Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

# A Revolutionary Approach to COVID 19 Detection Using a Novel Transformer Based Architecture

**Arshi Husain[1] and Virendra P Vishwakarma[2]**
*[1]IEC College Of Engineering & Technology and [2]Guru Gobind Singh Indraprastha University*

***Abstract:-*** The Covid19 pandemic continues to be a global crisis, resulting in tragic loss of millions of lives. This has spurred researchers to explore deep learn-ing (DL) techniques for Covid 19 diagnosis, aiming to assist medical pro-fessionals in the screening process offering valuable second opinions to clinicians. To that end, we introduce a novel DL architectural design for Covid 19 detection, which combines the strengths of vision transformers (ViTs) for capturing long range dependencies with Efficient Net's (EffNet) fine-grained classification capabilities. Built upon the EffNet-B0 backbone, the ViT-based model with the ViT-B/16 configuration extracts global con-text and long-distance feature information from input images, yielding powerful feature representations. 94.78 % is the accuracy achieved with our proposed model, demonstrating its effectiveness following experimental verification. The efficacy of our model has been empirically substantiated in comparison to state-of-the-art (SOTA) approaches. Despite its initial fo-cus on natural language processing (NLP), our substantial accuracy demon-strates that the ViT model exhibits promising performance and holds great potential for broader applications in computer vision (CV) tasks.

***Keywords****: Covid 19, Vision transformer, Efficient Net.*

## 1. Introduction

Since December 2019, there has been a sudden surge in coronavirus disease 2019 (Covid 19), rapidly spreading worldwide and leading to a widespread outbreak. As of August 6, 2023, the worldwide count registers more than 769 million confirmed cases and over 6.9 million reported deaths. Rapid tests were created to identify Covid 19 in a short timeframe of 30 minutes. Individuals residing in distant regions and grappling with the illness face significant hurdles due to the high cost and limited availability of these tools. The procedures received substantial criticism because of their elevated likelihood of producing false-negative results. [1]. Recent theoretical progress has highlighted that within clinical diagnosis, Real-time reverse transcriptase-polymerase chain reaction (RT-PCR) is widely recognized as the gold standard for diagnosing and confirming cases of Covid 19 in clinical settings [2]. However, in addition to RT-PCR, a comprehensive diagnosis of Covid 19 ideally should incorporate the evaluation of chest Xrays (CXR) or computed tomography (CT) results in patients. Acquiring these images is quick and inexpensive, and radiologists can examine them to look for visible signs of the infection [3]. CXR or CT scans frequently display comparable characteristics in individuals exhibiting Covid 19 symptoms, showing bilateral peripheral consolidation in their lung images [4]. Medical imaging has been a dependable method for non-invasive medical diagnosis from the beginning [5]. In the rapidly progressing field of computer science today, applications employing DL techniques have become more widespread, impacting various facets of our daily lives. In the fight against Covid 19, numerous DL methods have surfaced to aid in diagnosis, making a substantial contribution to the medical response. Lately, ViT have emerged as a noteworthy advancement in the field of computer CV, marking a transfer of the algorithm used in NLP to the CV domain. This shift challenges the supremacy of Convolutional Neural Networks (CNNs) and profoundly impacts CV researchers. In subsequent periods, researchers have applied the ViT structure to explore a diverse array of applications. For instance, the U-transformer framework has shown remarkable effectiveness in complex organ segmentation tasks [6], building on

the advancements introduced by the ViT structure. Additionally, RTMIC [7] is an innovative framework capable of automatically generating captions for CXR images and providing medical diagnoses. Inspired by these developments, our proposed novel DL model, achieves improved accuracy and performance in Covid 19 detection by drawing the strengths of ViT for capturing intricate long-range dependencies and EffNet's exceptional fine-grained classification capabilities.

The subsequent sections of this work are organized as follows: Section 2 conducts a survey of pertinent literature, reviewing previous studies in the field. Section 3 delin-eates the dataset used and the preliminaries integral to our proposed model. The as-sessment of the presented model, along with a comparative analysis against contem-porary SOTA techniques, is explored in Section 4. Finally, Section 5 concludes the work by examining the potential future directions of our novel framework.

