Deep Learning algorithms for automatic COVID-19 detection on chest X-ray images

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of current version 09/11/2022.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Deep Learning Algorithms for Automatic COVID-19 Detection on Chest X-Ray Images

Sergio Cannata[1], Student Member, IEEE, Annunziata Paviglianiti[1], Member, IEEE, Eros Pasero[1], Member, IEEE, Giansalvo Cirrincione[2], Fellow, IEEE, and Maurizio Cirrincione[3], Fellow, IEEE

[1]Department of Electronics and Telecommunications, Politecnico di Torino, Turin, 10129 ITALY
[2]University of Picardie Jules Verne, Amiens, 80000 FRANCE
[3]University of the South Pacific, Suva, 1168 FIJI

Corresponding author: Sergio Cannata (e-mail: sergio.cannata@polito.it).

ABSTRACT Coronavirus disease (COVID-19) was confirmed as a pandemic disease on February 11, 2020. The pandemic has already caused thousands of victims and infected several million people around the world. The aim of this work is to provide a Covid-19 infection screening tool. Currently, the most widely used clinical tool for detecting the presence of infection is the reverse transcription polymerase chain reaction (RT-PCR), which is expensive, less sensitive and requires the resource of specialized medical personnel. The use of X-ray images represents one of the latest challenges for the rapid diagnosis of the Covid-19 infection. This work involves the use of advanced artificial intelligence techniques for diagnosis using algorithms for classification purposes. The goal is to provide an automatic infection detection method while maximizing detection accuracy. A public database was used which includes images of COVID-19 patients, patients with viral pneumonia, patients with pulmonary opacity, and healthy patients. The methodology used in this study is based on transfer learning of pre-trained networks to alleviate the complexity of calculation. In particular, three different types of convolutional neural networks, namely, InceptionV3, ResNet50 and Xception, and the Vision Transformer are implemented. Experimental results show that the Vision Transformer outperforms convolutional architectures with a test accuracy of 99.3% vs 85.58% for ResNet50 (best among CNNs). Moreover, it is able to correctly distinguish among four different classes of chest X-ray images, whereas similar works only stop at three categories at most. The high accuracy of this computer-assisted diagnostic tool can significantly improve the speed and accuracy of COVID-19 diagnosis.

INDEX TERMS Biomedical imaging, COVID, Deep learning, Image classification, Medical diagnostic imaging, Vision Transformer

I. INTRODUCTION

On December 31st, 2019, Chinese health authorities reported an outbreak of pneumonia cases of unknown aetiology in the city of Wuhan (Hubei Province, China). Shortly thereafter, on January 9th, 2020, the China CDC (the Center for Disease Control and Prevention of China) identified a new coronavirus (tentatively named 2019-nCoV) as the etiological cause of these diseases. Chinese health authorities have also confirmed the inter-human transmission of the virus. On 11th February, the World Health Organization (WHO) announced that the disease transmitted since 2019-nCoV has been called COVID-19 (Corona Virus Disease). The Coronavirus Study Group (CSG) of the International Committee on Taxonomy of Viruses has officially classified with the name of SARS-CoV-2 the virus provisionally named by the international health authorities 2019-nCoV and responsible of cases of COVID-19 (Corona Virus Disease). The CSG - responsible for defining the official classification of viruses and the taxonomy of the Corona viridae family, after evaluating the novelty of the human pathogen and on the basis of phylogeny, taxonomy and established practice, has formally associated this virus with the coronavirus it causes severe acute respiratory syndrome (SARS-CoVs, Severe acute respiratory syndrome coronaviruses) classifying it as Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) [1]. After assessing the severity levels and global spread of

the SARS-CoV-2 infection, WHO declared that the COVID-19 epidemic can be considered a pandemic.

II. RELATED WORK

After the World Health Organization (WHO) declared the rapid spread of the aggressive COVID-19 virus, the world of scientific research went to great lengths to propose a solution for the early diagnosis of the virus [2]. Indeed, the rapid detection of COVID-19 can help control the spread of the disease.

Nowadays, the most used and most reliable method of diagnosing infection is the Reverse Transcription-Polymerase Chain Reaction (RTPCR). A sample is taken by nose / mouth and pharyngeal swab and analysed by real-time molecular methods through the amplification of the viral genes most expressed during the infection. This analysis can only be carried out in highly specialized laboratories, identified by the health authorities and requires on average from 2 to 6 hours to return a result. Another category of tests that have a lower sensitivity and specificity than the previous molecular tests, are the antigen swabbing. This type of test is based on the search for viral proteins (antigens) in respiratory samples. The sampling methods are the same as for molecular tests (nasal and throat swab) but the response time is shorter (about 15 minutes). Finally, the serological tests highlight the presence of antibodies against the virus and tests reveal that there has been exposure to the virus; but only in a few cases can they detect that an infection is in progress. In the current state of scientific development, serological tests cannot replace molecular tests based on the identification of viral RNA [3].

