MÜHENDİSLİK ve DOĞA BİLİMLERİ FAKÜLTESİ

Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi
International Journal of Engineering Research and Development

UMAGD, (2024) 16(2), 760-776.

10.29137/umagd.1469472

Cilt/Volume:16 Sayı/Issue:02 Haziran/June 2024

Araştırma Makalesi / Research Article

Improved Vision Transformer with Lion Optimizer for Lung Diseases Detection

Akciğer Hastalıklarının Tespiti için Lion Optimizer ile Geliştirilmiş Görü Transformatörü

İshak Pacal [1]

[1]*Iğdır Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü,76000, Iğdır, TÜRKİYE*

***Başvuru/Received**:16/04/2024 **Kabul/Accepted**: 14/05/2024. **Çevrimiçi Basım/Published Online:** 30/06/2024*
***Son Versiyon/Final Version**: 30/06/2024*

Abstract

Lung infections, such as pneumonia, bronchitis, tuberculosis, and notably COVID-19 caused by the SARS-CoV-2 virus, have caused widespread devastation globally, resulting in a significant loss of life. Timely and precise diagnosis of these respiratory diseases is crucial in controlling their spread and reducing their deadly impact. However, diagnostic errors can occur due to factors like physician workload and the need for a second opinion. To address these challenges, artificial intelligence-based diagnostic systems, utilizing deep learning algorithms, particularly in the radiology field, have been proposed. In this research, we introduced a novel model based on Multi-Axis Vision Transformer (MaxViT), which boasts a reduced parameter count, decreased GPU computational load, real-time diagnostic capabilities, and improved accuracy. Furthermore, we conducted a detailed performance comparison of optimization algorithms, including SGD, Adam, and Lion, with higher results indicating that the Lion optimizer notably enhances the diagnostic capabilities of the proposed MaxViT model, especially in detecting lung infections. Our proposed approach underwent rigorous experimentation using the COVID-QU-Ex dataset, recognized as the most current, comprehensive, and balanced dataset for lung infections and COVID-19. Our method achieved diagnostic accuracy of 97.14%, surpassing existing models while maintaining significantly fewer parameters.

Key Words
"Detection of Lung infections, COVID-19, MaxViT, Vision transformer, Deep Learning"

Öz

Pnömoni, bronşit, tüberküloz ve özellikle SARS-CoV-2 virüsü tarafından neden olan COVID-19 gibi akciğer enfeksiyonları, küresel olarak yaygın yıkıma neden oldu ve önemli bir can kaybına yol açtı. Bu solunum yolu hastalıklarının zamanında ve doğru teşhisi, yayılmanın kontrol altına alınması ve ölümcül etkisinin azaltılması açısından hayati öneme sahiptir. Ancak, hekim iş yükü ve ikinci bir görüşe duyulan ihtiyaç gibi faktörler nedeniyle teşhis hataları ortaya çıkabilir. Bu zorlukları ele almak için, özellikle radyoloji alanında derin öğrenme algoritmalarını kullanan yapay zeka tabanlı teşhis sistemleri önerilmiştir. Bu araştırmada, Çok-Eksenli Görüntü Dönüştürücüler temelli yeni bir model sunduk, bu model, azaltılmış parametre sayısı, azaltılmış GPU hesaplama yükü, gerçek zamanlı teşhis yetenekleri ve artan doğruluk gibi özelliklere sahiptir. Ayrıca, SGD, Adam ve Lion dahil olmak üzere optimizasyon algoritmalarının detaylı bir performans karşılaştırmasını yaptık ve etkili sonuçlar, Lion optimizatörünün MaxViT modelinin teşhis yeteneklerini özellikle akciğer enfeksiyonlarını tespitte önemli ölçüde artırdığını göstermektedir. Önerdiğimiz yaklaşım COVID-QU-Ex veri kümesi kullanılarak sıkı bir şekilde deneyime tabi tutuldu ve bu veri kümesi, akciğer enfeksiyonları ve COVID-19 için en güncel, kapsamlı ve dengeli veri kümesi olarak kabul edilmektedir. Yöntemimiz, mevcut modelleri aşarak %97,14'lük bir teşhis doğruluğuna ulaştı ve bunu yaparken belirgin şekilde daha az parametre kullandı.

Anahtar Kelimeler
"Akciğer enfeksiyonlarının tespiti, COVID-19, MaxViT, Görü transformatörü, Derin Öğrenme"

***Sorumlu Yazar:** ishak.pacal@igdir.edu.tr*

1. Introduction

Lung infections are a frequent disease that may be brought on by a number of microorganisms, including bacteria, viruses, fungi, and parasites (Cookson et al., 2018). Pneumonia, bronchitis, and tuberculosis are the most typical lung infections (Cookson et al., 2018). Numerous fatalities have recently been brought on by COVID-19 and COVID-19-induced lung infections. COVID-19 is a respiratory disease that emerged in the Chinese city of Wuhan in 2019 and caused by a virus called SARS-CoV-2 (Platto et al., 2020). This virus is airborne between humans and usually produces mild or moderate symptoms. However, the elderly and people with chronic diseases can have serious complications. COVID-19 symptoms contain weakness, runny nose, sore throat, nasal congestion, weakness, skin rash, nausea, diarrhea, headache, fever, cough, or visual disturbances (Ciotti et al., 2020). However, symptoms do not occur in everyone, and some people may not experience symptoms.

A significant detrimental effect of the COVID-19 epidemic has been felt globally, and many countries, albeit slightly (Velavan & Meyer, 2020). Health systems have been under a great burden to combat the pandemic and countries have implemented different measures to combat the pandemic. These include measures such as maintaining social distance, wearing masks, staying at home, avoiding going to public places, and businesses that are kept open on a limited basis. PCR tests are generally employed to identify the pathogen causing COVID-19, but these tests can sometimes be misleading (Yuki et al., 2020). X-rays are primarily used to effectively diagnose COVID-19 and measure the degree of infection in the lungs, and if necessary, computed tomography (CT) images are used (Yuki et al., 2020). Thus, the definitive diagnosis of the patient is made. X-ray images are both less harmful to health and an economical and faster technique compared to CT.

Physicians and radiologists who are experts in their fields can diagnose diseases using x-rays or other imaging modalities. X-ray images are widely used for the detection of lung infections and COVID-19 due to their efficiency (Soomro et al., 2022). Moreover, these images are typically easily accessible in the majority of medical institutions and offer a more efficient alternative to traditional laboratory examination, thus contributing to their overall benefits. A specialist physician can diagnose a patient's COVID-19 or lung infections by looking at the x-ray image (Cleverley et al., 2020). In some cases, it may misdiagnose the patient, especially in cases where the physician has an intense work tempo, absent-mindedness, or is not a secondary physician. Therefore, computer-aided diagnostic systems (CADx) are developed, especially in the field of radiology (Pacal et al., 2020). Recently, there have been serious developments in CAD systems (Aslan, 2024; Ayan et al., 2022; Pacal, 2024a). Particularly with recent advancements in artificial intelligence and the growing significance of deep learning algorithms, numerous new developments have emerged in the literature. As in the diagnosis of many diseases, it can be said that artificial intelligence has gained the most important place in the diagnosis of lung infections and COVID-19 (Tahir et al., 2021).

Artificial intelligence, a continually evolving field, has long been employed in disease diagnosis. Recently, deep learning has become particularly prominent in image analysis, securing a significant role in diagnostic systems and yielding successful outcomes (Karaman et al., 2022). Currently, deep learning is the most widely used methods, gaining popularity due to its success in various fields, including defense industry, autonomous vehicles, natural language processing, object detection, medical image processing (Aslan & Özüpak, 2024; Dönmez, 2024; Işık & Paçal, 2024; Kılıçarslan et al., 2024). Compared to classical machine learning approaches, deep learning is a discipline that provides effective results with large amounts of data, can automatically make feature discoveries, and contains many more complex and successful features (PACAL, 2022). While deep learning-based methods are extensively used in medical imaging, they have facilitated intense research efforts in combating lung infections and COVID-19 (Subramanian et al., 2022). While deep learning reduces the burden of specialist physicians and employees, it can learn the experiences of many specialist physicians and can be used effectively in the diagnosis of disease (Pacal & Karaboga, 2021).

1.1. Background

To evaluate X-ray images and identify lung-related disorders, deep learning can be employed. Deep learning algorithms can be trained with learned examples and then recognize images that are like those examples. These algorithms can be used to detect signs of lung infections and COVID-19 in X-ray images (Bhattacharyya et al., 2022). For example, algorithms can be used to detect abnormal changes or tumors in the lung. Also, deep learning algorithms are used to support radiologists' diagnostic process. These algorithms can help radiologists speed up their work and make more accurate diagnoses. The literature features numerous studies that use deep learning algorithms on X-ray images to diagnose lung infections and COVID-19. Almost all these studies utilized Convolutional neural network (CNN) architects, which is a popular architecture of deep learning algorithms.

CNNs are a robust deep learning architecture that has become widely adopted, particularly in image analysis and medical image analysis. Numerous research studies have leveraged CNN techniques to achieve positive outcomes in the diagnosis of lung infections and COVID-19. Several of these studies that have employed deep learning techniques are highlighted as follows. Aslani and Jacob (2023) reviewed deep learning for detecting COVID-19 utilizing chest radiography. The evaluation examined many articles that employed 2D/3D deep CNNs for COVID-19 identification. The study explains how to detect COVID-19 and points out a few drawbacks of the suggested techniques. (Podder et al., 2023) proposed a new framework called LDDNet, which is based on the DenseNet201 model and is optimized for lung diseases. LDDNet has additional layers of batch normalization, dense and dropout layers, and 2D global average pooling compared to the basic DenseNet201 model. Bhattacharyya et al. (2022) proposed an integrated method

in order to identify COVID-19 from chest X-ray images. In the presented method, the first step is C-GAN based method to obtain lung images, the second step is using CNN networks for discriminant extraction, and the last step is using several machine learning algorithms for classification. The presented method provides greater accuracy compared to other approaches. Similarly, Deb et al. (2022) proposed a method to diagnose COVID-19 using a method consisting of chest X-ray images. This method is based on ensemble learning, and popular networks such as VGGNet, GoogleNet, DenseNet and NASNet have been used in practice.