## 2. Related works

In the past few decades, the CV community has witnessed a substantial growth and progress, primarily propelled by the rise of DL. Several robust networks [8], [9], [10], [11], [12] have demonstrated exceptional success in extensive image classification tasks over the past few decades [13]. To automate the diagnosis of Covid 19, researchers in [14] utilized the EffNet B4 framework for transfer learning. They introduced a global average pooling 2D layer to address overfitting and decrease the overall parameter count. Over the past decades, numerous robust networks [8], [9], [10], [11], [12] have achieved notable success in large-scale image classification tasks [13]. In a related context, the authors in [15] utilized an ensemble approach by incorporating widely used pretrained CNN models such as InceptionV3 [16], MobileNetV2 [17], ResNet101 [18], NASNet [19], and Xception [20]. These models undergo fine-tuning on the CXR images database, and the final layer representations from each model are combined. Subsequently, these concatenated representations are inputted into a multi-layer perceptron (MLP) for the accurate diagnosis of Covid 19. Apart from the previously discussed architectural improvements, there have been noteworthy efforts [21], [22], [23] dedicated to optimizing overparameterized deep neural networks (DNNs), with a specific focus on achieving a balance between accuracy and efficiency. As an example, MobileNets [17], [24] and EffNets [25] are notable instances that utilized techniques from neural architecture search (NAS) and have showcased impressive performance. While prior research on CNN models has yielded positive outcomes, several unresolved issues persist. A significant limitation lies in the prevalent use of pre-trained models developed for RGB image datasets, which are unsuitable for single-channel images. Additionally, the substantial computational cost associated with these models, necessitating millions of fine-tuning operations for numerous parameters, renders them impractical for devices with limited resources [26]. The remarkable success of transformers in advancing NLP [27], [28] has led to a surge of efforts [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39] to smoothly incorporate architectures inspired by transformers into the domain of visual tasks. This resulted in the creation of a range of ViT models designed for different CV tasks, such as low-level vision tasks, object detection [40], classification [41], image retrieval [42]. In a recent investigation [41], it is asserted by the researchers that the use of CNNs is no longer mandatory in CV. They substantiate their argument by employing transformers directly on image sequences. A new training approach presented in DeiT [30] extends the ViT to enhance data efficiency through direct training on the ImageNet-1K dataset. Authors in [39], introduced an innovative extension to the ViT design by integrating a pyramid structure. This pyramid structure facilitates the creation of multi-scale feature maps, offering advantages in addressing tasks that require dense predictions at the pixel level.