In recent times, the attention for the diagnosis of infection is focusing on imaging tests. Chest X-ray (CXR) and computed tomography (CT) are the most popular imaging techniques for diagnosing COVID-19 disease. The historical conception of diagnostic imaging systems has been fully explored through several approaches ranging from automation engineering to deep learning [4]. Although some studies [5] show an increase in sensitivity when analyzing CT scans as opposed to CXR, this study focuses on chest X-ray images due to their readiness and wide availability, which is not always the case for CT images [6].

The Convolutional Neural Network (CNN) is one of the most popular and effective approaches in the diagnosis of COVID-19 from digitized images. Several reviews have been carried out to highlight recent contributions to COVID-19 detection [7]. Pre-trained CNN models were used for feature extraction using SVM classifiers with various kernel functions [8]. Then, several pre-trained CNN models were further trained using chest X-ray images for COVID-19 detection. The accuracy of the classification was used to evaluate the performance of the proposed methods. The pre-trained deep CNN models used in the study were ResNet18, ResNet50, ResNet101, VGG16, and VGG19. Since testing the study, the deep characteristics model (ResNet50) and SVM with linear kernel function produced an accuracy score of 94.7%, which was the highest of all results. Test results for fine-tuning the ResNet50 model and end-to-end training of the developed CNN model were 92.6% and 91.6% respectively. Since the number of COVID-19 X-Ray samples is limited, transfer learning (TL) appears as the reference method for classifying disease data to develop accurate automated diagnosis models. In this context, networks are able to acquire knowledge from pre-trained networks on large-scale image datasets or alternative data-rich sources. The classification algorithm based on transfer learning acquired results with an accuracy of 97.66% and an F1-score of 97.61% [9].

The studies suggested that transfer learning can allow the network to extract significant features related to the COVID-19 disease diagnosis. In fact, several works have applied this idea in order to rapidly develop a reliable tool to assist medical experts in diagnosing COVID-19. The wide popularity of convolutional neural networks made them the first choice for a number of works, in which said architectures manage to identify COVID-struck lungs on X-ray images. Shazia *et al.* [10] compared performances of several CNNs, presenting a test accuracy of 99.48% obtained by DenseNet121. However, the classification task only dealt with the goal of distinguishing between COVID and viral pneumonia, with the first (and most relevant) class being represented in the test set with just 157 images, versus more than 4000 images for pneumonia. Many other studies tackle the problem of classifying COVID and non-COVID X-ray images, typically viral/bacterial pneumonia, or they add normal lung CXR as a third category [11] [12].

The advent of the Vision Transformer has led many researchers to perform the same kind of task with such recent neural model, and assess its performance against CNNs. ViT's capability to connect local patches of information on a single image and build up the picture context has led the Vision Transformer to often surpass its convolutional competitors. Thus, many other works have deployed the Vision Transformer for COVID detection. Krishnan applied the ViT to distinguish between COVID and non-COVID chest X-ray images, assessing performance against some CNNs, reaching a final accuracy of 97.6% [13]. D. Shome *et al.* applied the model for both three and two classes, including pneumonia as a third category. The study managed to achieve accuracies of 92% and 98%, respectively [14]. Mondal *et al.* obtained a 96% test accuracy when using the Vision Transformer for classifying chest X-ray images into the same three categories, namely COVID, pneumonia and normal lungs [15]. Park *et al.* added a convolutional backbone for feature extraction, and developed a system to assess COVID severity. However, their work classified X-ray images over three categories – namely COVID, normal and a generic class of "other infection". Authors present test accuracy separately for the three classes with a confidence interval of 95%, and its value never goes above 94.2% (which is the best result for the normal class) [16]. All of the aforementioned papers present studies that are mainly focused on distinguishing between COVID and non-

COVID patients, or they include viral pneumonia as a third class when tackling multi-class classification. Almaki *et al.* took into consideration both viral and bacterial pneumonia, thus working over four classes, and combined their custom CNN with a few selected machine learning algorithms. Yet, their method only reached a final test accuracy of 97.29% [17]. The work presented in this document tackles classification over four categories – including lung opacity as a fourth class – and proves that the Vision Transformer is capable of distinguishing an additional class of pulmonary diseases with considerably high levels of accuracy and specificity. To the best of our knowledge, no other similar methods were able to reach such level of accuracy over four different classes of chest X-ray images in the case of automatic COVID-19 detection.

