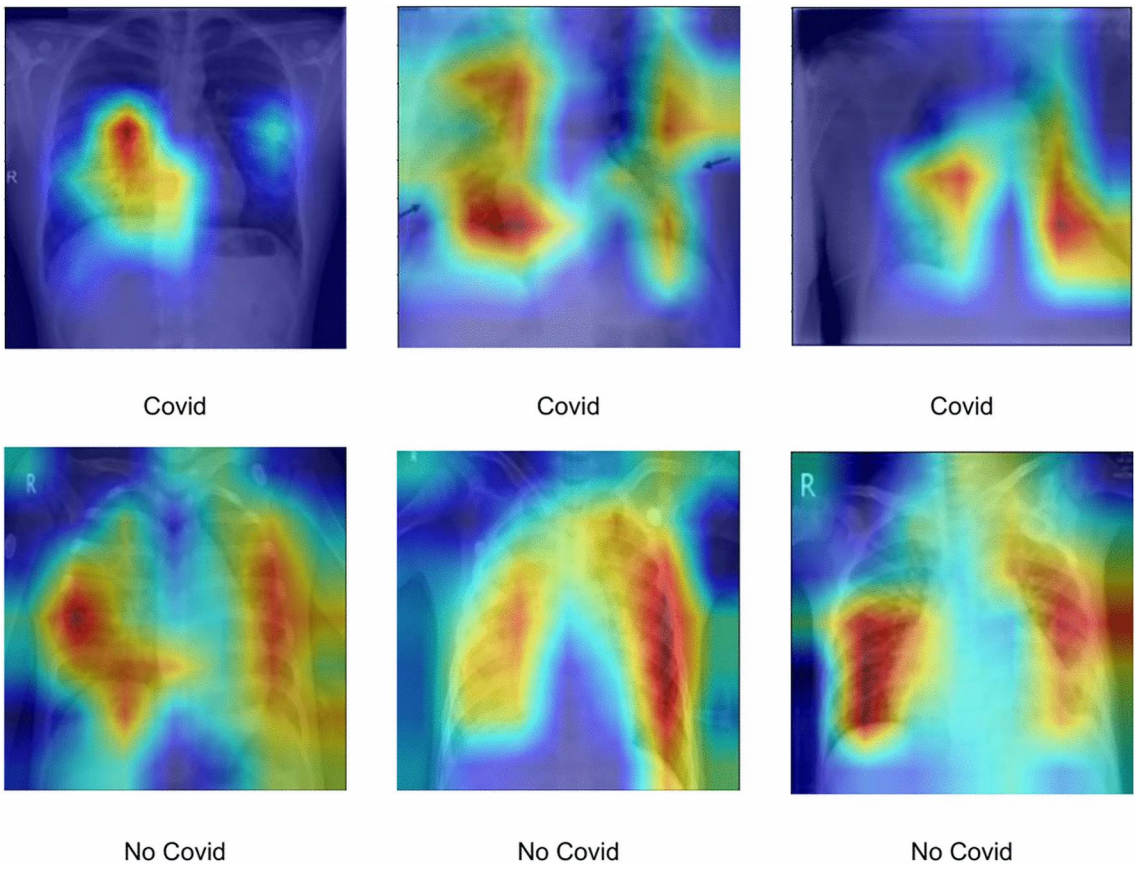

Fig. 11. Visualization of lung infections in X-ray images using Grad-CAM on LungMaxViT model with COVID-19 dataset.