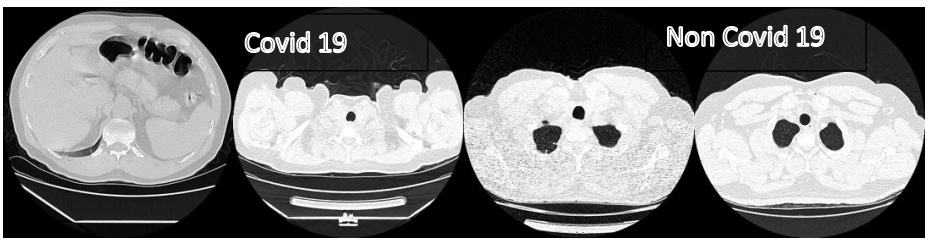


**Fig. 1. Sample images from dataset**

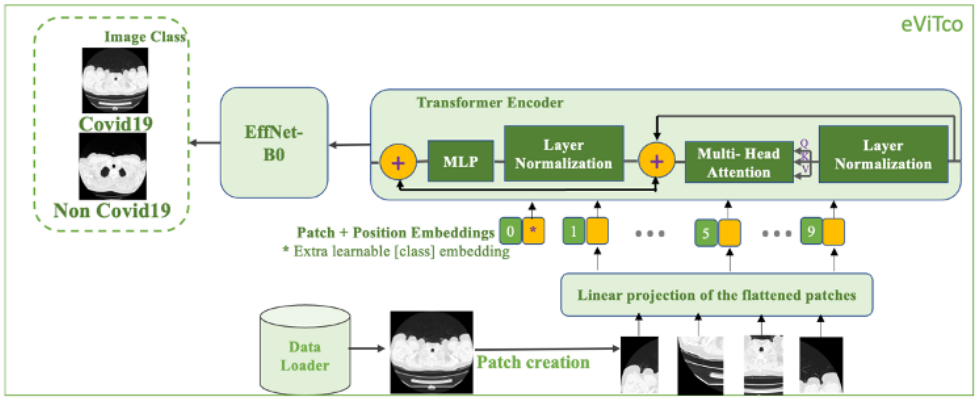


**Fig. 2. Proposed architecture**

**Table 1. EfficientNet B0 architecture**

| Stage | Operator | Resolution | Channel | Layer |
|---|---|---|---|---|
| 1 | Conv3x3 | 224x224 | 32 | 1 |
| 2 | MBConv1,3x3 | 112x112 | 16 | 1 |
| 3 | MBConv6,3x3 | 112x112 | 24 | 2 |
| 4 | MBConv6,5x5 | 56x56 | 40 | 2 |
| 5 | MBConv6,3x3 | 28x28 | 80 | 3 |
| 6 | MBConv6,5x5 | 14x14 | 112 | 3 |
| 7 | MBConv6,5x5 | 14x14 | 192 | 4 |
| 8 | MBConv6,3x3 | 7x7 | 320 | 1 |
| 9 | Conv1x1 with pooling & FC | 7x7 | 1280 | 1 |

**Table 2. Parameter's values set for ViT.**

| Parameter | Value |
|---|---|
| Encoder stride | 16 |
| Number of transformer layers | 12 |
| Number of attention heads for each layer | 12 |
| Activation function | Gelu |
| Hidden size | 768 |
| Dimensions of MLP output | 3072 |
| Dropout ratio for the attention probabilities | 0.1 |
| Initializer range | 0.02 |

## 3. Materials and Methods

### 3.1. Dataset

The dataset utilized in this work is a large public Covid 19 (SARS-CoV-2) lung CT scan dataset, containing total of 8,439 CT scans which consists of 7,495 positive cases indicating Covid 19 infection and 944 negative cases representing normal and non-Covid 19 conditions. The data is provided in the form of 512×512px PNG images

and has been sourced from actual patients in radiology centers within teaching hospitals in Tehran, Iran. A few sample images from the dataset utilized are shown in Figure 1.

### 3.2. Preliminaries

Here, we provide a concise overview of the fundamental elements that constitute the framework of our proposed hybrid model. The ViT and EffNet serve as the core structures of our hybrid model and contribute to its overall performance.

***ViT***

A ViT [41] is a DNN that uses the transformer framework, incorporating an attention mechanism from NLP [27], to efficiently analyze patterns in visual data. Inspired by the success of transformers in NLP, ViTs utilize self-attention mechanisms on image patches or tokens. This enables them to capture both local and large-scale dependencies, improving the modeling of global information in images. Unlike the original transformer with both encoder and decoder, ViT is optimized as an encoder-only architecture for processing visual data, particularly in tasks like image classification, where the input image is divided into fixed-size patches for linear sequence processing. The image's extracted patches serve as a sequential input for the transformer. These patches are flattened and transformed into a latent vector of $D$ dimensions to generate patch embeddings. A trainable embedding is integrated into the sequence of embedded patches within the ViT mode. The last state of the transformer layer associated with this class token, succinctly conveys classification information from the image. The resulting sequence, incorporating patch embeddings and positional embeddings, is then inputted to the encoder. The classification head is represented by an MLP during pre-training and replaced by a linear layer in fine-tuning. The ViT's transformer encoder (TE) consists of interleaved multi-headed self-attention (MSA) and MLP blocks, with skip connections after each block.

***EffNet***

EffNets have attracted attention for their efficacy in image classification, utilizing a novel CNN scaling technique introduced in [25]. This technique uniformly scales the width, resolution, and depth of CNNs to enhance performance. The EffNet family comprises eight models, with EffNet-B0 as the baseline for subsequent models (B1 to B7). Designed by a NAS for a balance between accuracy and efficiency, EffNet-B0 employs compound scaling across width, depth, and resolution. When dealing with higher-resolution images, increasing the network depth aids in capturing larger receptive fields that encompass more pixels in larger images. Table 1 above represents the architecture of EfficientNet-B0. The MBConv block represents an Inverted Residual Block, originally featured in MobileNetV2, enhanced with a Squeeze and Excitation optimization technique. This synchronization optimizes the network for effective handling of various image resolutions, leading to improved performance [25].

### 3.3. Methodology

To facilitate better understanding, we have broken down our hybrid model into three separate components, each precisely explained below and the the proposed architecture has been depicted in Figure 2.