With the use of pre-processed chest radiography images, Ahmad et al. (2022) demonstrated a cutting-edge multimodal deep learning strategy that can identify between infections with and without COVID-19. The model can categorize various types of pneumonia. Gafoor et al. (2022) proposed deep learning technologies for the diagnosis of COVID-19. This study has observed that deep learning algorithms effectively diagnose patients infected with COVID-19. Nayak et al. (2023) presented a new shallow CNN model called LW-CORONet, which extracts significant characteristics from chest X-ray (CXR) images. The model, with just five trainable layers, achieves high classification accuracy in both multi-class and binary cases across two large chest X-ray (CXR) datasets. Another study proposed an integrated method for the diagnosis of COVID-19 patients on X-rays by effectively using deep learning techniques and optimization algorithms together. While this method consists of CNN models such as ResNet, VGG, DenseNet, it has been stated that the proposed method in experimental studies is more successful than other studies (Dhiman et al., 2022). Devasia et al. (2023) presented a model that fines-tunes and applies transfer learning to the EfficientNetB4 architecture to detect active pulmonary tuberculosis using chest X-rays. The approach includes a multilabel methodology to detect lung zone. Sedik et al. (Sedik et al., 2022) proposed a method for the diagnosis of COVID-19 patients with a hybrid structure using CNN and LSTM architectures together. In this method, training and testing stages were carried out for both CT and X-ray images. The proposed method has yielded successful results in the diagnosis of COVID-19 disease. Alshmrani et al. (2023) introduced a deep learning architecture designed for the multi-class classification of various lung diseases such as Pneumonia, Lung Cancer, Tuberculosis, Lung Opacity, and COVID-19. The study involved a large number of chest X-ray (CXR) images that were resized, normalized, and randomly divided to suit the needs of deep learning.

Vision transformers represent a recent advancement in deep learning, especially within the field of computer vision (Pacal, 2024b). These neural networks are a subset of those constructed using the transformer design, which was initially created to address issues with natural language processing. Given their remarkable performance in image classification and object detection tasks, vision transformers are an interesting development in the world of medical imaging. Vision transformers may be used to examine the images and detect the presence of diseases in COVID-19 with X-ray images in the setting of lung infections. The transformer architecture allows the network to understand the meaning of the image and make predictions based on that context. This is important in medical imaging, where subtle changes in the image may indicate the presence of a disease.

1.2. Contributions

Deep learning algorithms have had tremendous success in the diagnosis of diseases of the lungs. Nearly all these studies are CNN-based studies, while a very few and effective part of them are vision transformer-based studies, which is one of the most popular architectures of deep learning nowadays. In this research, we present a proficient method using transformers for the diagnosis of lung infections and COVID-19 compared to existing studies in the literature. The main differences and contributions that distinguish our study from other studies are as follows.

- We utilize the latest vision transformer models in literature to diagnose lung infections and COVID-19, offering the most extensive study available to researchers.
- We utilize the MaxViT design for the initial time in identifying lung infections related to COVID-19, yielding the most promising outcomes.
- We introduce a unique multi-axis vision transformer model with real-time disease detection that is efficient, has fewer parameters, and performs better for lung infections and COVID-19 detection.
- This study provides a general comparison of CNNs and Vision Transformer architectures.
- We provided a comprehensive comparison of popular optimization algorithms, including Stochastic Gradient Descent (SGD), Evolved Sign Momentum (Lion) and Adaptive Moment Estimation (Adam), and furthermore, through an extensive evaluation of these optimization algorithms, this study concludes that the Lion optimizer significantly enhances the diagnostic accuracy of the MaxViT model. This contribution underscores the critical importance of selecting the appropriate optimizer for effectively fine-tuning deep learning models in medical diagnosis.
- We apply all these experimental studies to the COVID-QU-Ex dataset, which is the most up-to-date, novel, well balanced and large for lung infections and COVID-19.

2. Methodology

2.1. Vision Transformer

The transformer design, which was first described in the 2017 publication "Attention Is All You Need" by Google researchers, serves as the foundation for a particular sort of neural network architecture called a Vision Transformer (Vaswani et al., 2017). Instead of processing input sequences sequentially, as is the case with standard recurrent neural networks, the transformer design processes input sequences in parallel. The processing of images and the extraction of features from them can be done using a Vision Transformer in the context of computer vision (Dosovitskiy et al., 2020). The Vision Transformer model demonstrates remarkable performance in

image classification tasks, boasting state-of-the-art results. The model then applies self-attention mechanisms to the sequence of image patches to extract features. DeiT (Data-Efficient Image Transformer): This is an extension of the ViT which is specifically designed to be more data-efficient. The authors proposed to use a smaller architecture and smaller batch sizes during training to improve the model's performance on image classification tasks (Touvron et al., 2020). Swin Transformer: It is a variation of the Vision Transformer where the model uses a more complex spatial attention mechanism (Liu et al., 2021). T2T-ViT (Token-to-Token Vision Transformer): T2T-ViT is an extension of the Vision Transformer that incorporates token-to-token attention mechanisms, which allows the model to better capture the spatial relationships (Yuan et al., 2021). These are some of the examples of Vision Transformer architectures that have been proposed in recent research. It has been demonstrated that they perform well on computer vision tasks such as image classification. Recently, many transformer models have been suggested by many researchers for images because transformer models have become more advantageous and effective than CNN architectures in many respects (Tu et al., 2022).

2.2 MaxViT: Multi-Axis Vision Transformer

MaxViT, a multi-axis vision transformer, exemplifies a machine learning model designed to analyze visual data using transformer architecture (Tu et al., 2022). Integrating additional forms of focus, like spatial and channel attention, into the transformer framework empowers the model to boost its effectiveness in computer vision tasks. By including attention mechanisms that concentrate on particular regions or aspects of the input, the model is created to analyze visual data more efficiently. MaxViT's capability to handle massive and complicated visual information, such as high-resolution photos and films, is one of its main features. This makes it suitable for many different computer vision applications, including semantic segmentation, object identification, and image categorization. The model may also adapt to the unique properties of the input data thanks to its capacity to incorporate numerous attention mechanisms, which can enhance performance. The visual organization of the Vision Transformer-based MaxVit architecture is shown in Figure 1. The MaxVit architecture incorporates MBConv, block, and grid attention layers to create a novel form of base structure block in place of the conventional hierarchical design found in ConvNet architectures like ResNet. By drawing influence from sparse techniques, the MaxVit architecture is a novel sort of attention module. This reduces the secondary complexity of ordinary attention while preserving all local knowledge. The sequential design of MaxVit offers more simplicity and flexibility, and it also shows better performance compared to previous methods, as one of its biggest advantages is that each module can be combined individually or in any order. On the other hand, parallel designs do not provide this benefit. MaxViT's innovative architecture, Max-SA, can be constructed by hierarchically stacking various tiers of Max-SA, as depicted in Figure 1 (Tu et al., 2022). This is facilitated by the adaptability and scalability inherent in Max-SA. Through the integration of global and local receptive fields spanning the entire network, from shallow to deep layers, MaxViT demonstrates superior model capacity and generalization capabilities. Consequently, the MaxViT algorithm outperforms its rivals, marking a significant advancement in the field.

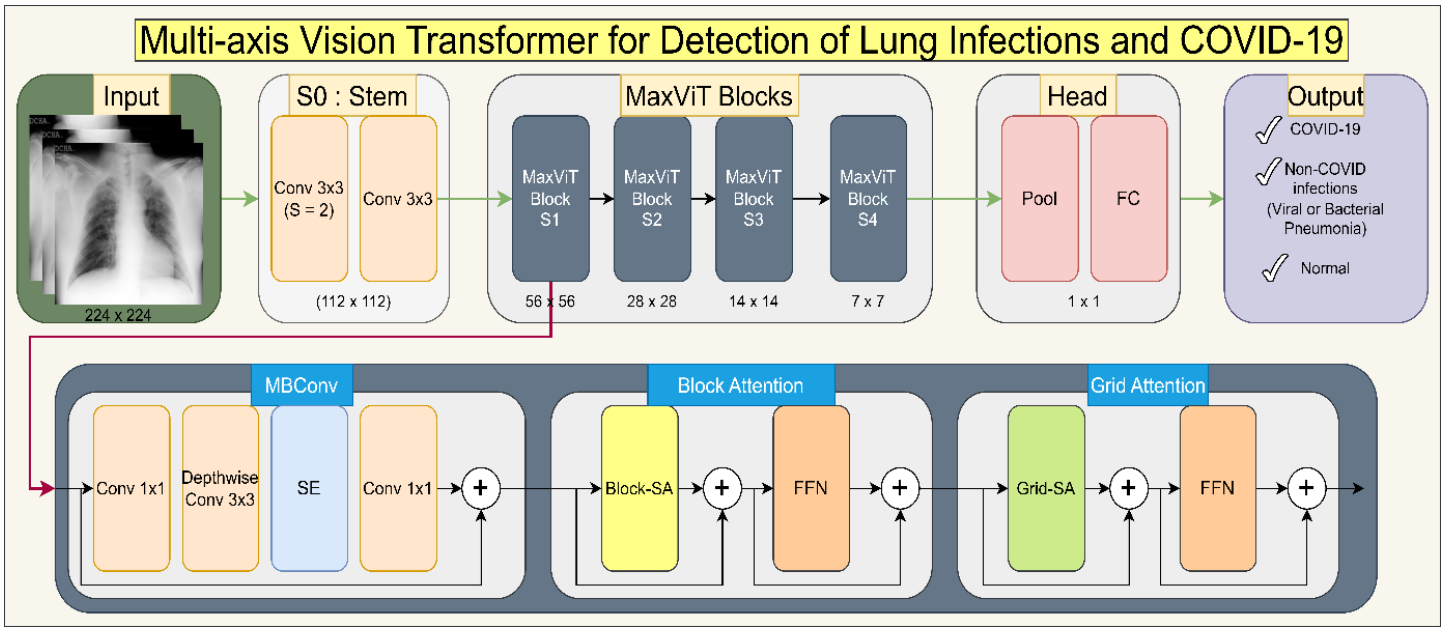


Figure 1. Overview of the Proposed MaxViT-based Lung Infection and COVID-19 Disease Detection Model (Tu et al., 2022).

