

# COVID-19 Diagnosis in 3D Chest CT Scans with Attention-Based Models

Citation for published version (APA):

Hartmann, K., & Hortal, E. (2023). COVID-19 Diagnosis in 3D Chest CT Scans with Attention-Based Models. In J. M. Juarez, M. Marcos, G. Stiglic, & A. Tucker (Eds.), *Artificial Intelligence in Medicine: AIME 2023* (Vol. 13897, pp. 229-238). Springer, Cham. [https://doi.org/10.1007/978-3-031-34344-5\\_27](https://doi.org/10.1007/978-3-031-34344-5_27)

## Document status and date:

Published: 01/01/2023

## DOI:

[10.1007/978-3-031-34344-5\\_27](https://doi.org/10.1007/978-3-031-34344-5_27)

## Document Version:

Publisher's PDF, also known as Version of record

## Document license:

Taverne

## Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

## General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.umlib.nl/taverne-license](http://www.umlib.nl/taverne-license)

## Take down policy


If you believe that this document breaches copyright please contact us at:

[repository@maastrichtuniversity.nl](mailto:repository@maastrichtuniversity.nl)

providing details and we will investigate your claim.



# COVID-19 Diagnosis in 3D Chest CT Scans with Attention-Based Models

Kathrin Hartmann and Enrique Hortal<sup>(✉)</sup> 

Department of Advanced Computing Sciences, Maastricht University, Maastricht,  
The Netherlands

kathrin.hartmann@student.maastrichtuniversity.nl,  
enrique.hortal@maastrichtuniversity.nl

**Abstract.** The three-dimensional information in CT scans reveals notorious findings in the medical context, also for detecting symptoms of COVID-19 in chest CT scans. However, due to the lack of availability of large-scale datasets in 3D, the use of attention-based models in this field is proven to be difficult. With transfer learning, this work tackles this problem, investigating the performance of a pre-trained TimeSformer model, which was originally developed for video classification, on COVID-19 classification of three-dimensional chest CT scans. The attention-based model outperforms a DenseNet baseline. Furthermore, we propose three new attention schemes for TimeSformer improving the accuracy of the model by 1.5% and reducing runtime by almost 25% compared to the original attention scheme.

**Keywords:** Vision Transformer · Medical imaging · Attention Schemes · COVID-19 · 3D CT scan

## 1 Introduction

One recent research field that has rapidly evolved due to the pandemic is the identification of COVID-19 symptoms in lung images. Previous work includes approaches for the classification, detection and segmentation of COVID-19 images among others [14]. COVID-19 classification with deep learning models leads to especially good results when the models are trained on three-dimensional CT scans which are likely to reveal most information about the disease as the symptoms of COVID-19 might be present at different depth levels in the lung [14]. This leads us to hypothesize that, incorporating depth dependencies can help with the task at hand.

At the same time, Vision transformers [3], which are an adaption of the classical transformer architecture for images, are gaining popularity in computer vision. Similarly, the use of attention-based models in the medical context has grown significantly in the last couple of years [14]. With enough data available, these models have been able to outperform classical Convolutional Neural Networks in several tasks in the medical field [14]. However, due to the lack of availability of large-scale datasets, especially three-dimensional ones, research in

COVID-19 classification has mostly been focusing on 2D CT scans using CNN-based models, as attention-based models need a large amount of data to be trained on [3].

To investigate this research gap of attention-based model classification on 3D images in the medical field, we aim to apply attention-based models on 3D chest CT images to identify lungs affected by COVID-19. With transfer learning (using models pre-trained on images from a different domain), we want to overcome the challenges posed when using attention-based models on small datasets. We aim to achieve more accurate results compared to traditional Convolutional Neural Network approaches by embedding information globally across the overall image using attention schemes. Apart from that, we want to investigate how our different, newly developed attention schemes perform compared to previously developed attention schemes both performance and time-wise (reducing the computational power required). To the best of our knowledge, 3D attention-based models have not been applied to the proposed task. The only 3D approaches in the literature are using CNN or U-Net approaches such as [6] and those using a 3D attention-based model are not intended for medical image classification but 3D segmentation such as [15].