III. DATABASE DESCRIPTION: CHEST X-RAY IMAGES

The dataset used in this work is a collection of chest X-ray images that were (and still are) gathered by researchers from different countries, with the specific purpose of creating a publicly available database for COVID-related research. The version used for this work was collected in October 2021, and it consists of 3616 COVID-19 positive cases, 10,192 normal images, 6012 pulmonary opacity (non-COVID lung infection), and 1345 viral pneumonia [18][19]. All images were downloaded as png-formatted RGB images with a size of 299x299x3.

The lungs are the two organs responsible for supplying oxygen to the body and for the elimination of carbon dioxide from the blood, or the gaseous exchanges between air and blood (a process known as hematosis). Located in the thoracic cavity, they are surrounded by a serous membrane, the pleura, which is essential for the performance of their functions.

The lungs are separated by a space between the spine and the sternum, the mediastinum, which includes the heart, esophagus, trachea, bronchi, thymus and great vessels.

Each of the two lungs has at the upper end, an apex that extends upwards to the base of the neck and, at the lower end, rests on the diaphragmatic muscle. Their main blood is to receive the load of carbon dioxide and waste products from the peripheral circulation and to clean it up: once cleansed the blood is then sent to the heart, from where it is sent to organs and tissues. An example of healthy lungs X-ray image is shown in Fig. 1.

FIGURE 1. **Healthy lungs on a chest X-ray image.**

In general, in pneumonia, the lungs fill with fluid and become inflamed, causing difficulty breathing. For some cases, breathing problems can become severe and require hospitalization with oxygen and ventilator treatments. Pneumonia caused by COVID-19 tends to take hold in both lungs. The air sacs in the lungs fill with fluid, limiting their ability to absorb oxygen and causing shortness of breath, cough, and other symptoms. Even after the disease has passed, lung lesions can cause breathing difficulties that could take months to improve. An example can be observed in Fig. 2.

FIGURE 2. **COVID-19 lungs on a chest X-ray image.**

Pulmonary opacity is represented by spots that appear on the lungs and usually do not exceed 3 cm in diameter. In most cases they are benign, meaning they are not cancerous. A pulmonary nodule is usually seen by means of chest X-ray or computed tomography (CT). They may appear as single nodules or there may be several. A cancerous lung lump is usually larger than 3 cm and can be irregular in shape. Such nodules can be seen in Fig. 3.

FIGURE 3. Pulmonary opacity lungs on a chest X-ray images. A pulmonary node is highlighted with a red circle.

Viral pneumonia is defined as a pathological entity in which there is the viral cause of abnormal oxygen and carbon dioxide gas exchanges in the alveoli, secondary to virus-mediated inflammation and / or immune response [20]. In X-ray images, areas of the chest are generally visible as lighter, whitish spots in the regions affected by pneumonia, as shown in Fig. 4.

FIGURE 4. Pulmonary opacity lungs on a chest X-ray images. A pulmonary node is highlighted with a red circle.

IV. METHODOLOGY

A. DEEP LEARNING ARCHITECTURES: A CONCEPTUAL FACE-OFF

Ever since their pioneering use [21], Convolutional Neural Networks (CNNs) have proved themselves to be extremely powerful tools when it comes to image classification. The basic principle is the application of convolutional layers, which are able to extract significant features from images by means of a sequence of operations over a selected area of the image itself.

A typical Convolutional Neural Network shall appear as depicted in Fig. 5.

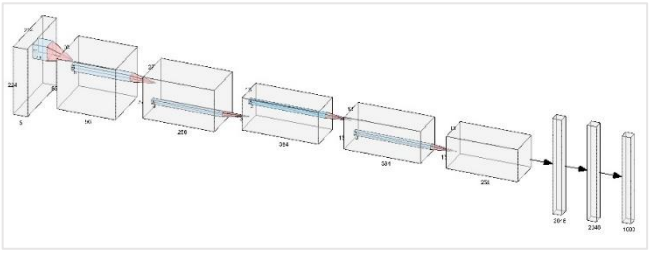


FIGURE 5. Example of a convolutional neural network.

Each convolutional layer is generally followed by a pooling layer, which basically modifies the size of the input in order to make it suitable for the next one. At the very top, a fully-connected or dense layer is present, with the purpose of classifying the input image into one of the given categories, generally applying a softmax function to the input. As powerful as they are, CNNs do exhibit some issues, such as the inability to retain information about the composition and position of specific elements within an image, and to pass such information on to subsequent layers. For this reason, several architectures were developed and presented in recent years. Specifically, Transformers have aroused great interest, especially in NLP applications [22]. In this work the focus is centered on what is probably the most popular version of the Transformer architecture for image classification, the Vision Transformer, or ViT [23]. The peculiar structure and basic working principles of this network are represented in Fig. 6.