2.3. Proposed Model

Deep learning algorithms for object classification are widely applied in medical image interpretation, demonstrating encouraging potential for upcoming advancements. Vision transformer architectures have achieved high success in recent years, surpassing CNN architectures in many areas (Pacal, 2024a). Most proposed deep learning algorithms utilize universal datasets such as ImageNet (Russakovsky et al., 2015), MS-COCO (Lin et al., 2014), and Pascal-VOC (M. et al., 2010) for image processing. Therefore, to assess the effectiveness of a given algorithm or architecture in medical image analysis, training and testing must be performed. In this research,

we evaluated the effectiveness of various vision transformer-based designs and a few CNN architectures in detecting COVID-19. Taking into account the performance provided by the MaxVit architecture, we suggest re-scaling it to make it more effective for COVID-19 detection.

Self-attention works better than local convolution by enabling global interactions in neural networks. However, the quadratic complexity of self-attention makes its application across whole space impracticable. The multi-axis approach of the Max-SA method is presented as a solution to this problem. By splitting the input feature map into discrete P x P windows, this method allows self-attention to function just on the local spatial dimension inside each window. This technique, called "block attention," successfully reduces the computational burden of implementing self-attention on a global scale. Local interconnections within the network framework are facilitated by it.

$$RelativeAttention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (1)$$

$$Block: (H,W,C) \rightarrow \left(\frac{H}{P}\ x\ P, \frac{W}{P}\ x\ P, C\right) \rightarrow \left(\frac{HW}{P^2}, P^2, C\right) \quad (2)$$

$$Grid: (H,W,C) \rightarrow \left(G\ x\ \frac{H}{G}, G\ x\ \frac{W}{G}, C\right) \rightarrow \left(G^2, \frac{HW}{G^2}, C\right) \rightarrow \left(\frac{HW}{G^2}, G^2, C\right) \quad (3)$$

According to equations Eq.1 to Eq.3, for a specified x value, $x \in R^{HxWxC}$ the concept of block attention on specific local features is elucidated in Eq. 4, whereas Eq. 5 presents the Grid Attention module designed for capturing global features.

$$x \leftarrow x + Unblock\left(RelAttention\left(Block(LN(x))\right)\right) and\ x \leftarrow x + GRN(LN(x)) \quad (4)$$

$$x \leftarrow x + Ungrid\left(RelAttention\left(Grid(LN(x))\right)\right) and\ x \leftarrow x + GRN(LN(x)) \quad (5)$$

In the Relative Attention process, we omit the QKV input format. Here, LN stands for Layer Normalization, and GRN denotes a novel version of MLP. Grid attention is a method that divides the tensor into a regular GxG grid to produce windows of variable sizes in order to achieve sparse global attention. Grid attention is defined by equation 2, wherein when self-attention acts across the grid axis (GxG), tokens are blended globally and dilatedly. With the same variables and FLOPs as the current attention module, the suggested Max-SA module can easily replace it. It also makes building simpler because it doesn't require cyclic shifting, masking, or padding. Figure 2 illustrates the multi-axis approach to self-attention computation with a 4x4 window/grid size (Tu et al., 2022). The grid-attention module utilizes a sparse, uniform grid spread across the entire 2D space to focus on pixels, while the block-attention module conducts self-attention within defined windows. Both strategies, employing fixed attention regions, exhibit linear complexity concerning input size. The self-attention operation spatially blends identical colors.

In order to build a single block that incorporates both local and global interaction mechanisms, MaxViT combines components from the Maxvit block, grid block attention, and local block attention. The network is organized hierarchically, much like standard ConvNets, with Conv3x3 layers being used for downsampling at the beginning (S0 conv-stem). It consists of four body phases, each of which doubles the channels but reduces the resolution by half from the step before it. Every stage's initial MBConv block uses downsampling, while MaxViT blocks are used throughout the backbone. The default expansion and shrink rates for the inverted bottleneck are 4 and 0.25, respectively. Attention blocks maintain a fixed attention head size of 32. As blocks per stage (B) and channel dimensions increase, the model size grows accordingly.

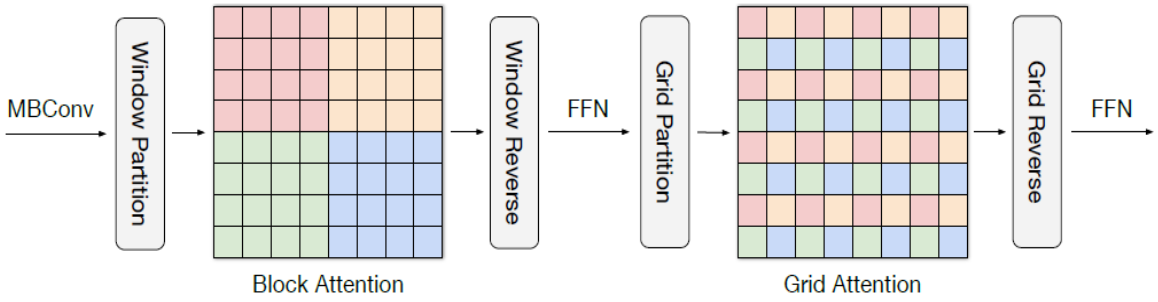


Figure 2. A Depiction of the Method for Calculating Self-attention using a Multi-axis Approach, with a Window/grid size of 4x4 (Tu et al., 2022).

The values indicated by the letters "B" and "C" represent the quantity of blocks and channels, accordingly, at each model's step. The model's attention layers are all set up to use 32 attention heads. The MBConv module's Squeeze-and-Excitation (SE) component is always run at a 4x expansion rate and a 0.25 shrinking rate. Moreover, the model's original structure comprises two convolutional

layers. We adapted the proposed solution relying on the MaxViT model in order to maximize the model's efficacy while reducing the number of parameters for the three-class COVID-QU-Ex dataset. In this scaling, the block count was the only thing that was lowered in comparison to the smaller model, while the block and channel numbers (B and C) were reduced in comparison to the larger model. This basically meant that the repetition counts of layers S1–S4 (which stand for block and channel quantities) were significantly reduced, which allowed for a drop in the number of parameters and improved global and local feature extraction from X-ray images. Compared to the larger model, we decreased the block count in the Convolutional stem from 2 to 1 and the channel number from 192 to 64. On the other hand, we kept the same number of channels as the smaller model and merely decreased the block size from two to one. These substantial alterations in block and channel numbers led to the proposal of a smaller model than the tiny model, resulting in improved accuracy and a higher Frames Per Second (FPS) rate, thereby enhancing inference speed.

2.4. Deep Learning Optimizers

Deep neural network training relies primarily on two main learning approaches: supervised and unsupervised learning. Supervised learning is applied when the network has access to both input data and corresponding output labels, making it suitable for tasks like classification and regression. In contrast, unsupervised learning is employed in scenarios where only input data is available, typically for clustering and uncovering underlying data structures. The fundamental algorithm used for supervised learning in deep neural networks is back-propagation, which comprises forward and backward propagation phases. However, standard gradient descent in backpropagation may encounter challenges such as local optima and a high number of iterations. To address these issues, mini-batch gradient descent, including stochastic gradient descent (SGD), is often utilized, providing memory efficiency and faster convergence. SGD processes individual training samples, making it suitable for large datasets and enabling frequent parameter updates for improved optimization.

Due to their critical role in model training, optimizers are essential to deep learning. Finding the best model configurations is made easier by them as they assist in overcoming difficulties brought on by high-dimensional parameter spaces. Faster convergence is made possible, especially with big and deep networks, by advanced optimizers like Adam and RMSprop with adjustable learning rates and momentum approaches. In order to ensure more stable and effective training procedures, optimizers also address gradient-related problems including vanishing and bursting gradients. As a result, choosing an optimizer is crucial in deep learning because it directly affects model performance as well as training effectiveness.

Adam is a very effective and flexible optimization algorithm that is commonly used in place of more conventional stochastic gradient descent techniques. It exhibits high computational effectiveness and low memory requirements by dynamically updating the learning rate for each parameter [8]. Adam uses a momentum-based parameter update method that is similar to gradient descent. Adam is a flexible and popular option for optimizing complex models, especially in deep learning applications, because to its combination of strategies.

SGD is a prevalent optimization technique in deep learning and machine learning. Due to its ability to calculate the loss function gradient for individual training samples, it requires minimal memory usage and is capable of processing large datasets effectively. The main problem with SGD is choosing the right learning rate to avoid oscillations and obtain the global optimum. It is crucial to use strategies like learning rate schedules and momentum to enhance its performance because its noisy updates can both help in avoiding local minima and impede convergence. Finding the ideal hyperparameter values is frequently a crucial component of efficiently implementing SGD in practice.

The Lion optimizer, also known as "EvoLved Sign Momentum," is a substantial improvement in the field of machine learning optimization. By elegantly concentrating on momentum monitoring and sign-based updates, it streamlines the procedure while using less memory and ensuring consistent update magnitudes across all dimensions. Lion is a potential optimizer for the community at large because it performs exceptionally well across a range of machine learning models and applications. For bigger batch sizes, its memory-efficient stochastic gradient-descent method different from adaptive optimizers like Adam offers efficiency and magnitude control, which is depicted in the Algorithm 1.