## 2 Related Works

### 2.1 COVID-19 Datasets

A few datasets containing 3D chest CT scan images like the ones presented in [10] and [11], MIA-COV19 [9], COV19 CT DB [13] and CC-CCII [17] have been collected. However, most of them are not publicly available. On its part, COV19 CT DB contains 3D CT scans of lungs infected with COVID-19 from around 1000 patients but no other healthy or scans from lungs with other medical conditions are considered. Finally, the CC-CCII dataset contains three classes, namely Common Pneumonia, COVID-19 and Normal lung scans. This database is open access and was pre-processed and restructured in [5]. In this work, we are using and adjusting this pre-processed CC-CCII dataset to conduct our research.

### 2.2 Convolutional Neural Networks for COVID-19 Detection

Convolutional Neural Networks (CNNs) are very popular for image classification and are essential for computer vision in medical imaging. Naturally, CNNs have been widely used for COVID-19 classification over the last three years, also using chest CT scans. Several works state that ResNet [4] is the best-performing network when comparing performance to other nets [1, 10, 12]. Other research [5], however, shows that DenseNets are able to outperform ResNets when using 3D convolutions. In this respect, a DenseNet121 achieves the highest scores on the dataset CC-CCII [17]. In view of the above, the DenseNet121 model is utilized as the baseline model in this work.

### 2.3 Attention-Based Models for COVID-19 Detection

A few approaches have been developed using 3D CT scans. Two of them are the works presented in [7] and [18] which use the Swin transformer in their network to distinguish 3D CT scans based on the classes “COVID-19” and “healthy”. Both models are trained on the MIA-COV19D dataset. The work in [18] uses a U-Net for lung segmentation and then classifies the segmented lung scans with a Swin transformer network. Authors in [7] propose two networks: The first one determines the importance of single slices in a scan based on symptoms shown in it via the Wilcoxon signed-rank test [16] on features extracted with a Swin transformer block. The second network is hybrid, consisting of a CNN that extracts features of each CT scan slice and two different Swin transformers: one captures within slice dependencies and the other is used to identify between slice dependencies. In this way, the full context of the 3D scan is captured. Both of the proposed models outperform a DenseNet201 that was trained for comparison. In our work, on the contrary, we are not using a Swin transformer but an original vision transformer model that was adjusted for three-dimensional, originally video, input. We are leveraging the 3D information in the CT scans by applying several attention schemes that process the CT scans in different ways to explore both spatial and temporal dependencies.

## 3 Methodology

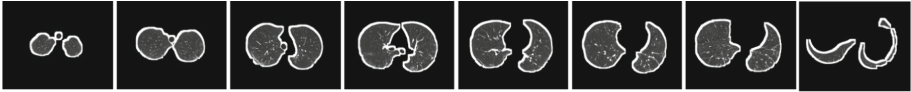
We propose the first application of a pure vision transformer-based model for COVID-19 CT scan classification that is using the 3D information in the CT scans. This is done by applying a pre-trained TimeSformer model [2] on a pre-processed dataset.

### 3.1 Dataset

The CC-CCII dataset [5, 17] is a publicly available 3D chest CT scan dataset that we modify for our research purpose with appropriate corrections. The dataset contains three different classes: lungs diagnosed with Common Pneumonia (CP), lungs diagnosed with Novel Corona Virus (NCP), and lungs without any condition (Normal). In this study, only the first two classes, namely CP and NCP, are considered. Slices containing mistakes in the order of the lung slices were discarded. Apart from that, for consistency and to reduce the required computational power, we sample the number of slices in a scan to 32 (scans with less than 32 slices are also discarded), crop, and resize the lung slices. Our final dataset contains a total of 1874 scans of width  $\times$  height  $\times$  number slices =  $160 \times 128 \times 32$ , 824 of them in class CP and 1047 in class NCP. A part of a randomly selected sample of a lung scan from the final dataset can be seen in Fig. 1.

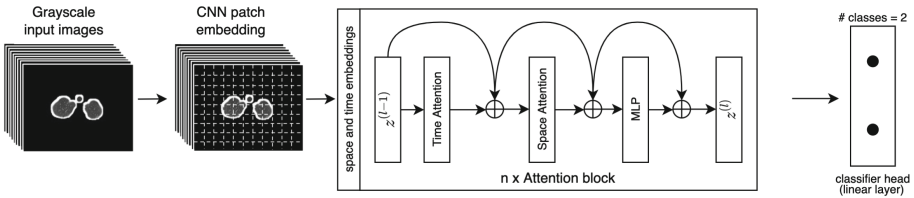
### 3.2 TimeSformer

To be able to efficiently train a model that can distinguish between diseases of 3D lung scans, we examine a domain outside of 3D medical imaging that also requires



**Fig. 1.** The extract of a random lung scan from the utilized dataset. The scans, from left to right, correspond with lung slices from top to bottom.