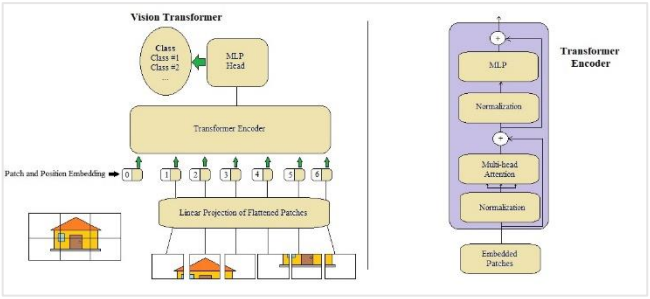


FIGURE 6. Vision Transformer architecture. On the left side, the "0" item indicates a classification token.

Input images are divided into patches, and the linear projection of flattened patches is embedded, in order to preserve positional information. Embedded patches are arranged into a sequence and then fed to the Encoder, which exploits the multi-head attention technique to extract information, patterns and relationships among image patches. Eventually, outputs are fed to a Multi-Layer Perceptron to perform classification.

In order to better clarify the similarities and differences between CNNs and ViT, a brief of the architectures of three among the most relevant and popular Convolutional Networks will follow. InceptionV3 originated as a module for GoogLeNet [24], with the purpose of allowing for deeper networks without increasing too much the number of parameters.

1x1 convolution blocks were introduced to reduce dimensionality.
1x1 convolutional layers act as rectified linear activators as well, so their purpose is two-fold.
The next CNN chosen for analysis and comparison is Xception [25], which was described as an "extreme" version of InceptionV3 with the exploitation of the so-called *depthwise separable convolution*, and a subsequent redefinition of the Inception module.
The basic underlying concept is the assumption that cross-channel correlation and spatial correlation can be mapped separately. This leads to the idea of using a 1x1 convolution to map cross-channel correlations at first, and apply 3x3 convolutions to map spatial correlations later on. This has been proved to slightly outperform InceptionV3.
Remarkably, the "middle flow" section of the network presents a skip connection, which is indeed the key element of the next CNN to be described, ResNet50.
First introduced by He et al. [26], ResNet had the peculiarity of using *skip connections* to tackle the problem of vanishing gradients. This approach was successful and resulted in a number of variations of the original topology.

B. AN INSIGHT INTO MULTI-HEAD ATTENTION

For a better understanding of what the Vision Transformer performs on images, it is worth to delve deeper into the concept of *self-attention*. This idea is derived from the field of Natural Language Processing. In fact, when it comes to translating a text into another language, it is necessary to be aware of the position of each word with respect to each other, and to the context of the sentence in order to give the proper meaning to each word. In this sense, self-attention tries to mimic the thoughts and procedures behind a language translation process.
The whole mechanism starts by assigning to each input three vectors to represent it, namely *key*, *query* and *value*. All of them are obtained by multiplying the input vector by a set of weights (which need to be initialized). Subsequently, each query is multiplied – through a dot product – by each key, and the resulting output is scaled by a factor equal to the square root of the dimension of the key vector. The result goes through a softmax operator, and is later multiplied by the *value* vector. The outcome of this sequence of operations is the *attention score* of the given input. Each block performing such sequence is called *head*.
Keys, queries and values can be linearly projected to $d_{k,}$ d_q and d_v dimensions, where d_k, d_q and d_v are the dimensions of key, query and value vectors, respectively. Thus, attention can be evaluated N times in parallel on each projected version of keys, queries and values, in what is called *multi-head attention*.
After creating image patches, the Vision Transformer embeds the linear project of flattened patches and feeds the embedding to the Encoder Transformer block, allowing for multi-head attention to capture feature and relationships among patches.

C. AN OVERVIEW OF THE VISION TRANSFORMER ARCHITECTURE

Although the peculiar architecture of the Vision Transformer revolves around heads and self-attention, it does not stop at that. In fact, the inputs to each head are embedded both linearly and then again by using sine and cosine functions at different frequencies. This allows to capture information about the position of the single patch with respect to the entire image. A learnable classification token (indicated with "0" on the left side of Fig. 6) is prepended to the sequence of embedded patches so that the network can perform the classification task. In fact, the state of this token at the output of the Multi-layer Perceptron shall represent the input image, retaining significant information for classification.
Subsequently, embedded inputs go through a normalization layer before actually being fed to the Encoder block, and are combined to the Encoder output via a residual connection. The result is once again normalized and then fed to a Multi-Layer Perceptron, which consists of two fully-connected layers with a GELU activation function, reported in (1).

$$GELU(x) = xP(x \leq X) = x\Phi(x) = x \cdot \frac{1}{2}[\mathrm{erf}(x/\sqrt{2})] \quad (1)$$

This section of the network is responsible for performing the actual image classification on the basis of all the information that was extracted and processed by the Transformer heads.