Algorithm 1 Lion Optimizer

given β_1, β_2, λ, η, f
initialize θ_0, $m_0 \leftarrow 0$
while θ_t not converged **do**
  $g_t \leftarrow \nabla_\theta f(\theta_{t-1})$
  update model parameters
  $c_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
  $\theta_t \leftarrow \theta_{t-1} - \eta_t(\text{sign}(c_t) + \lambda\theta_{t-1})$
  update EMA of g_t
  $m_t \leftarrow \beta_2 m_{t-1} + (1-\beta_2) g_t$
end while
return θ_t

2.5. Dataset

The significance of the data set in demonstrating the desired success of a deep learning architecture is undoubtedly beyond question. While classic machine learning approaches focus on manual feature extraction and small data, the main difference between the two approaches is that deep learning architectures require large-scale data sets and automatic feature extraction. The increasing number of medical images in hospitals and research centers in recent years has made deep learning the most effective method in medical image processing, as deep learning architectures have a naturally data-hungry characteristic for features. In this study, the use of a publicly available COVID-19 data set is aimed at making the experimental results and discussions more objective. Although there are several public data sets in the literature, there is no common evaluation criterion. Considering this situation, COVID-QU-Ex (Tahir et al., 2021) includes both a common evaluation criterion and X-ray images that are unique to the training, validation, and test folders for lung infections and COVID-19. Therefore, it is possible to assess the actual effectiveness of the model through its training and testing outcomes on test data and compared with other current methods. Some sample images taken randomly from the dataset is illustrated in Figure 3.

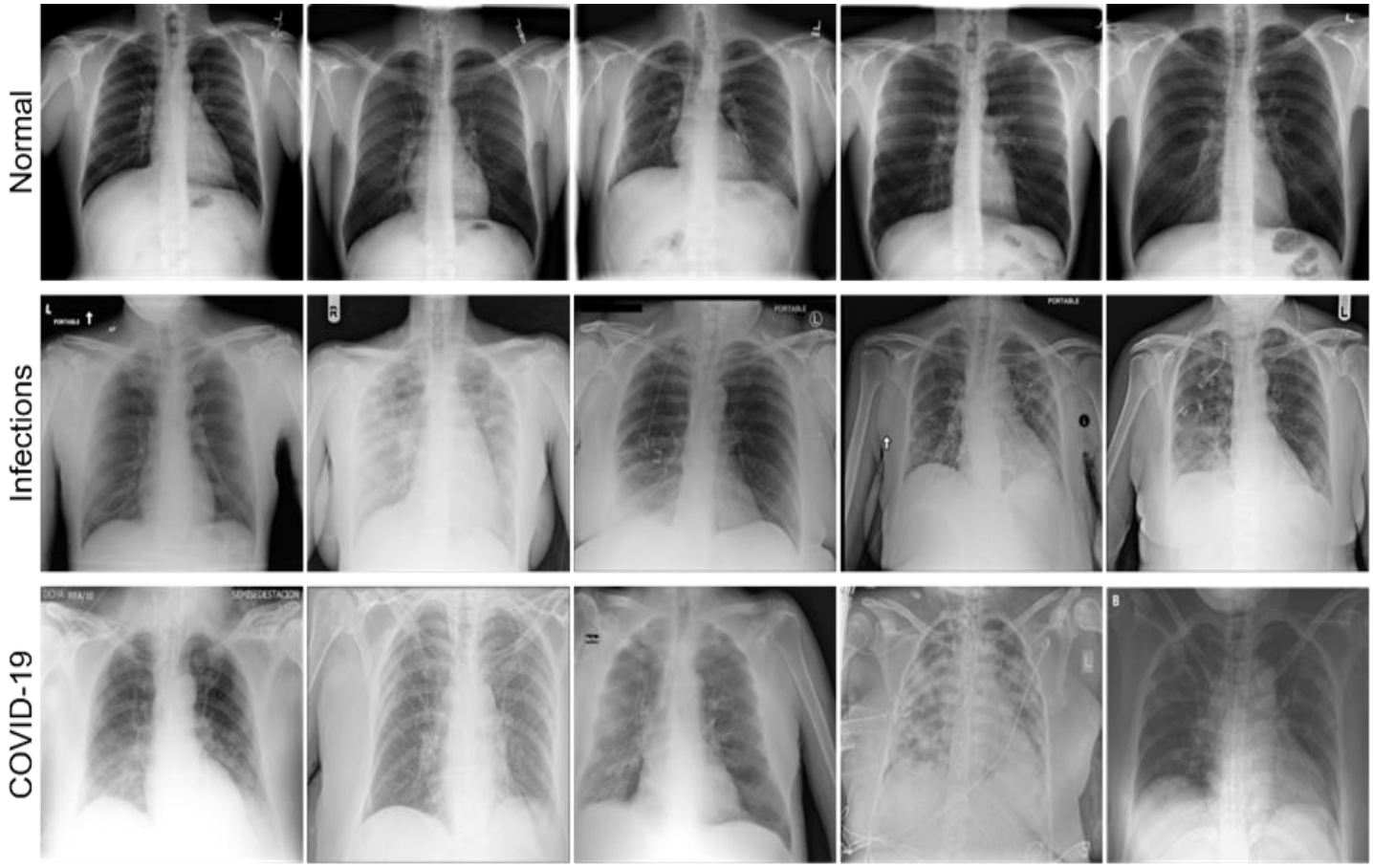


Figure 3. Some Sample Images from the Different Classes in the COVID-QU-Ex Dataset.

3. Results and Discussion

The findings and analyses of several CNN and Transformer-based models for the identification of lung infections and disorders linked to COVID-19 are presented in this section. Along with the outcomes of the proposed approach, experimental findings pertaining to well-known optimizers are also presented.

3.1. Experimental Design

For optimal outcomes, deep learning algorithms necessitate training on a computer equipped with GPU capabilities. GPUs have the ability to perform parallel processing and tensor operations faster than CPUs due to the CUDA cores within them. This research involved conducting experiments on a computer with specific configurations: utilizing a Linux-based Ubuntu 22.04 OS, an NVIDIA RTX 3090 GPU, and 32 GB of DDR5 RAM. Python served as the primary programming language, coupled with PyTorch as the chosen deep learning framework for executing the experiments.

3.2. Data Pre-Processing and Transfer Learning

This study uses a publicly available COVID-19 dataset to improve the objectivity of our research findings. COVID-QU-Ex [9] is unique in that it offers a standardized evaluation criterion and includes X-ray images for training, validation, and testing, covering lung infections and COVID-19 cases. This dataset enables a fair assessment of our model's performance and facilitates comparisons with other methods. Table 1 provides an overview of the dataset's image distribution across training, validation, and test sets.

Table 1. Details of the COVID-QU-Ex data set

Class Type	Training	Validation	Test	Total
COVID-19	7658	1903	2395	11956
Non-COVID infections (viral or bacterial pneumonia)	7208	1802	2253	11263
Normal (healthy)	6849	1712	2140	10701
Total	21715	5417	6788	33920

The COVID-QU-Ex dataset's primary advantage lies in its predefined training and evaluation image partitions, facilitating straightforward comparisons and precise assessments of model performance vis-à-vis other research studies. Comprising an extensive amalgamation of images from diverse datasets, it encompasses 11,956 COVID-19 cases, 11,263 instances of COVID infections (viral or bacterial pneumonia), and 10,701 normal (healthy) cases. Notably, as illustrated in Table 2, the class distributions within the training, validation, and test datasets are meticulously balanced, eliminating any data imbalance issues. This equilibrium fosters optimal learning of each class by the model and minimizes the risk of bias towards any category.

To improve the robustness and flexibility of our models in this work, we utilized methods for data augmentation. Data augmentation includes altering the original images in various ways, producing fresh synthetic instances, and lowering the danger of overfitting. During model training, we notably used cropping, flipping, rotation, copy-paste, shearing, and scaling. This significantly increased the dataset and improved the model's capacity to generalize to examples that hadn't been seen before. The goal of this augmentation was to improve the precision and dependability of our models for detecting lung infections, which would ultimately lead to more effective screening and diagnosis. By exploiting the model's acquired information and representations from millions of various images, the pre-trained weights from the ImageNet dataset were also utilized in transfer learning. The performance of the pre-trained model dataset was further enhanced through fine-tuning, which reduced training time and processing requirements.

3.2. Performance Metrics

Object classification algorithms or architectures utilize various criteria to assess the efficacy of a model from diverse angles. Commonly employed metrics encompass accuracy, precision, recall, and the F1-score. These metrics are prevalent in scholarly discourse for comparative analysis and comprehensive evaluation of deep learning models. Computation of these metrics necessitates the utilization of the true positive, true negative, false positive, and false negative data constituting the confusion matrix. True positives denote accurately predicted instances of the positive class, whereas true negatives signify appropriately anticipated instances of the negative class. False positives and false negatives denote instances where the prediction is erroneously labeled as positive or negative, respectively. While accuracy gauges the proportion of correct predictions relative to all predictions, precision assesses the ratio of true positive predictions to all positive predictions. Recall, also referred to as sensitivity, displays the proportion of predictions for a particular class that are actually true positive predictions. The weighted average of recall and precision, on the other hand, makes up the F1-score. These metrics can be formally stated in Table 2.

Table 2. Performance Metrics

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Number of total predictions}}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