3D inputs: video classification. One video can be seen as 2D images (frames) stacked together, which also corresponds to our application of 3D CT scans. To build efficient vision transformer models for video classification, authors in [2] have developed TimeSformer, a model that takes 3D inputs (videos), divides the video frames into patches and feeds these patches to a transformer network that consists of  $n$  attention blocks. Within these blocks, embeddings and attention schemes are applied to efficiently classify the videos. Our modified version of their architecture can be seen in Fig. 2.

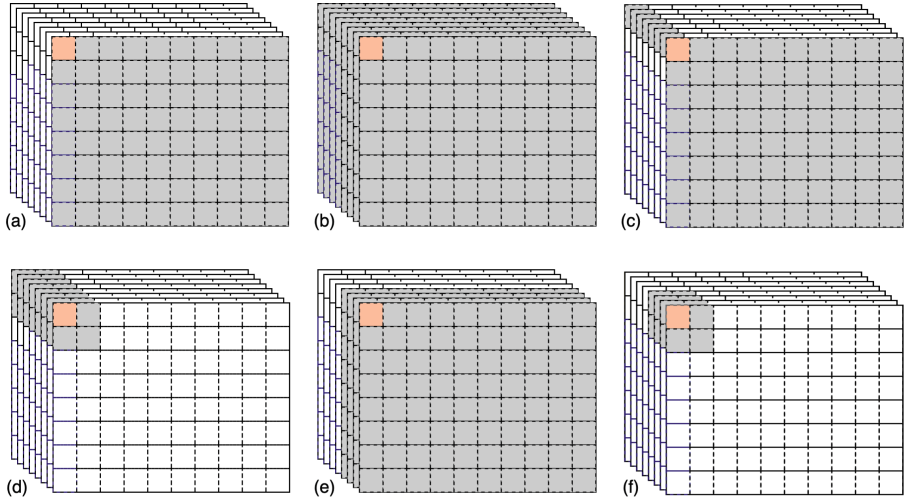


**Fig. 2.** Our TimeSformer setup. 3D grey-scale input images are embedded into patches and fed into a Transformer encoder. The encoder first applies spacial and time embeddings and then inputs these embeddings into  $n$  attention blocks. The architecture of the attention block depends on the attention scheme. Eventually, the outcome of the encoder is classified by a linear layer.

### 3.3 Original Attention Schemes

Authors in [2] compare the implementation of three different attention schemes, namely joint-space-time attention, space-only attention and divided-space-time attention. Joint-space-time-attention calculates the attention between all patches of all layers (see Fig. 3 (b)). This approach is, however, computationally intensive, as the attention between all existing patches is calculated. For this reason, other attention schemes were proposed. Space-only attention, on the contrary, calculates attention only between patches on the same layer (Fig. 3 (a)). This way of calculating attention is less computationally intensive than joint-space-time attention but does not consider dependencies across layers and therefore, ignores the “depth” information. The third attention scheme, divided-space time attention is developed to focus on both spacial and time information in the video. It calculates the attention within all patches at the same temporal position (across frames) and the attention with all patches in the current layer (spatial attention within the frame) (Fig. 3 (c)). Both of these attentions are calculated separately

and then combined and fed into a multilayer perceptron. The paper also proposes temporal embedding in addition to spatial embedding to give the model more depth information.



**Fig. 3.** The attention schemes presented in [2]: space-only attention (a), joint-space-time attention (b), divided-space-time attention (c); and the three newly proposed schemes: space-limited attention (d), time-limited attention (e) and space-and-time-limited attention (f). The patch under analysis is highlighted in orange while the patches considered in each attention scheme are represented in grey. (Color figure online)