D. EXPERIMENTAL SETUP AND TESTS

The first step of the process consisted in training and testing the three aforementioned CNNs over the chest X-ray dataset. All of the networks were trained by exploiting the transfer learning technique, which allows to retain and freeze network weights derived from previous training sessions over specific datasets. In this case, weights obtained over the ImageNet21k database were used. Subsequently, only some of the top layer weights were set to be trainable. Indeed, the differences among the chosen architectures caused the number of trainable layers to change from one neural network to another. The reduced number of *unfrozen* layer weights were trained and tested over the chest X-Ray database.
As far as the Vision Transformer is concerned, it is worth noting that the fine tuning process for this architecture is rather different with respect to CNNs: in fact, all of ViT's weights are subtly modified during this process, and no layers are actually *frozen*.
The fine-tuning method allowed to significantly reduce the overall amount of time and computational resources dedicated to the training and testing phases for all convolutional networks.
All networks were trained and tested on a PC with a CPU@3.70GHz with TensorFlow 2.5.0 and Keras. Hyperparameters were configured in the very same fashion for all architectures: initial learning rate was set to 0.0001; fine-tuning learning rate to 0.00001; the chosen optimizer was Adam; batch size was set to 32; dropout coefficient equal to 0.5; loss function of choice was the Sparse Categorical Cross Entropy function. This choice for the loss

function is motivated by the fact that the classes of our expression are mutually exclusive, that is, each image belongs to exactly one class. The same number of 60 epochs was set both for the initial training epochs and for fine-tuning epochs, for a total of 120 epochs.

A difference was set in the layer selected to start the fine-tuning process, just as previously mentioned, as follows: the fine-tune was set at layer 308 (over 311 total layers) for InceptionV3; at layer 128 (over 132 total layers) for Xception; at layer 172 (over 175 total layers) for ResNet50.

The next step was the deployment of the Vision Transformer architecture, or ViT, which can be seen as some kind of equivalent of the BERT Transformer [27] applied to vision and image classification. Once again, transfer learning was exploited in order to reduce the total amount of time spent on training, validating and testing.

The following hyperparameters were used: the base architecture is ViT-B_32, which is based on the "base" version of BERT (12 layers, a hidden size set to 768, 12 heads and a patch size of 32x32 for the input); batch size was set to 32; learning rate was set to 0.00001; the selected optimizer was Rectified Adam; loss function of choice was the Categorical Cross Entropy function; label smoothing was set to 0.2; overall number of epochs was 30. All settings and hyperparameters were chosen in order to make a fair comparison against CNNs, and to take into account the significant architectural differences between CNNs and the Vision Transformer.

In order to contrast data imbalance among classes, data augmentation with random horizontal flipping and random rotation (set to 0.2), as well as Mitchell-Netravali [18] filtering were applied to the database images before feeding them to the CNNs. On the other hand, no operation of any kind was performed before feeding the images to the Vision Transformer, with the exception of resizing them from an initial resolution of 299x299x3 to 224x224x3 in order to fit the ViT input layer.

The dataset was split using 70% of data for training, 10% for validation and 20% for testing. The split was kept identical for each model.

V. RESULTS AND DISCUSSION

Table I shows a comparison of the results of the Vision Transformer versus the convolutional networks.

TABLE I
PERFORMANCE COMPARISON BETWEEN VISION TRANSFORMER AND CONVOLUTIONAL NEURAL NETWORK ARCHITECTURES

Network Architecture	Test Accuracy
Inception V3	0.7936
Xception	0.8362
ResNet50	0.8558
ViT	**0.9930**

ResNet50 exhibits the best performance among the convolutional neural networks of choice, with a test accuracy of about 86%. However, the Vision Transformer architecture clearly outperforms the selected CNNs on this specific image classification task with an outstanding accuracy of 99.3%.

A few more indicators are shown in Table II and Fig. 7, 8, 9 and 10 to further describe the performance of the ViT architecture over the four classes of the database: precision, recall, F1-score, a visualization of the attention map and the confusion matrix, all of which are metrics and parameters commonly used to assess the performance of deep learning architectures over given tasks – like image classification. The *Support* column in Table II refers to the number of images per category that were used to test the Vision Transformer ability to assign images to each category.