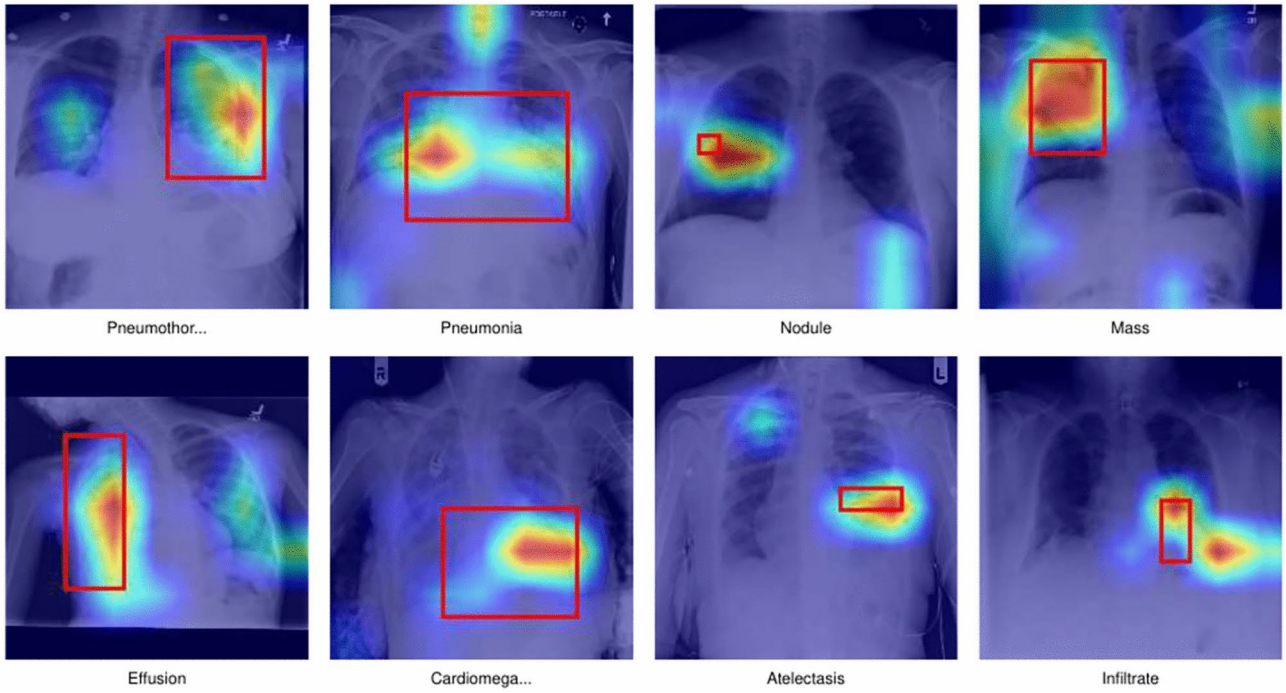

Fig. 12. Visualization of lung infections in X-ray images using Grad-CAM on LungMaxViT model with Chest X-rays14 dataset.