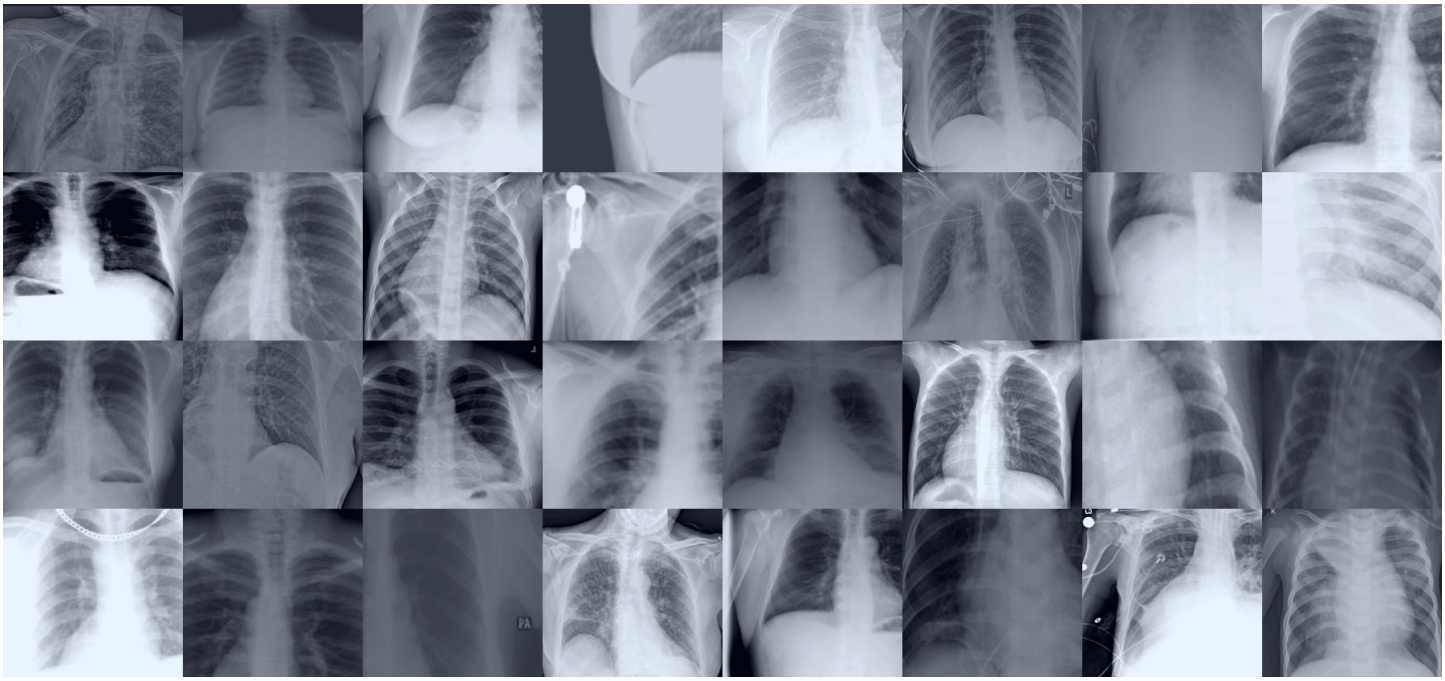
$$Recall = Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

3.3. Training Procedure

To optimize the efficiency and precision of training deep learning models, several strategies and components are employed. These are transfer learning and methods for enhancing data. Furthermore, factors like input image dimensions, batch size, epochs, optimizer

selection, learning rate, and repetition of data augmentation are pivotal. To enhance diagnostic accuracy of the models and make their comparability objective, we employed the most fundamental data augmentation techniques in this work, including scale, smoothing, mixup, color jitter, and flip, in the same way for each model. The data augmentation technique was carried out only during the training, that is, offline data augmentation was not used. For transfer learning, weights of the ImageNet dataset were used in all models. Transfer learning is the most important machine learning technique used to offer faster convergence of training and higher accuracy. Other parameters used in the study are as follows. The training input value of each model is selected as 3x224x224, but a fixed 299x299 size is selected for Xception and 240x240 is selected for FlexiViT architectures because of the nature of these architectures. Briefly, 224x224 resolution was used for almost all the models. An example batch of X-ray images entered in the training is depicted in Figure 4. The optimization algorithm used was Stochastic Gradient Descent (SGD) with an initial learning rate of 0.01. Each model underwent 300 epochs with a momentum of 0.9. Additionally, a warm-up stage was included, where the weight decay gradually increased from



2.0e-05 to 1.0e-05.

Figure 4. A Batch of X-ray Images Used During the Training Process.

3.4. Experimental Results

Deep learning models have outperformed other methods in object classification and detection so far. Performance comparison typically relies on commonly used datasets in the fields of object classification and detection, like ImageNet and MSCOCO, and the models' performance is ranked accordingly. It is only through experimental results that we can determine which model will offer better performance for more specific areas. While the SOTA models generally offer the best performance, in some cases, popular models can perform better.

Table 3. Experimental Results for CNN Models

Model	Accuracy	Precision	Recall	F1-score
ResNet18	0.9553	0.9549	0.9549	0.9549
ResNet34	0.9580	0.9576	0.9574	0.9575
ResNet50	0.9454	0.9449	0.9454	0.9451
VGG13	0.9571	0.9565	0.9568	0.6566
VGG16	0.9583	0.9578	0.9577	0.9577
VGG19	0.9537	0.9532	0.9531	0.9531
EfficientNetv2-Small	0.9646	0.9645	0.9641	0.9643
EfficientNetv2-Medium	0.9592	0.9587	0.9584	0.9586
MobileNetv3-Small	0.9216	0.9223	0.9215	0.9219
MobileNetv3-Large	0.9135	0.9132	0.9138	0.9135
DenseNet121	0.9574	0.9569	0.9569	0.9569
DenseNet169	0.9582	0.9576	0.9580	0.9578
DenseNet201	0.9602	0.9596	0.9599	0.9597
Inception_v3	0.9601	0.9597	0.9591	0.9594
Xception	0.9630	0.9624	0.9622	0.9623
Vovnet39a	0.9396	0.9392	0.9397	0.9394
Vovnet57a	0.9381	0.9376	0.9381	0.9378

In this study, we utilized the most popular and current CNN architectures to date, including ResNet (He et al., 2016), DenseNet (Huang et al., 2016), Xception (Chollet, 2017), VGG (Simonyan & Zisserman, 2015), EfficientNet (Tan & Le, 2019), MobileNets (Howard et al., 2017), Inception (Szegedy et al., 2014) and VovNet (Lee et al., 2019) structures. Additionally, we used vision transformer architectures, which have become as popular as CNNs in image analysis since 2021 and have produced more successful results than CNNs in many areas. We conducted a comprehensive comparison of nearly all vision transformer architectures in the literature, such as ViT (Dosovitskiy et al., 2020), Swin (Liu et al., 2021), MaxViT (Tu et al., 2022) , DeiT (Touvron et al., 2020), CoaT (Xu et al., 2021) , MobileViT (Mehta & Rastegari, 2021), CrossViT (Chen et al., 2021), PiT (Heo et al., 2021), BeiT (Bao et al., 2021), and FlexiViT (Beyer et al., 2022). Table 3 provides experimental results for CNN models. Experimental results were obtained by running the codes of each model in github repositories such as Facebook, Google, Microsoft, Apple and Huggingface.

Table 3 presents performance metrics of popular CNN models for diagnosing lung infections and COVID-19 diseases on the COVID-QU-Ex dataset. Upon examination of Table 3, it can be seen that almost all models demonstrate high performance, with an accuracy of over 91%. Upon closer inspection of the CNN models, it is observed that the EfficientNetV2-small model performs significantly better than the other models, while the models in the MobileNet architecture, which are shallower compared to other architectures, exhibit lower performance. The VGG architecture, which is older than other models, still appears to be powerful and provides similar results to newer architectures. On the other hand, it can be seen that the ResNet architecture still offers high performance. When the models for each architecture are ranked in terms of their performance, it can be said that shallower models show better performance. Based on accuracy and other metrics, the EfficientNetV2-small architecture is found to provide the most successful results with values such as accuracy: 0,9646, precision: 0,9645, recall: 0.9641, and F1-score: 0.9643, while the MobileNetv3-Large model exhibits the lowest performance with values such as accuracy: 0.9135, precision: 0.9132, recall: 0.9138, and F1-score: 0.9135 across all metrics. Following the EfficientNet_v2-small model, Xception, Inception_v3, and DenseNet-201 models showed an accuracy of over 96%, providing results that were closest to and most successful compared to the EfficientNet_v2-small model. The outcomes of the vision transformer models are shown in Table 4, which have recently become more popular deep learning architectures compared to other architectures.

Table 4. Experimental Results for Vision Transformer Models

Model	Accuracy	Precision	Recall	F1-score
ViT-L_32	0,9504	0,9498	0,9496	0,9497
ViT-L_16	0,9596	0,9593	0,9591	0,9592
ViT-B_32	0,9645	0,9640	0,9637	0,9638
ViT-B_16	0,9648	0,9642	0,9638	0,9640
Swin-B	0,9634	0,9632	0,9624	0,9628
Swin-T	0,9671	0,9666	0,9667	0,9666
MaxVit-L	0,9629	0,9626	0,9622	0,9624
MaxVit-B	0,9624	0,9621	0,9616	0,9618
MaxVit-S	0,9642	0,9635	0,9636	0,9635
MaxVit-T	0,9636	0,9629	0,9629	0,9629
DeiT3-H	0,9658	0,9652	0,9653	0,9652
DeiT3-L	0,9577	0,9579	0,9564	0,9571
DeiT3-B	0,9589	0,9583	0,9579	0,9581
DeiT3-S	0,9619	0,9615	0,9613	0,9614
CoaT-T	0,9487	0,9488	0,9484	0,9486
DeiT-B	0,9653	0,9647	0,9647	0,9647
DeiT-S	0,9636	0,9630	0,9632	0,9631
DeiT-T	0,9579	0,9572	0,9573	0,9572
MobileVit-S	0,9614	0,9611	0,9607	0,9609
MobileVit-xS	0,9626	0,9615	0,9621	0,9618
CrossVit-S	0,9602	0,9598	0,9596	0,9597
CrossVit-T	0,9595	0,9589	0,9590	0,9589
PiT-S	0,9576	0,9569	0,9572	0,9570
PiT-B	0,9643	0,9641	0,9634	0,9637
PiT-T	0,9499	0,9494	0,9498	0,9496
BeiT-B	0,9582	0,9576	0,9576	0,9576
BeiT-L	0,9664	0,9659	0,9658	0,9658
FlexiViT-B	0,9554	0,9552	0,9548	0,9550
FlexiViT-L	0,9646	0,9641	0,9639	0,9640
MaxViT-Proposed	**0,9692**	**0,9688**	**0,9687**	**0,9687**

Table 4 presents performance metrics for the diagnosis of lung infections and COVID-19 diseases on the COVID-QU-Ex dataset using deep learning architectures, specifically the vision transformer architecture which is the most popular architecture lately. We provide the widest experimental results in the literature for almost all vision transformer models ranging from the oldest and first vision transformer architecture, ViT, to the newest and most powerful architectures such as MaxViT, DeiT-3, and FlexiViT for the diagnosis of lung infections and COVID-19. While the success of these architectures can be found only on the ImagNet dataset, comparing their performance on medical images such as X-rays is quite difficult. This study particularly presents the performance of vision transformer models on a very wide range of medical X-ray images. Table 4 shows that almost all models have an accuracy of greater than 94%, showing a much higher performance compared to CNN models. When examining the vision transformers, the proposed MaxViT model is seen to be more successful than other models, while the Swin-T model is the second most successful model with a small difference. Models belonging to ViT architecture with 16 patch numbers showed higher performance compared to models with 32 patch numbers. The DeiT model is seen to be more successful than the DeiT-3 model, while the most current transformer model, FlexiViT, is less successful than some other transformer models. On the other hand, it is observed that architectures such as ViT, BeiT, and Swin are still among one of the most successful architectures. In short, it can be said that the most current transformer models are not the most successful models for medical datasets, and this situation is generally valid only for the ImageNet dataset. The proposed MaxViT model has the best performance metrics, scoring 0.9692 accuracy, 0.9688 precision, 0.9687 recall, and 0.9687 F1-score, compared to 0.9671 accuracy, 0.9666 precision, 0.9667 recall, and 0.9666 F1-score for the Swin-T model. After these models, it is observed that models such as BeiT-L, DeiT-B, DeiT3-H, MaxViT-S, and ViT-B_16 provide similar and successful results, while on the other hand, the lowest performance belongs to the CoaT-T model. However, it can be said that all transformer models show much higher performance compared to CNNs.