TABLE II
PRECISION, RECALL AND F1-SCORE FOR VISION TRANSFORMER

	Precision	Recall	F1-score	Support
COVID (Class: 0)	0.97	0.94	0.96	353
Lung Opacity (Class: 1)	0.87	0.93	0.90	602
Normal (Class: 2)	0.95	0.92	0.94	1019
Viral Pneumonia (Class: 3)	0.96	0.98	0.97	135

These results show that the Vision Transformer is highly capable to correctly classify images in each category. This is further verified when evaluating metrics like precision, recall and F1-score, which allow for a better insight of the Vision Transformer performance. In fact, given the overall number of images classified by the network into a given category, precision expresses the ratio of how many of those images have been correctly included by the Transformer into that category. On the other hand, considering the number of images that actually belong to a specific category, recall indicates the proportion of those images that were correctly associated to that class by the neural network. Typically, a high precision score implies a poor recall value, and viceversa. For this reason, F1-score is also taken into account, since it represents a sort of combination of precision and recall into a single metric. As it is possible to observe, all three parameters are remarkably high for the Vision Transformer architecture over all four classes.

The attention map, which can be defined as a matrix that represents the relative importance of layer activations at different spatial locations with respect to the given task, can be observed in Fig. 7, 8 and 9 for all 12 layers of the Vision Transformer. Even though it might not convey any particular information to the human eye, the attention map can help to analyze the behaviour of a network, since it visualizes some aspects of the input image that are interpreted as relevant

features by the architecture, thus leading it to assign such image into a specific class.

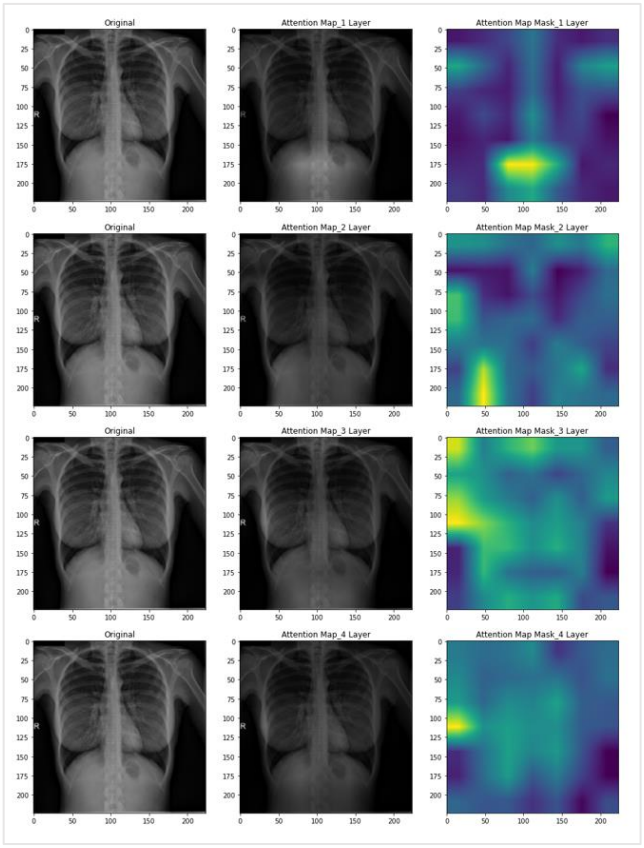


FIGURE 7. Example of Vision Transformer attention maps on a COVID-19 chest X-ray image. Attention maps for layers 1 to 4 are shown.

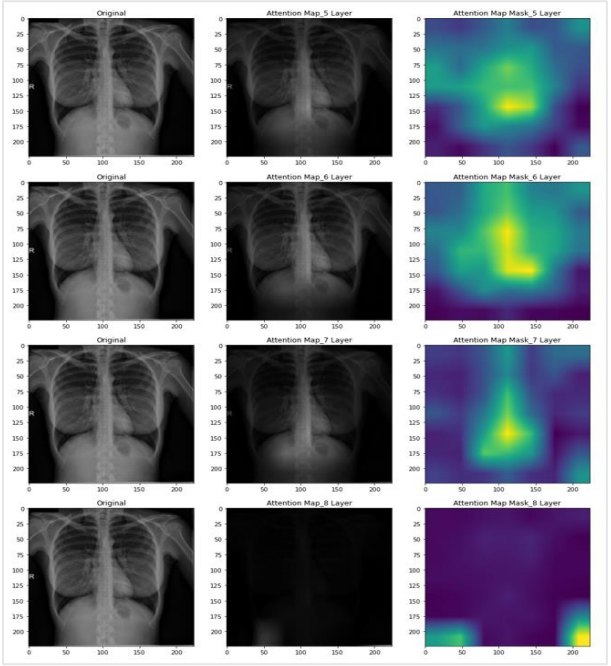


FIGURE 8. Attention maps on the same COVID-19 chest X-ray image, Maps for layers 5 to 8 are shown.