***ViT ImageProcessor***

In preprocessing pipeline, the ViT imageprocessor component from the transformer's library, applies various operations on input images to transform them into a format that can be fed into the transformer architecture. The images are resized to a uniform size of 224 X 224 pixels to ensure consistent dimensions across all the images in the set. Then, the images are converted from PIL format to PyTorch tensors using the transforms. This operation converts the image data into a tensor representation, which can be processed by the neural network. The transformed images are then utilized for further processing.

***ViT feature extractor***

Our work explores the potential of ViTs as feature extractors, starting with the ViT-B/16 variant in the first stage. This variant, with a base architecture featuring a 16x16 patch size and a 224x224 input image size, involves

dividing images into patches, converting them into a flat format, generating low-dimensional linear embeddings, adding positional embeddings, inputting the sequence into a transformer encoder (TE), and obtaining feature vectors as output. These features capture visual information for the final binary classification task of CT images. A concise summary of these steps is presented below.

***EffNet-B0 integration***

During the third stage, the output of the TE is linked to the input of EffNet B0. This connection facilitates the transfer of the learned, comprehensive feature representations from the ViT to the EffNet. EffNet then conducts feature extraction to capture local details and fine-grained features. The model gains a comprehensive image understanding by merging ViT and EffNet strengths, utilizing both global and local information to enhance classification performance. Inspired by [16], EffNet employs a compound scaling technique utilizing a compound coefficient ∅, systematically adjusting layer depth, dimensions, and input resolution.

$$\text{width: } \lambda^{\emptyset} \tag{1}$$

$$\text{depth: } \psi^{\emptyset} \tag{2}$$

$$\text{resolution: } \chi^{\emptyset} \tag{3}$$

$$\text{s.t. } \psi \cdot \lambda^{1.5} \cdot \chi^2 \approx 2.5, \ \psi \geq 1, \lambda \geq 1, \chi \geq 1$$

The constants $\psi, \lambda,$ and $\chi$ determined through grid search, govern the allocation of resources for scaling depth, width, and resolution. Intuitively, the compound coefficient $\emptyset$ dictates the extent of resource allocation for model scaling, allowing for both increases ($\emptyset \geq 1$) and decreases ($\emptyset \leq -1$) In this study, we employed the B0 variant of the EffNet architecture, pre-trained on a distinct ImageNet classification problem with 1000 classes. To tailor the output size for binary classification, we freeze the pre-trained EffNet layer weights. The fully connected layer (FCL) is removed, and a new linear layer with an input size of 1280 (EffNet output) and an output size of 2 (indicating Covid19 or non-Covid19) is added. Classification yields a probability vector for each class. Following EffNet-B0 model outputs, cross-entropy loss is calculated to measure the difference between predicted outputs and actual labels, and gradients are computed for model parameter adjustments. The gradients, obtained through backpropagation, are employed to update the model's parameters, facilitating their learning and adjustment. Backpropagation propagates the gradients from the loss function backward through the model, aiming to minimize the overall loss.

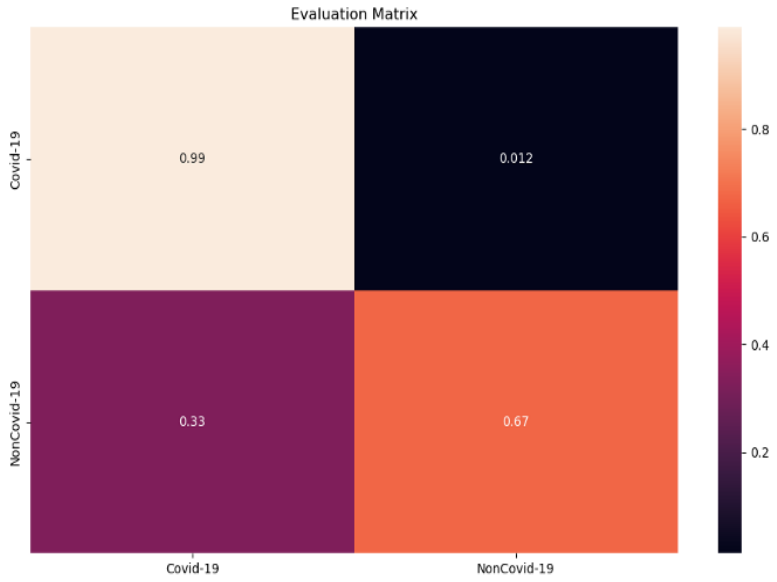


**Fig. 3. Confusion matrix**

**Table 3. Experimental Results.**

| Evaluation Metrics | Result |
|---|---|
| Sensitivity | 0.9541 |
| Precision | 0.9099 |
| F1 score | 0.95 |
| Accuracy | 94.78 % |