### 3.4 New Attention Schemes

The attention schemes described above were developed for video classification. In our use case, we hypothesize that the parts affected by COVID-19 or Common Pneumonia may be spread in depth over a bigger area than a single patch. From the proposed schemes, divided-space-time attention only considers one patch in the time dimension while space-only attention does not consider depth information. As joint-space-time attention is highly computationally intensive, we want to evaluate other attention schemes able to 1) capture both spatial and time dependencies and 2) reduce the time computational power required. We propose three new attention calculation schemes for our use case: space-limited attention, time-limited attention and space-and-time-limited attention. Per each patch, the space-limited attention considers the total time dimension but only focuses on a subarea around the patch under analysis (see Fig. 3 (d)). To that end, non-overlapping squares of adjacent patches, for example,  $4 \times 4$  patches or  $2 \times 2$  patches, are utilized on each slice. Time-limited attention does the opposite: while considering the total space dimension, a limited number of adjacent slices

in the time (depth in our case) direction are considered (Fig. 3 (e)). Finally, a combination of both attention schemes is proposed as space-and-time-limited attention. This attention scheme uses non-overlapping cubes of patches, smaller than the original width and height (Fig. 3 (f)).

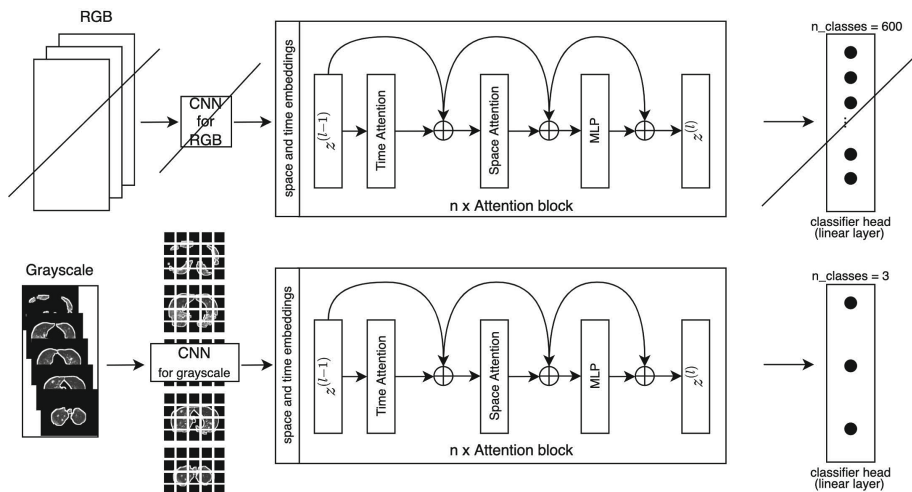
### 3.5 Experimental Setup

As the baseline, a DenseNet121 [8] is trained from scratch on our dataset. This model is selected as it utilizes 3D convolutions to capture depth information and therefore, can be considered a fairer comparison to our attention-based models than 2D approaches. The DenseNet is trained with a batch size of 16 for a maximum of 50 epochs with a patience factor of 15 for early stopping and a learning rate decay factor of  $0.1 \frac{\text{epoch}}{20}$ , starting off from a learning rate of 0.1.

Due to the large amount of data necessary to train attention-based models, we utilize a pre-trained TimeSformer model made available by [2] that we fine-tune on our dataset. Pre-trained weights from training on the video dataset Kinetics-600 are used. It is worth mentioning that we modified the architecture to fit our dataset input. Thus, the CNN input layer from TimeSformer, with 3 (RGB) channels, is replaced by a 1-channel (grey-scale) input. Similarly, the output layer is modified to accommodate two classes, instead of the 600 classes on the Kinetics-600 dataset. Figure 4 shows our modified TimeSformer model. As an initial evaluation, this modified divided-space-time attention (*dst*) model is compared with the DenseNet model explained in Sect. 2.2.

Consecutively, the proposed space-limited (*sl*), time-limited (*tl*) and space-and-time-limited (*stl*) attention schemes are evaluated. For consistency, the implementation of these models is as similar as possible to the initial divided-space-time attention approach. The modified input and output layers of TimeSformer and the number of attention blocks are consistent across the three models. The sole implementation difference across the proposed models is the attention blocks. Depending on the attention scheme applied, the original *dst* attention blocks are replaced by the proposed ones. The same pre-trained weights from training on the video Kinetics-600 dataset, as in the previous experiment, are used. This means that the model was pre-trained with a different attention scheme (divided-space-time attention) and is now fine-tuned on one of the newly proposed ones. In total, 12 attention blocks are used for all experiments.