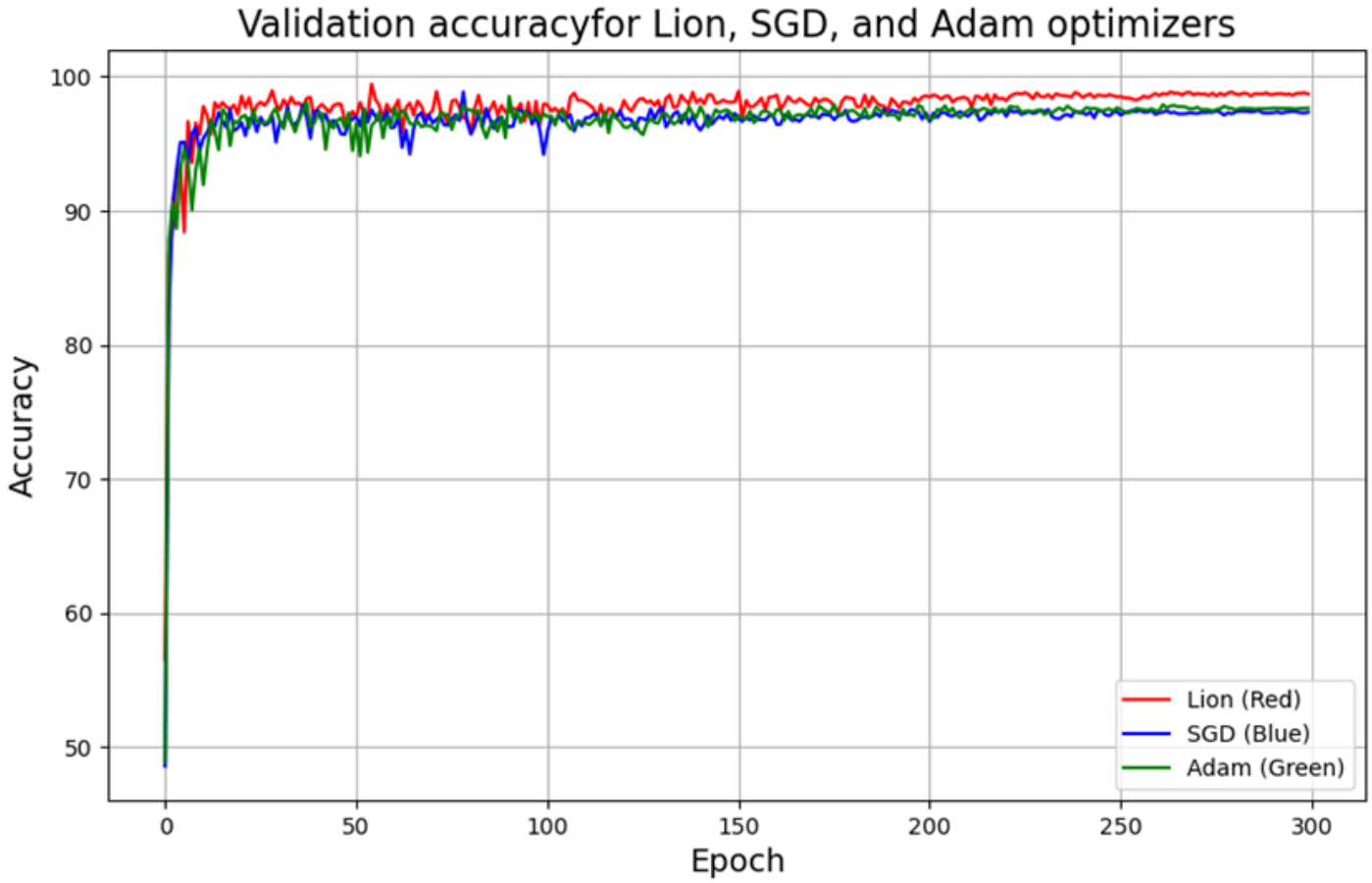


Figure 5. Performance of the Proposed Model with Various Optimizers.

3.5. Results for Optimizers

The experimental findings are presented in this section concerning the performance of various optimizers, SGD, Adam, and Lion optimizer, when applied to our Proposed Model. These optimizers play a critical role in training machine learning models, influencing their convergence speed and overall effectiveness. Through a series of experiments, we aim to assess how each of these optimizers impacts the performance of our model, shedding light on their respective strengths and weaknesses in optimizing the Proposed Model. Table 5 provides results and valuable insights into the choice of optimizer for achieving the best results in our detection of lung related diseased.

Table 5. Proposed Model with Various Optimizers

Optimizer	Accuracy	F1-score	Convergence epoch
Adam	0.9692	0,9687	81
SGD	0.9698	0.9693	79
Lion	**0.9714**	**0.9709**	**55**

As seen in the Figure 5 and Table 5, we introduce our novel proposed model, thoughtfully enhanced with the cutting-edge Lion optimizer, tailored to address the critical task of diagnosing lung infections and COVID-19. Our study encompasses a meticulous examination of this model's performance, comparing it against established optimizers like Adam and SGD, within the context of medical diagnostics. Remarkably, the Lion optimizer emerges as the star performer, demonstrating remarkable accuracy at 0.9714 and an F1-score of 0.9709. Equally noteworthy is its expeditious convergence, requiring just 55 epochs to reach optimal results. These findings underscore the Lion optimizer's potential to revolutionize the field of medical diagnostics, particularly in the crucial arena of lung infection and COVID-19 diagnosis, elevating it to a prominent position within our research contribution. Furthermore, it is worth noting that the Lion optimizer achieves a stable level of accuracy around the 200th epoch, whereas in the case of SGD and Adam, stability in accuracy is achieved after the 250th epoch. As a result, the proposed model demonstrates a more effective performance when coupled with the Lion optimizer.

3.6. Proposed Model Over Pure Maxvit Variants

We provide a thorough comparison of the proposed MaxViT model with the original models found in the MaxViT paper. The MaxViT-Xlarge, MaxViT-Large, MaxViT-Small, MaxViT-Base, MaxViT-Small, and MaxViT-Tiny models are presented in the MaxViT paper, which are obtained by scaling the S0-Convolutional Stem and MaxViT blocks based on the MaxViT architecture. Following these rules, we propose a model that is smaller than the MaxViT-Tiny model. Table 6 provides a detailed comparison of the proposed model and other MaxViT models.

Table 6. The Effectiveness of the Proposed Model versus Pure MaxViT Models (GPU: 1xRTX 3090, Batch-size: 16).

Model Name	Img-size	Accuracy	Model Parameters (M)	GPU Usage (GB)	Training time
MaxViT-Large	224x224	0.9629	210.8	21.24 (%86)	545s/epoch
MaxViT-Base	224x224	0.9624	118.7	15.62 (%64)	382s/epoch
MaxViT-Small	224x224	0.9642	68.2	9.690 (%39)	213s/epoch
MaxViT-Tiny	224x224	0.9636	30.4	6.997 (%28)	144s/epoch
MaxViT-Proposed	224x224	**0.9714**	**16.8**	**5.857 (%24)**	**98s/epoch**

In Table 6, the Img-size column indicates the resolution of the training and test pictures, the accuracy and F1-score columns indicate how well each model performed on the COVID-QU-Ex dataset, and the model parameters column indicates number of parameters per each model in millions. In addition, the batch-size was fixed at 16, and the GPU consumption for each model and the time taken for each model to be trained for one epoch in seconds were provided in the table. Upon examining Table 5, it can be seen that the proposed method outperforms other models in terms of both high F1 and accuracy values, as well as much lower GPU consumption and training time compared to other models. Additionally, when considering complexity in terms of total parameter count, it is seen that the proposed method offers higher performance with up to 80% fewer parameters compared to the closest model, MaxViT-Tiny. Moreover, when considering inference speed, the proposed method processes a 224x224 image at a speed of 5ms or 200FPS, while the closest model, MaxViT-Tiny, processes it at a speed of 8ms or 125 FPS. This indicates that the proposed model is much faster than all other models in terms of inference speed and is 60% faster than the closest model. The class-wise performance of the proposed MaxViT model is shown in confusion chart in Figure 7.

As depicted in the Figure 6, the proposed model's performance evaluation was marked by a robust and rigorous class-wise comparison, illuminating its remarkable accuracy and precision in classifying instances across three vital categories: COVID-19, Infections, and Normal. Impressively, in the case of COVID-19, the model accurately identified 2361 cases while making only 10 and 24 misclassifications as Infections and Normal, showcasing its exceptional sensitivity and specificity. Likewise, within the Infections class, the model's prowess shone through, with 2159 correct predictions and only 13 and 81 instances erroneously categorized as COVID-19 and Normal, demonstrating its ability to discern subtle nuances in medical data. Furthermore, the Normal class exhibited strong performance, with the model correctly assigning 2074 instances and committing just 9 and 57 misclassifications as COVID-19 and Infections, emphasizing its reliability in distinguishing between health states.