**Table 4. Comparative analysis with SOTA using transformers as base.**

| Method | Modality | Overall Accuracy |
|---|---|---|
| ViT base [43] | CT | 76.6 % |
| ViT base [44] | CT | 78.8 % |
| ViT base [45] | CT | 87.26 % |
| ResNet50+ViT-base [46] | CXR | 88.9 % |
| Dense transformer (Swim transformer base) [46] | CXR | 91.8 % |
| Data-efficient image transformers [47] | CXR | 92 % |
| **Proposed model** | **CT** | **94.78 %** |

**Table 5. Comparative analysis with SOTA using EffNet as base.**

| Method | Modality | Overall Accuracy |
|---|---|---|
| Voting based EffNet [48] | CT | 87.6 % |
| EffNet- B4[49] | CT | 89.7 % |
| EffNet- B3 [50] | CXR | 93 % |
| EffNet- B3 [51] | CXR | 93.9 % |
| **Proposed model** | **CT** | **94.78 %** |

## 4. Experimental Results

### 4.1. Implementation Details

The configuration of the device employed to train and assess the model comprises an Apple M1 chip running at 3.2GHz with 8GB of RAM. The execution of our novel model utilizes the PyTorch framework and integrates essential PyTorch libraries, including the PyTorchVision library on Jupyter Notebook. In this study, we chose to integrate the ViT-B16 architecture into our hybrid model. The configuration of its architecture is outlined as follows: The model parameters are set using a pre-trained model on ImageNet-21k, and subsequently, fine-tuning is performed on ImageNet-2012. The Shuffle parameter is configured as true to randomize the sample order during

training, and the number of workers parameter is set to 4. This facilitates the efficient utilization of computational resources by distributing the data loading across multiple subprocesses. The parallel processing capability enables the ViT model to efficiently handle data loading and preprocessing in the background while concurrently performing model computations. For effective processing of training data, a batch size of 16 was utilized. Comprehensive architectural details, as shown in Table 2, encompass all other aspects of the ViT model. The second component of our hybrid model incorporates EffNet-B0. The CrossEntropyLoss function is chosen as the loss function, evaluating loss by comparing predicted class probabilities with true class labels. Additionally, the Adam optimizer is employed for optimizing model parameters. To ensure adequate model convergence and capture underlying data patterns, we conducted training for a total of 10 epochs. Dataset partitioning involved a random approach, assigning 80 percent of total images to the training set and the remaining images to the test set.

### 4.2. Results

The results obtained from the application of our model are shown in Table 3. Figure 3. illustrates the confusion matrix, providing a visual representation of the model's classification performance. Furthermore, we performed a comparative analysis between our proposed method and various SOTA techniques for Covid 19 detection incorporating the implementation of architectures based on transformers as well as with SOTA employing EffNet based networks. While variations in datasets may limit direct performance comparisons, our study showcases the relative improvement achieved by combining ViTs and EffNets. We surpassed existing research that relied solely on either pure ViT or EffNets for the same problem statement. This finding underscores the potential of our approach and offers valuable insights, indicating the most promising results among the listed techniques in terms of performance. It suggests that a hybrid network, effectively integrating transformers and CNNs, can leverage the strengths of both models, outperforming pure CNNs and pure transformers in the task of **Covid** 19 image classification. The comparison of our results with the existing studies that have employed transformer-based networks and EffNet based architectures for diagnosing Covid 19 has been demonstrated in Table 4 and Table 5 respectively.

## 5. Conclusion and Future work

A novel model has been introduced in this work for the automated prognosis of Covid 19 through the analysis of CT images. The dataset utilized comprises 8439 CT scan images. The noteworthy result underscores the capability to employ the ViT architecture for diverse CV applications. We have empirically demonstrated the effec-tiveness of our novel model in comparison to SOTA approaches, evaluated through a variety of metrics. In our future endeavors, we aim to evaluate the model's perfor-mance on CXR images also and expand the assessment to encompass larger datasets. We aim to attain even greater levels of precision through ongoing research and optimization efforts.