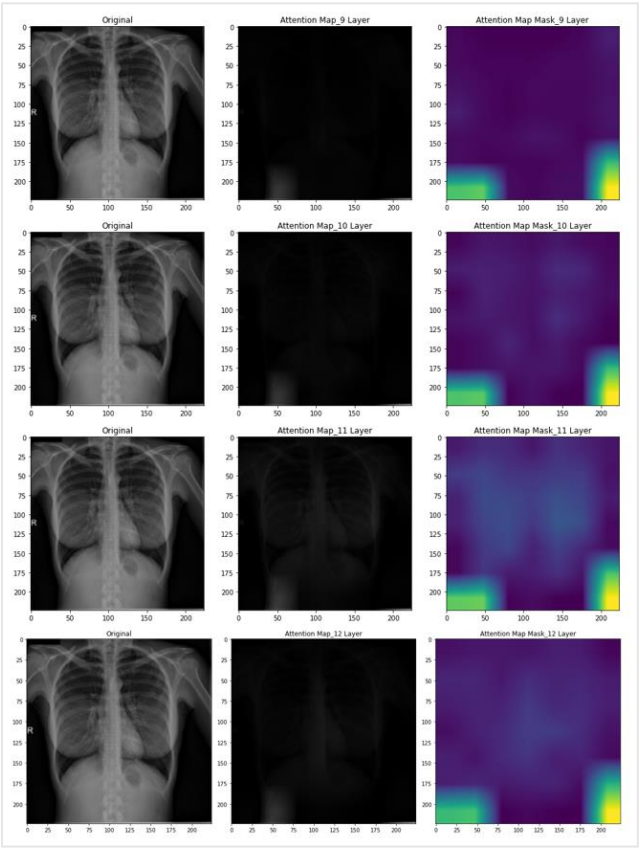


FIGURE 9. Attention maps on the same COVID-19 chest X-ray image, Last maps, for layers 9 to 12, are displayed.

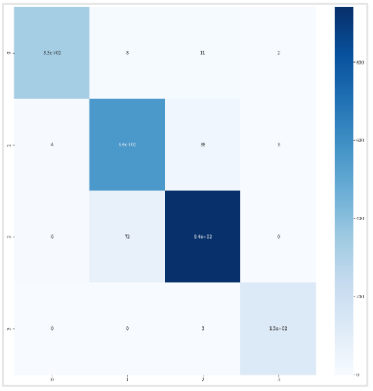


FIGURE 10. Vision Transformer confusion matrix

As a closing remark, it is worth highlighting that the Vision Transformer is able to reach a significantly higher accuracy with respect to Convolutional Neural Networks after iterating for only 30 epochs, as opposed to 60 + 60 = 120 overall epochs to train and fine tune the other architectures.

VI. CONCLUSIONS

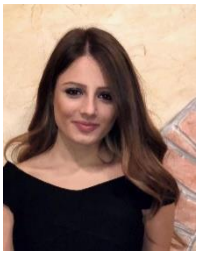
Results presented so far have proved that a specific architecture, the Vision Transformer, is able to achieve a significantly better performance with respect to other network configurations on this peculiar application. This paves the way to the exploitation of the Vision Transformer, and attention-based networks in general, for the purpose of assisting, accelerating and automatizing clinical diagnosis. Indeed, a fast accurate and reliable tool to promptly identify lung infections can assume crucial relevance when the disease of interest is the cause of the current pandemic situation. However, one of the main characteristics of deep learning, and of neural networks in general, is the lack of transparency, meaning that the mechanisms that lead such algorithms when making decisions are often obscure. This often leads to situations in which the network excels at performing its tasks on a given dataset but is unable to generalize over different scenarios. This becomes particularly significant when the primary purpose of the algorithm is to provide a fast and reliable response when assisting physicians in clinical diagnosis. For this reasons, future work in this direction should be dedicated to shed some light on what might lead a deep neural network into making a specific choice when facing alternatives, trying to bring clarity into what has traditionally been perceived as a black box. For instance, an interesting path to follow could be represented by the kind of ideas described by Hassani et al. [29], who tried to exploit the convolutional networks' capabilities to extract significant features from images and feed those information to a Transformer, as opposed to using the patching and embedding method. Another way could be the analysis of how initial data is divided into clusters, and a subsequent comparison against the outcome of the classification task performed by the transformer, in order to investigate the factors that push the network to put a given image into a certain category.