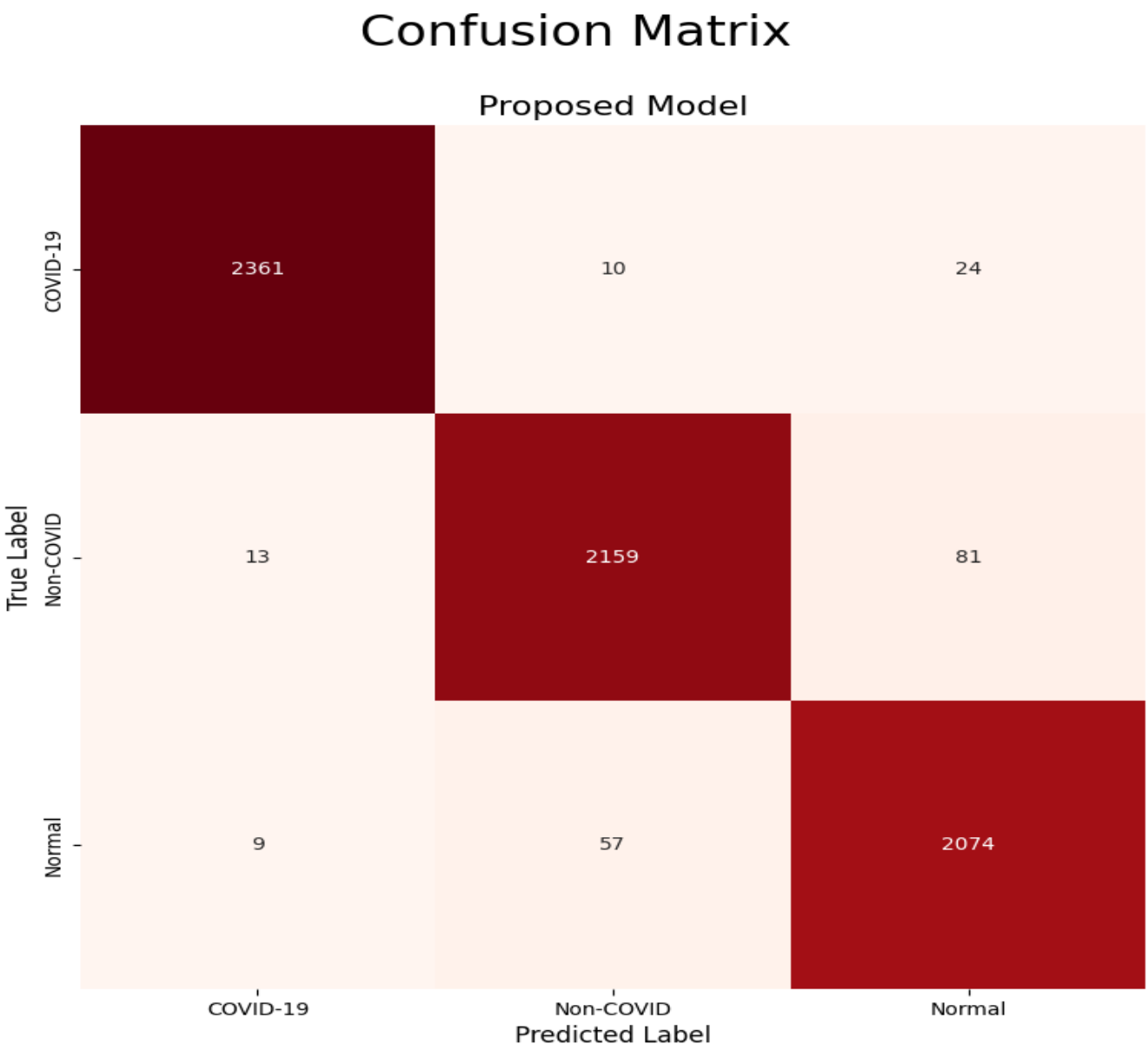


Figure 6. Confusion chart for Proposed Model

3.7. Comparison with SOTA Methods

This section provides a thorough evaluation of the suggested model in comparison to the most recent methodologies described in the literature. Evaluation measures, datasets, and imaging modalities for each model are provided in order to compare how well each model performs compared to the others. Data on the suggested method's comparison with other approaches on the COVID-QU-Ex dataset are presented in Table 7. Since the COVID-QU-Ex dataset is a new and novel largest dataset, there are limited studies on this dataset in the literature. In a comparative analysis of models on the COVID-QU-Ex dataset, it is evident that the proposed "MaxViT" model, our model, stands out significantly with an impressive accuracy of 0.9714. Notably, our model is also distinguished by its lightweight nature compared to several state-of-the-art (SOTA) methods such as VGG16, DenseNet201, Inception_v3, DeiT3-S, CrossVit-T, and PiT-S, which, while achieving commendable accuracies, most of them are not considered lightweight models. Even when compared to recent advances like the models by Ibrokhimov and Kang (0.9660) and Yue et al. (0.9617), our MaxViT model surpasses their performance, showcasing its efficacy in COVID-19 classification tasks while maintaining a lightweight design, which is crucial for practical deployment and resource efficiency. In contrast, other models, such as those presented by Nafisah et al., Kuzinkovas and Clement, Ahmet and Lin, and Constantinou et al., fall behind in terms of accuracy, further emphasizing the superiority of our lightweight MaxViT model.

Table 7. Proposed Model with SOTA Methods on COVID-QU-Ex Dataset

Study	Method	Accuracy
VGG16	CNN-VGG	0,9583
DenseNet201	CNN-DenseNet	0,9602
Inception_v3	CNN-Inception	0,9601
DeiT3-S	Transformer-DeiT3	0,9619
CrossVit-T	Transformer-CrossViT	0,9595
PiT-S	Transformer-PiT	0,9576
Ibrokhimov and Kang, (2022)	VGG19, ResNet50	0.9660
Yue et al. (2022)	LRA-Net	0.9617
Nafisah et al. (2023)	EfficientNet	0.9313
Kuzinkovas and Clement, (2023)	CNN-based	0.9468
Ahmet and Lin, (2023)	CNN-based	0.8800
Constantinou et al. (2023)	CNN-based	0.9600
Our Model	MaxViT	**0.9714**

4. Conclusion and Future Directions

Considering the profound global impact of lung infections, notably exemplified by the unprecedented challenges posed by the COVID-19 pandemic, there exists an urgent imperative for the development of timely and precise diagnostic tools. This comprehensive study has responded to this pressing need by exploring the potential of deep learning algorithms within CAD systems for respiratory diseases.

Introducing an innovative approach, we present the MaxViT model, strategically engineered with reduced parameters and real-time diagnostic capabilities. Our findings demonstrate an impressive accuracy rate of 97.14% on the COVID-QU-Ex dataset, marking a significant advancement in diagnostic precision. Notably, the lightweight design of the MaxViT model positions it as a transformative asset in the field. Moreover, the integration of the Lion optimizer has further amplified its diagnostic efficacy, particularly in identifying lung infections with exceptional accuracy. In addition, this study offers a comprehensive comparison between CNNs and ViT-based architectures, shedding light on their respective strengths and weaknesses. These insights lay the groundwork for the development of highly efficient diagnostic tools for the early detection and containment of lung infections, including the formidable COVID-19, thereby mitigating their global impact and fostering proactive healthcare strategies.

Future research endeavors aim to further evaluate the performance of the MaxViT model across larger datasets and diverse populations. Additionally, exploring the integration of various data sources, such as imaging and clinical records, could serve as a pivotal step in assessing the clinical success of the MaxViT model. Enhancing the interpretability of deep learning models remains a primary focus, with the development of methodologies to elucidate decision-making processes, fostering trust among healthcare professionals and facilitating seamless integration into clinical practice. Furthermore, the tailored development of the Lion optimizer and exploration of novel optimization techniques hold promise in advancing diagnostic performance and efficiency. Finally, examining how the MaxViT model adapts to evolving healthcare demands and technological advancements is paramount for future investigation.

References

Abdul Gafoor, S., Sampathila, N., Madhushankara, M., & Swathi, K. S. (2022). Deep learning model for detection of COVID-19 utilizing the chest X-ray images. Cogent Engineering, 9(1). https://doi.org/10.1080/23311916.2022.2079221

Ahmad, M., Usama, ·, Bajwa, I., Mehmood, Y., Muhammad, ·, & Anwar, W. (n.d.). Lightweight ResGRU: a deep learning-based prediction of SARS-CoV-2 (COVID-19) and its severity classification using multimodal chest radiography images. Neural Computing and Applications. https://doi.org/10.1007/s00521-023-08200-0

Ahmed, U., & Lin, J. C. W. (2023). Robust adversarial uncertainty quantification for deep learning fine-tuning. Journal of Supercomputing, 79(10), 11355–11386. https://doi.org/10.1007/s11227-023-05087-5

Alshmrani, G. M. M., Ni, Q., Jiang, R., Pervaiz, H., & Elshennawy, N. M. (2023). A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images. Alexandria Engineering Journal, 64, 923–935. https://doi.org/10.1016/J.AEJ.2022.10.053

Aslan, E. (2024). LSTM-ESA HİBRİT MODELİ İLE MR GÖRÜNTÜLERİNDEN BEYİN TÜMÖRÜNÜN SINIFLANDIRILMASI. Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi, 11(22), 63–81. https://doi.org/10.54365/adyumbd.1391157

Aslan, E., & Özüpak, Y. (2024). Classification of Blood Cells with Convolutional Neural Network Model. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, 13(1), 314–326. https://doi.org/10.17798/bitlisfen.1401294

Aslani, S., & Jacob, J. (2023). Utilisation of deep learning for COVID-19 diagnosis. Clinical Radiology, 78(2), 150–157. https://doi.org/10.1016/J.CRAD.2022.11.006

Ayan, E., Karabulut, B., & Ünver, H. M. (2022). Diagnosis of Pediatric Pneumonia with Ensemble of Deep Convolutional Neural Networks in Chest X-Ray Images. Arabian Journal for Science and Engineering, 47(2), 2123–2139. https://doi.org/10.1007/S13369-021-06127-Z/FIGURES/12

Bao, H., Dong, L., Piao, S., & Wei, F. (2021). BEiT: BERT Pre-Training of Image Transformers. Mim, 1–18.