## Data Availability

The dataset substantiating the findings in this study is openly accessible at:

https://www.kaggle.com/datasets/mehradaria/covid19-lung-ct-scans

## References

[1] M. Hemalatha, ‘A hybrid random forest deep learning classifier empowered edge cloud architecture for COVID-19 and pneumonia detection’, *Expert Syst Appl*, vol. 210, Dec. 2022, doi: 10.1016/j.eswa.2022.118227.

[2] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, and J. Liu, ‘Chest CT for Typical Coronavirus Disease 2019 (COVID-19) Pneumonia: Relationship to Negative RT-PCR Testing’, *Radiology*, vol. 296, no. 2, pp. E41–E45, Aug. 2020, doi: 10.1148/radiol.2020200343.

[3] F. A. Breve, ‘COVID-19 detection on Chest X-ray images: A comparison of CNN architectures and ensembles[Formula presented]’, *Expert Syst Appl*, vol. 204, Oct. 2022, doi: 10.1016/j.eswa.2022.117549.

[4] H. Y. F. Wong *et al.*, ‘Frequency and Distribution of Chest Radiographic Findings in Patients Positive for COVID-19’, *Radiology*, vol. 296, no. 2, pp. E72–E78, Aug. 2020, doi: 10.1148/radiol.2020201160.

[5] M. Scarpiniti, S. Sarv Ahrabi, E. Baccarelli, L. Piazzo, and A. Momenzadeh, 'A novel unsupervised approach based on the hidden features of Deep Denoising Autoencoders for COVID-19 disease detection', *Expert Syst Appl*, vol. 192, Apr. 2022, doi: 10.1016/j.eswa.2021.116366.

[6] J. Chen *et al.*, 'TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation'. [Online]. Available: https://github.com/Beckschen/

[7] H.-I. Suk, M. Liu, P. Yan, and C. Lian, Eds., *Machine Learning in Medical Imaging*, vol. 11861. in Lecture Notes in Computer Science, vol. 11861. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-32692-0.

[8] F. Demir, 'DeepCoroNet: A deep LSTM approach for automated detection of COVID-19 cases from chest X-ray images', *Appl Soft Comput*, vol. 103, May 2021, doi: 10.1016/j.asoc.2021.107160.

[9] C. Szegedy *et al.*, 'Going Deeper with Convolutions', Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.4842

[10] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[11] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, 'Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks', Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.12348

[12] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, 'Densely Connected Convolutional Networks', Aug. 2016, [Online]. Available: http://arxiv.org/abs/1608.06993

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, 'ImageNet: A Large-Scale Hierarchical Image Database'. [Online]. Available: http://www.image-net.org.

[14] G. Marques, D. Agarwal, and I. de la Torre Díez, 'Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network', *Applied Soft Computing Journal*, vol. 96, Nov. 2020, doi: 10.1016/j.asoc.2020.106691.

[15] A. Gupta, Anjum, S. Gupta, and R. Katarya, 'InstaCovNet-19: A deep learning classification model for the detection of COVID-19 patients using Chest X-ray', *Appl Soft Comput*, vol. 99, Feb. 2021, doi: 10.1016/j.asoc.2020.106859.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, 'Rethinking the Inception Architecture for Computer Vision', Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.00567

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, 'MobileNetV2: Inverted Residuals and Linear Bottlenecks'.

[18] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[19] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, 'Learning Transferable Architectures for Scalable Image Recognition', in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Dec. 2018, pp. 8697–8710. doi: 10.1109/CVPR.2018.00907.

[20] F. Chollet, 'Xception: Deep learning with depthwise separable convolutions', in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Institute of Electrical and Electronics Engineers Inc., Nov. 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.

[21] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, 'ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design', Jul. 2018, [Online]. Available: http://arxiv.org/abs/1807.11164

[22] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, 'SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size', Feb. 2016, [Online]. Available: http://arxiv.org/abs/1602.07360

[23] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, 'GhostNet: More Features from Cheap Operations', Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.11907

[24] A. Howard *et al.*, 'Searching for MobileNetV3', May 2019, [Online]. Available: http://arxiv.org/abs/1905.02244

[25] M. Tan and Q. V. Le, 'EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks', May 2019, [Online]. Available: http://arxiv.org/abs/1905.11946

[26] H. I. Hussein, A. O. Mohammed, M. M. Hassan, and R. J. Mstafa, 'Lightweight deep CNN-based models for early detection of COVID-19 patients from chest X-ray images', *Expert Syst Appl*, p. 119900, Aug. 2023, doi: 10.1016/j.eswa.2023.119900.