All the above-mentioned experiments are conducted using the dataset described in Sect. 3.1. The train-validation-test split is set fixed during pre-processing to avoid having scans of the same patient in different sets. We randomly split the patients in the dataset into training, validation and test data such that the ratio of scans in each set is 60-20-20. After pre-processing, we have a final input dataset with 1181 scans in the training dataset, 361 scans in the validation dataset and 332 scans in the test dataset. The classes are also balanced as much as possible, with 827 and 1047 CP and NCP instances respectively. All the experiments have been conducted in the same machine, using an NVIDIA® Tesla V100 32GB GPU.



**Fig. 4.** Modification of TimeSformer model to make it suitable for our dataset. The 3-channel CNN input layer that divides the input images into patches is replaced by a 1-channel grey-scale input layer. The attention blocks remain unchanged, containing the calculation of time attention, space attention and a multilayer perceptron, connected with skip connections. The output layer is modified to classify two classes instead of the original 600.

### 3.6 Evaluation Metrics

The performance of the models is statistically analyzed by using the following evaluation metrics:

Accuracy ( $acc$ ), calculated as the number of correctly classified instances divided by the total number of them:

$$acc = (\# \text{ instances correctly classified}) / (\# \text{ total instances}) \quad (1)$$

Precision ( $prec$ ), calculated by dividing the number of true positive (TP) samples by the number of true positive and false positive (FP) samples:

$$prec = (TP) / (TP + FP) \quad (2)$$

Recall ( $rec$ ), calculated by dividing the number of true positive instances by the number of true positive and false negative (FN) ones:

$$rec = (TP) / (TP + FN) \quad (3)$$

Additionally, we also calculate specificity. This metric assesses how the negative class performs and it is calculated as the number of true negatives (TN) instances divided by the number of TN and FP:

$$spec = (TN) / (TN + FP) \quad (4)$$



Finally, the weighted average F1 score is calculated. This metric measures a combination of precision and recall for each class and weights it by the number of instances in the class as:

$$weighted\_avg\_f1 = \sum_{n=1}^C f1_{C_n} * W_{c_i} \quad (5)$$

where

$$W_{c_i} = (\# \text{ instances in class } c_i) / (\# \text{ total instances}) \quad (6)$$

$$f1 = 2 * (prec * rec) / (prec + rec) \quad (7)$$

Apart from these statistical metrics, the training runtime was also measured. This information is a good indicator of the computational power required to train each of the models proposed.

## 4 Results and Discussion

Table 1 shows the resulting metrics after running the proposed baseline, the 3D DenseNet121 (*3D DN*) and five different attention schemes on the dataset described in Sect. 3.1. We successfully fine-tuned the customized divided-space-time attention (*dst*) and space-only attention (*so*) models presented in [2] and the three newly proposed schemes, namely space-limited (*sl*), time-limited (*tl*) and space-and-time-limited attention (*stl*). The joint-space-time attention scheme from [2] could not be evaluated on this dataset due to computational power limitations. Additionally, it is worth mentioning that, also due to computational power restrictions, the attention schemes evaluated could only be run with a batch size of 16. The best results for all models were achieved when fine-tuning for 20 epochs. After exhaustive experimentation, the best-performing window size for space-limited attention was a window of  $2 \times 2$  and the depth with the highest accuracy for time-limited attention was 8. Thus, these results are combined and a cube size of  $2 \times 2 \times 8$  is used for training the model with space-and-time-limited attention.

**Table 1.** The results for the 3D DenseNet121 baseline and the fine-tuning TimeSformer models with five different attention schemes.

Attention type	acc	prec	recall	specificity	weighted avg f1	runtime (mins)
3D DN	0.777	0.818	0.818	0.713	0.777	<b>32</b>
TSf dst	0.798	0.818	0.862	0.698	0.796	75
TSf so	0.798	0.812	<b>0.872</b>	0.682	0.796	54
TSf sl	<b>0.813</b>	<b>0.844</b>	0.852	0.752	<b>0.813</b>	57
TSf tl	0.633	0.758	0.586	0.705	0.637	68
TSf stl	0.783	0.839	0.798	<b>0.760</b>	0.784	55

Among all the attention-based models evaluated, the proposed *TSf sl* (TimeSformer with space-limit attention) scheme outperforms the rest, achieving around 4% improvement in accuracy over the 3D DenseNet121 [8] baseline model (from 0.777 to 0.813) and a 1.5% improvement over the original schemes, namely *dst* and *so*. This model also surpasses the original models in precision (higher than 3% improvement), specificity (between 5.4 and 7%) and the weighted average F1 (1.7%). Additionally, this model is able to reduce the training runtime by almost 25% compared to the original attention scheme divided-space-time attention (from 75 to 57 min in our setup).