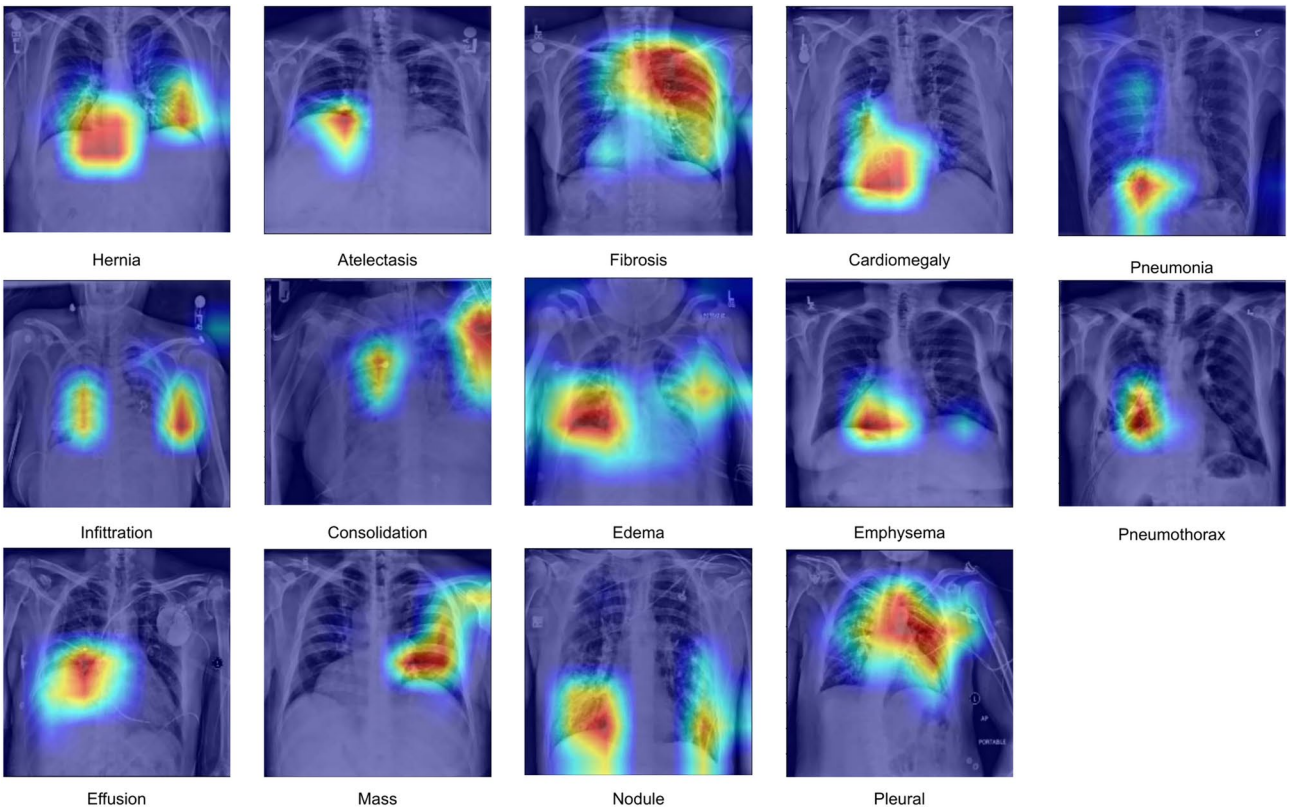

Fig. 13. Visualization of lung infections in X-ray images using Grad-CAM on LungMaxViT model with Chest X-rays14 dataset.

Conclusion

It is well known that lung disease is a major cause of mortality worldwide. This paper proposes a novel approach to predict the lung diseases from Chest X-ray images by blending the backbone of CNN with a multi-axis transformer for better accuracy and robust performance. In this study, Chest X-ray14 and COVID-19 datasets are examined to demonstrate the proposed model's effectiveness and generalization for various lung diseases. After image preprocessing, the LungMaxViT framework is pre-trained by imagenet-1k to obtain the appropriate initial parameters. Then, four transfer learning models and LungMaxViT are applied to the two datasets. The deep learning models, including ResNet50, MobileNetV2, ViT, Maxvit and lungMaxViT show the prediction accuracy results of 94.5%, 95.7%, 94.6%, 96.5% and 96.8%, respectively, for COVID-19 as well as AUC results of 87.3% , 87.6%, 87.4%, 92.6% and 93.2% for Chest X-ray14. F1-score outperforms all other models for the two datasets, demonstrating that the proposed model performs with much improved efficiency over the other models. The proposed LungMaxViT does not only predict COVID-19 from the other lung lesions but also has better performance than the other models for 14 classifications of lung diseases. The proposed model has better multi-class classification results than the other classical deep learning methods based on the used evaluation metrics, showing its superiority and efficiency in the early identification of COVID-19 and other lung diseases from Chest X-ray images. Furthermore, the proposed hybrid model shows its capability to provide more reasonable and accurate explainable identification results in terms of heat maps.

Medical foundation models, exhibiting considerable generalization and adaptability, have immense potential in solving a wide range of downstream tasks, as they can help to accelerate the development of accurate and robust models, reduce the dependence on large amounts of labeled data, and preserve the privacy and confidentiality of patient data. For future work, we will study the novel fine-tuned medical foundation model and various multi-modal self-supervised models to classify various medical image datasets, containing other types of diseases that may be detectable using medical imaging.

Data availability

The COVID-19 datasets generated during and analysed during the current study are available in the kaggle repository, https://www.kaggle.com/datasets/anasmohammedtahir/covidqu. The Chest X-rays14 datasets generat ed during and/or analysed during the current study are available in the nihcc repository, https://www.kaggle.com/datasets/nih-chest-xrays/data/data.

Received: 23 September 2024; Accepted: 14 February 2025
Published online: 24 February 2025