[27] A. Vaswani *et al.*, 'Attention Is All You Need', Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762

[28] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. [Online]. Available: https://github.com/tensorflow/tensor2tensor

[29] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, 'Transformer in Transformer', Feb. 2021, [Online]. Available: http://arxiv.org/abs/2103.00112

[30] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, 'Training data-efficient image transformers & distillation through attention', Dec. 2020, [Online]. Available: http://arxiv.org/abs/2012.12877

[31] J. Guo *et al.*, 'CMT: Convolutional Neural Networks Meet Vision Transformers'. [Online]. Available: https://gitee.com/mindspore/models/tree/master/research/cv/CMT.

[32] Y. Tang *et al.*, 'An Image Patch is a Wave: Phase-Aware Vision MLP'.

[33] J. Guo *et al.*, 'Hire-MLP: Vision MLP via Hierarchical Rearrangement', Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.13341

[34] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, 'Conditional Positional Encodings for Vision Transformers', Feb. 2021, [Online]. Available: http://arxiv.org/abs/2102.10882

[35] H. Wu *et al.*, 'CvT: Introducing Convolutions to Vision Transformers', Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.15808

[36] Y. Tang *et al.*, 'Patch Slimming for Efficient Vision Transformers', Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.02852

[37] Z. Liu *et al.*, 'Swin Transformer: Hierarchical Vision Transformer using Shifted Windows', Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.14030

[38] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, 'Incorporating Convolution Designs into Visual Transformers', Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.11816

[39] W. Wang *et al.*, 'Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions', Feb. 2021, [Online]. Available: http://arxiv.org/abs/2102.12122

[40] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, 'Deformable DETR: Deformable Transformers for End-to-End Object Detection', Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.04159

[41] A. Dosovitskiy *et al.*, 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[42] H. Chen *et al.*, 'Pre-Trained Image Processing Transformer', Dec. 2020, [Online]. Available: http://arxiv.org/abs/2012.00364

[43] X. Gao, Y. Qian, and A. Gao, 'COVID-VIT: Classification of Covid-19 from CT chest images based on vision transformer models'. [Online]. Available: https://github.com/xiaohong1/COVID-ViT

[44] X. Gao *et al.*, 'COVID-VIT: Classification of Covid-19 from 3D CT chest images based on vision transformer model', in *Proceedings - 3rd International Conference on Next Generation Computing Applications, NextComp 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/NextComp55567.2022.9932246.

[45] H. A. Dehkordi, H. Kashiani, A. A. Hamidi Imani, and S. B. Shokouhi, 'Lightweight Local Transformer for COVID-19 Detection Using Chest CT Scans', in *ICCKE 2021 - 11th International Conference on Computer Engineering and Knowledge*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 328–333. doi: 10.1109/ICCKE54056.2021.9721517.

[46] J. Mei, 'Marrying Convolution and Transformer for COVID-19 Diagnosis Based on CT Scans', in *Proceedings of the International Joint Conference on Neural Networks*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/IJCNN55064.2022.9892015.

[47] S. A. Jalalifar and A. Sadeghi-Naini, 'Data-Efficient Training of Pure Vision Transformers for the Task of Chest X-ray Abnormality Detection Using Knowledge Distillation', in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1444–1447. doi: 10.1109/EMBC48229.2022.9871372.

[48] P. Silva *et al.*, 'COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis', *Inform Med Unlocked*, vol. 20, Jan. 2020, doi: 10.1016/j.imu.2020.100427.

[49] T. Anwar and S. Zakir, 'Deep learning based diagnosis of COVID-19 using chest CT-scan images', in *Proceedings - 2020 23rd IEEE International Multi-Topic Conference, INMIC 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/INMIC50486.2020.9318212.

[50] R. Chatterjee, A. Datta, R. Halder, and A. Chatterjee, 'An Efficient Pneumonia Detection from the Chest X-Ray Images'. [Online]. Available: https://www.researchgate.net/publication/354062844

[51] E. Luz, P. L. Silva, R. Silva, L. Silva, G. Moreira, and D. Menotti, 'Towards an Effective and Efficient Deep Learning Model for COVID-19 Patterns Detection in X-ray Images', Apr. 2020, doi: 10.1007/s42600-021-00151-6.