## 5 Conclusion

In this work, we proposed three newly developed attention schemes in addition to the attention schemes developed for TimeSformer [2]. These schemes are proposed with the aim of reducing the computationally intensive training process while maintaining or even improving the performance of our classification models. Our results indicate that our space-limited attention scheme yields better results compared to all other schemes and baseline for distinguishing between scans from lungs affected by COVID-19 and Common Pneumonia. This finding corroborates our hypothesis that capturing the time (depth in our case) dependencies play an important role in the detection of lung diseases.

However, as future work, it would be advisable for a more in-depth evaluation of the proposed attention schemes. These newly proposed attention schemes should be further investigated by validating more combinations of patch shapes, both in the spatial and temporal dimensions. Nevertheless, it is worth stressing that, even though the use of bigger patches could help identify the time and space dependencies more accurately, it will come at the expense of a higher computational power. Furthermore, with more computational power, it would be also possible to design more comparable experiments and train DenseNets and TimeSformer on the same batch size as the original works. Moreover, more fine-tuning of the models could be done to boost their performance. With more resources, it would also be possible to evaluate how the joint-space-time attention scheme performs compared to the proposed attention approaches. To conclude, another very interesting experiment we are planning to conduct is the evaluation of the proposed models on different 3D lung CT scan datasets and more classes to get further insights into how generally valid the results achieved in this work are. Finally, for the medical field, more research in the visualization of attention and the explanation of models for this application is of interest.

## References

1. Ardakani, A.A., Kanafi, A.R., Acharya, U.R., Khadem, N., Mohammadi, A.: Application of deep learning technique to manage covid-19 in routine clinical practice using CT images: results of 10 convolutional neural networks. *Computers in biology and medicine* 121 (2020)
2. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: *ICML*, vol. 2, p. 4 (2021)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)* (2020)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
5. He, X., et al.: Benchmarking deep learning models and automated model design for covid-19 detection with chest CT scans. *MedRxiv* (2021)
6. Hatamizadeh, A., Tang, Y., Nath, V., et al.: Unetr: transformers for 3D medical image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584 (2022)
7. Hsu, C.-C., Chen, G.-L., Wu, M.-H.: Visual transformer with statistical test for covid-19 classification. *arXiv preprint [arXiv:2107.05334](https://arxiv.org/abs/2107.05334)* (2021)
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708 (2017)
9. Kollias, D., Arsenos, A., Soukissian, L., Kollias, S.: Mia-cov19d: Covid-19 detection through 3-D chest CT image analysis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 537–544 (2021)
10. Li, L., Qin, L., Xu, Z., et al.: Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. *Radiology* **296**(2), 65–71 (2020)
11. Mishra, A.K., Das, S.K., Roy, P.: Bandyopadhyay, S.: Identifying covid19 from chest CT images: a deep convolutional neural networks based approach. *J. Healthcare Eng.* (2020)
12. Pham, T.D.: A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks. *Sci. Rep.* **10**(1), 1–8 (2020)
13. Shakouri, S., et al.: Covid19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed covid-19 diagnosis. *BMC Res Notes* (2021)
14. Shamshad, F., et al.: Transformers in medical imaging: a survey. *arXiv preprint [arXiv:2201.09873](https://arxiv.org/abs/2201.09873)* (2022)
15. Shin, Y., Eo, T., Rha, H., et al.: Digestive Organ Recognition in Video Capsule Endoscopy Based on Temporal Segmentation Network. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) *MICCAI 2022*. LNCS, vol. 13437, pp. 136–146. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-16449-1\\_14](https://doi.org/10.1007/978-3-031-16449-1_14)
16. Woolson, R.F.: Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, 1–3 (2007)
17. Zhang, K., Liu, X., Shen, J., et al.: Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell* **181**(6), 1423–1433 (2020)
18. Zhang, L., Wen, Y.: A transformer-based framework for automatic covid19 diagnosis in chest CTs. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 513–518 (2021)