References

1. Yang, L. et al. Covid-19: immunopathogenesis and immunotherapeutics. *Signal Transduct. Target. Ther.* **5**, 128. https://doi.org/10.1038/s41392-020-00243-2 (2020).
2. Ruuskanen, O., Lahti, E., Jennings, L. C. & Murdoch, D. R. Viral pneumonia. *Lancet* **377**, 1264–1275. https://doi.org/10.1016/S0140-6736(10)61459-6 (2011).
3. Jaeger, S. et al. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **4**, 475. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20 (2014).
4. Rajpurkar, P. et al. Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225https://doi.org/10.48550/arXiv.1711.05225 (2017).
5. Rajpurkar, P. et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS Med.* **15**, e1002686. https://doi.org/10.1371/journal.pmed.1002686 (2018).
6. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25** (2012).
7. Rahman, H. et al. A systematic literature review of 3d deep learning techniques in computed tomography reconstruction. *Tomography* **9**, 2158–2189. https://doi.org/10.3390/tomography9060169 (2023).
8. Van Boxtel, J., Vousten, V., Pluim, J. & Rad, N. M. Hybrid deep neural network for brachial plexus nerve segmentation in ultrasound images. In *2021 29th European Signal Processing Conference (EUSIPCO)*, 1246–1250. https://doi.org/10.23919/EUSIPCO54536.2021.9616329 (IEEE, 2021).
9. Mujahid, M. et al. Pneumonia classification from x-ray images with inception-v3 and convolutional neural network. *Diagnostics* **12**, 1280 (2022).
10. Ali, M. et al. Pneumonia detection using chest radiographs with novel efficientnetv2l model. *IEEE Access* **12**, 34691–34707. https://doi.org/10.1109/ACCESS.2024.3372588 (2024).
11. Alvi, S. B. K. et al. A lightweight deep learning approach for covid-19 detection using x-ray images with edge federation. *Digit. Health* **9**, 20552076231203604 (2023).
12. Quasar, S. R. et al. Ensemble methods for computed tomography scan images to improve lung cancer detection and classification. *Multimed. Tools Appl.* **83**, 52867–52897 (2024).
13. Brauwers, G. & Frasincar, F. A general survey on attention mechanisms in deep learning. *IEEE Trans. Knowl. Data Eng.* **35**, 3279–3298. https://doi.org/10.1109/TKDE.2021.3126456 (2021).
14. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929https://doi.org/10.48550/arXiv.2010.11929 (2020).
15. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022 (2021).
16. Tu, Z. et al. Maxvit: Multi-axis vision transformer. In *European Conference on Computer Vision*, 459–479. https://doi.org/10.1007/978-3-031-20053-3_27 (Springer, 2022).
17. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7132–7141. https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html (2018).
18. Hodson, E., Thayer, D. & Franklin, C. Adaptive gaussian filtering and local frequency estimates using local curvature analysis. *IEEE Trans. Acoust. Speech Signal Process.* **29**, 854–859. https://doi.org/10.1109/TASSP.1981.1163641 (1981).
19. Vaezi, M. M. & Bavarian, B. Contrast-dependent spread filters. In *Image Processing Algorithms and Techniques*, vol. 1244, 100–107. https://doi.org/10.1117/12.19500 (SPIE, 1990).

20. Jeong, H. & Kim, C.-I. Adaptive determination of filter scales for edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **14**, 579–585. https://doi.org/10.1109/34.134062 (1992).
21. Reza, A. M. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *J. VLSI Signal Process. Syst. Signal Image Video Technol.* **38**, 35–44. https://doi.org/10.1023/B:VLSI.0000028532.53893.82 (2004).
22. Zuiderveld, K. *Contrast Limited Adaptive Histogram Equalization*, 474–485. https://doi.org/10.5555/180895.180940 (Academic Press Professional, 1994).
23. Wang, X. et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2097–2106. https://openaccess.thecvf.com/content_cvpr_2017/html/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.html (2017).
24. Database, Q.-C. https://www.kaggle.com/aysendegerli/qatacov19-dataset (Accessed 14 March 2021).
25. Kaggle. Covid-19 radiography database. https://www.kaggle.com/tawsifurrahman/covid19-radiography-database (Accessed 14 March 2021).
26. 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*, 770–778. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html (2016).
27. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520 (2018).
28. Ozturk, T. et al. Automated detection of covid-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* **121**, 103792 (2020).
29. Wang, J., Peng, Y., Xu, H., Cui, Z. & Williams, R. O. The covid-19 vaccine race: challenges and opportunities in vaccine formulation. *Aaps Pharmscitech* **21**, 1–12 (2020).
30. Apostolopoulos, I. D. & Mpesiana, T. A. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Phys. Eng. Sci. Med.* **43**, 635–640 (2020).
31. Ibrokhimov, B. & Kang, J.-Y. Deep learning model for covid-19-infected pneumonia diagnosis using chest radiography images. *BioMedInformatics* **2**, 654–670 (2022).
32. Wang, X. et al. Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2097–2106 (2017).
33. Rajpurkar, P. et al. Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017).
34. Wang, H., Jia, H., Lu, L. & Xia, Y. Thorax-net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography. *IEEE J. Biomed. Health Inform.* **24**, 475–485 (2019).
35. Guan, Q. et al. Discriminative feature learning for thorax disease classification in chest X-ray images. *IEEE Trans. Image Process.* **30**, 2476–2487 (2021).
36. Ahmad, Z., Malik, A. K., Qamar, N. & Islam, S. U. Efficient thorax disease classification and localization using DCNN and chest X-ray images. *Diagnostics* **13**, 3462 (2023).
37. Samek, W., Wiegand, T. & Müller, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 (2017).
38. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html (2017).

Acknowledgements

The authors are grateful to the support of the NSFC (62372494), Guangdong Engineering Centre Project (2024GCZX001), Specialized Talent Training Program (2024001), Virtual Teaching Group (2022002) and Key Disciplines Project (2024ZDJS142).

Author contributions

X.F. and Y.L. conceived the experiment(s), X.F. and R.L. conducted the experiment(s), X.F., R.L., W.D., A.T. and Y.L. analysed the results. All authors reviewed the manuscript.

Additional information

Correspondence and requests for materials should be addressed to Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025