Beyer, L., Izmailov, P., Kolesnikov, A., Caron, M., Kornblith, S., Zhai, X., Minderer, M., Tschannen, M., Alabdulmohsin, I., & Pavetic, F. (2022). FlexiViT: One Model for All Patch Sizes.

Bhattacharyya, A., Bhaik, D., Kumar, S., Thakur, P., Sharma, R., & Pachori, R. B. (2022). A deep learning based approach for automatic detection of COVID-19 cases using chest X-ray images. Biomedical Signal Processing and Control, 71. https://doi.org/10.1016/j.bspc.2021.103182

Chen, C. F., Fan, Q., & Panda, R. (2021). CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. Proceedings of the IEEE International Conference on Computer Vision, 347–356. https://doi.org/10.1109/ICCV48922.2021.00041

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W. C., Wang, C. Bin, & Bernardini, S. (2020). The COVID-19 pandemic. Critical Reviews in Clinical Laboratory Sciences, 57(6), 365–388. https://doi.org/10.1080/10408363.2020.1783198

Cleverley, J., Piper, J., & Jones, M. M. (2020). The role of chest radiography in confirming covid-19 pneumonia. In The BMJ (Vol. 370). BMJ Publishing Group. https://doi.org/10.1136/bmj.m2426

Constantinou, M., Exarchos, T., Vrahatis, A. G., & Vlamos, P. (2023). COVID-19 Classification on Chest X-ray Images Using Deep Learning Methods. International Journal of Environmental Research and Public Health, 20(3). https://doi.org/10.3390/ijerph20032035

Cookson, W. O. C. M., Cox, M. J., & Moffatt, M. F. (2018). New opportunities for managing acute and chronic lung infections. Nature Reviews Microbiology, 16(2), 111–120. https://doi.org/10.1038/nrmicro.2017.122

Deb, S. D., Jha, R. K., Jha, K., & Tripathi, P. S. (2022). A multi model ensemble based deep convolution neural network structure for detection of COVID19. Biomedical Signal Processing and Control, 71. https://doi.org/10.1016/j.bspc.2021.103126

Devasia, J., Goswami, H., Lakshminarayanan, S., Rajaram, M., & Adithan, S. (123 C.E.). Deep learning classification of active tuberculosis lung zones wise manifestations using chest X-rays: a multi label approach. Scientific Reports |, 13, 887. https://doi.org/10.1038/s41598-023-28079-0

Dhiman, G., Chang, V., Kant Singh, K., & Shankar, A. (2022). ADOPT: automatic deep learning and optimization-based approach for detection of novel coronavirus COVID-19 disease using X-ray images. Journal of Biomolecular Structure and Dynamics, 40(13), 5836–5847. https://doi.org/10.1080/07391102.2021.1875049

Dönmez, E. (2024). Hybrid convolutional neural network and multilayer perceptron vision transformer model for wheat species classification task: E-ResMLP+. European Food Research and Technology, 250(5), 1379–1388. https://doi.org/10.1007/S00217-024-04469-0/TABLES/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. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem, 770–778. https://doi.org/10.1109/CVPR.2016.90

Heo, B., Yun, S., Han, D., Chun, S., Choe, J., & Oh, S. J. (2021). Rethinking Spatial Dimensions of Vision Transformers.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). Densely Connected Convolutional Networks. http://arxiv.org/abs/1608.06993

Ibrokhimov, B., & Kang, J.-Y. (2022). Deep Learning Model for COVID-19-Infected Pneumonia Diagnosis Using Chest Radiography Images. BioMedInformatics, 2(4), 654–670. https://doi.org/10.3390/biomedinformatics2040043

Işık, G., & Paçal, İ. (2024). Few-shot classification of ultrasound breast cancer images using meta-learning algorithms. Neural Computing and Applications, 1–13. https://doi.org/10.1007/S00521-024-09767-Y/TABLES/7

Karaman, A., Karaboga, D., Pacal, I., Akay, B., Basturk, A., Nalbantoglu, U., Coskun, S., & Sahin, O. (2022). Hyper-parameter optimization of deep learning architectures using artificial bee colony (ABC) algorithm for high performance real-time automatic colorectal cancer (CRC) polyp detection. Applied Intelligence. https://doi.org/10.1007/s10489-022-04299-1

Kılıçarslan, S., Diker, A., Közkurt, C., Dönmez, E., Demir, F. B., & Elen, A. (2024). Identification of multiclass tympanic membranes by using deep feature transfer learning and hyperparameter optimization. Measurement, 229, 114488. https://doi.org/10.1016/J.MEASUREMENT.2024.114488

Kuzinkovas, D., & Clement, S. (2023). The Detection of COVID-19 in Chest X-rays Using Ensemble CNN Techniques. Information, 14(7), 370. https://doi.org/10.3390/info14070370

Lee, Y., Hwang, J. W., Lee, S., Bae, Y., & Park, J. (2019). An energy and GPU-computation efficient backbone network for real-time object detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2019-June, 752–760. https://doi.org/10.1109/CVPRW.2019.00103

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8693 LNCS(PART 5), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.

M., E., L., V.-G., I., W. C. K., J., W., & A., Z. (2010). The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 88(2), 303–338.
Mehta, S., & Rastegari, M. (2021). MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. 3.

Nafisah, S. I., Muhammad, G., Hossain, M. S., & AlQahtani, S. A. (2023). A Comparative Evaluation between Convolutional Neural Networks and Vision Transformers for COVID-19 Detection. Mathematics, 11(6). https://doi.org/10.3390/math11061489

Nayak, S. R., Nayak, D. R., Sinha, U., Arora, V., & Pachori, R. B. (2023). An Efficient Deep Learning Method for Detection of COVID-19 Infection Using Chest X-ray Images. Diagnostics, 13(1). https://doi.org/10.3390/diagnostics13010131

PACAL, İ. (2022). Deep Learning Approaches for Classification of Breast Cancer in Ultrasound (US) Images. Journal of the Institute of Science and Technology, 1917–1927. https://doi.org/10.21597/jist.1183679

Pacal, I. (2024a). A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images. International Journal of Machine Learning and Cybernetics. https://doi.org/10.1007/s13042-024-02110-w

Pacal, I. (2024b). Enhancing crop productivity and sustainability through disease identification in maize leaves: Exploiting a large dataset with an advanced vision transformer model. Expert Systems with Applications, 238. https://doi.org/10.1016/j.eswa.2023.122099

Pacal, I., & Karaboga, D. (2021). A robust real-time deep learning based automatic polyp detection system. Computers in Biology and Medicine, 134. https://doi.org/10.1016/J.COMPBIOMED.2021.104519

Pacal, I., Karaboga, D., Basturk, A., Akay, B., & Nalbantoglu, U. (2020). A comprehensive review of deep learning in colon cancer. Computers in Biology and Medicine, 126. https://doi.org/10.1016/J.COMPBIOMED.2020.104003

Platto, S., Xue, T., & Carafoli, E. (2020). COVID19: an announced pandemic. Cell Death and Disease, 11(9). https://doi.org/10.1038/s41419-020-02995-9

Podder, P., Das, S. R., Mondal, M. R. H., Bharati, S., Maliha, A., Hasan, M. J., & Piltan, F. (2023). LDDNet: A Deep Learning Framework for the Diagnosis of Infectious Lung Diseases. Sensors, 23(1). https://doi.org/10.3390/s23010480

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sedik, A., Hammad, M., Abd El-Samie, F. E., Gupta, B. B., & Abd El-Latif, A. A. (2022). Efficient deep learning approach for augmented detection of Coronavirus disease. Neural Computing and Applications, 34(14), 11423–11440. https://doi.org/10.1007/s00521-020-05410-8

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–14.

Soomro, T. A., Zheng, L., Afifi, A. J., Ali, A., Yin, M., & Gao, J. (2022). Artificial intelligence (AI) for medical imaging to combat coronavirus disease (COVID-19): a detailed review with direction for future research. Artificial Intelligence Review, 55(2), 1409–1439. https://doi.org/10.1007/s10462-021-09985-z

Subramanian, N., Elharrouss, O., Al-Maadeed, S., & Chowdhury, M. (2022). A review of deep learning-based detection methods for COVID-19. In Computers in Biology and Medicine (Vol. 143). Elsevier Ltd. https://doi.org/10.1016/j.compbiomed.2022.105233

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going Deeper with Convolutions.

Tahir, A. M., Chowdhury, M. E. H., Khandakar, A., Rahman, T., Qiblawey, Y., Khurshid, U., Kiranyaz, S., Ibtehaz, N., Rahman, M. S., Al-Maadeed, S., Mahmud, S., Ezeddin, M., Hameed, K., & Hamid, T. (2021). COVID-19 infection localization and severity grading from chest X-ray images. Computers in Biology and Medicine, 139. https://doi.org/10.1016/j.compbiomed.2021.105002

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. http://arxiv.org/abs/1905.11946

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). Training data-efficient image transformers & distillation through attention. 1–22.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., & Li, Y. (2022). MaxViT: Multi-axis Vision Transformer. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13684 LNCS, 459–479. https://doi.org/10.1007/978-3-031-20053-3_27

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 2017-Decem(Nips), 5999–6009.

Velavan, T. P., & Meyer, C. G. (2020). The COVID-19 epidemic. Tropical Medicine and International Health, 25(3), 278–280. https://doi.org/10.1111/tmi.13383

Xu, W., Xu, Y., Chang, T., & Tu, Z. (2021). Co-Scale Conv-Attentional Image Transformers. Proceedings of the IEEE International Conference on Computer Vision, 9961–9970. https://doi.org/10.1109/ICCV48922.2021.00983

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F. E., Feng, J., & Yan, S. (2021). Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet.

Yue, G., Lin, J., An, Z., & Yang, Y. (2022). Loop Residual Attention Network for Automatic Segmentation of COVID-19 Chest X-ray Images. IEEE Access, PP, 1. https://doi.org/10.1109/ACCESS.2022.3227798

Yuki, K., Fujiogi, M., & Koutsogiannaki, S. (2020). COVID-19 pathophysiology: A review. Clinical Immunology, 215(April). https://doi.org/10.1016/j.clim.2020.108427