REFERENCES

[1] E.J. Lefkowiz *et al*., "Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)", *Nucleic Acids Res.*, vol. 46, no. D1, pp. D708–D717, Jan. 2018, doi: 10.1093/NAR/GKX932.
[2] "Coronavirus disease (COVID-19)." https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed Jul. 02, 2020).
[3] "Covid-19 PCR test: how does it work? are there any alternatives? | Auxologico." https://www.auxologico.com/covid-19-pcr-test-how-does-it-work-are-there-any-alternatives (accessed Oct. 29, 2021).
[4] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network", *Appl. Intell. 2020 512*, vol. 51, no. 2, pp. 854–864, Sep. 2020, doi: 10.1007/S10489-020-01829-7.
[5] A. Borakati *et al.*, "Diagnostic accuracy of X-ray versus CT in COVID-19: a propensity-matched database study", *BMJ Open*. 2020 Nov 6;10(11):e042946. doi: 10.1136/bmjopen-2020-042946. PMID: 33158840; PMCID: PMC7650091.
[6] C. Schaefer-Prokop, M. Prokop, "Chest Radiography in COVID-19: No Role in Asymptomatic and Oligosymptomatic Disease" *Radiology*. 2021 Mar;298(3):E156-E157. doi: 10.1148/radiol.2020204038. Epub 2020 Dec 8. PMID: 33290177; PMCID: PMC7734840.
[7] D. Dong *et al.*, "The Role of Imaging in the Detection and Management of COVID-19: A Review,"*IEEE Rev. Biomed. Eng.*, vol. 14, pp. 16–29, 2021, doi: 10.1109/RBME.2020.2990959.
[8] A.M. Ismael, A. Sengur, "Deep learning approaches for COVID-19 detection based on chest X-ray images", *Expert Systems with Applications*, vol.164, 2021, https://doi.org/10.1016/j.eswa.2020.114054.
[9] I. Katsamenis, E. Protopapadakis, A. Voulodimos, A. Doulamis, and N. Doulamis, "Transfer Learning for COVID-19 Pneumonia Detection and Classification in Chest X-ray Images", doi: 10.1101/2020.12.14.20248158.
[10] A. Shazia, T.Z. Xuan, J.H. Chuah *et al.*, "A comparative study of multiple neural network for detection of COVID-19 on chest X-ray", *EURASIP J. Adv. Signal Process*. 2021, 50 (2021). https://doi.org/10.1186/s13634-021-00755-1.
[11] S. Vineth Ligi, S.S. Kundu, R. Kumar, R. Narayanamoorthi, K.W. Lai, S. Dhanalakshmi, "Radiological Analysis of COVID-19 Using Computational Intelligence: A Broad Gauge Study.", *J Healthc Eng*. 2022 Feb 23;2022:5998042. doi: 10.1155/2022/5998042. PMID: 35251572; PMCID: PMC8890832.
[12] S.L.W. Ching, H.C. Joon, T.H.T. Clarence Augustine, A. Shazia, A.S. Muhammad, F. Amir, K. Azira, W. L. Khin, "An Overview of Deep Learning Techniques on Chest X-Ray and CT Scan Identification of COVID-19", *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 5528144, 17 pages, 2021. https://doi.org/10.1155/2021/5528144.
[13] K. S. Krishnan and K. S. Krishnan, "Vision Transformer based COVID-19 Detection using Chest X-rays," *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, 2021, pp. 644-648, doi: 10.1109/ISPCC53510.2021.9609375.
[14] D. Shome, T. Kar, S.N. Mohanty, P. Tiwari, K. Muhammad, A. AlTameem, Y. Zhang, A.K.J. Saudagar, "COVID-Transformer: Interpretable COVID-19 Detection Using Vision Transformer for Healthcare.", *Int J Environ Res Public Health*. 2021 Oct 21;18(21):11086. doi: 10.3390/ijerph182111086. PMID: 34769600; PMCID: PMC8583247.
[15] A.K. Mondal, A. Bhattacharjee, P. Singla, A.P. Prathosh, "xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography.", *IEEE J Transl Eng Health Med*. 2021 Dec 8;10:1100110. doi: 10.1109/JTEHM.2021.3134096. PMID: 34956741; PMCID: PMC8691725.
[16] S. Park, G. Kim, Y. Oh, J.B. Seo, S.M. Lee, J.H. Kim, S. Moon, J.K. Lim, J.C. Ye, "Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification.", *Med Image Anal*. 2022 Jan;75:102299. doi: 10.1016/j.media.2021.102299. Epub 2021 Nov 4. PMID: 34814058; PMCID: PMC8566090.
[17] Y.E. Almalki, A. Qayyum, M. Irfan, N. Haider, A. Glowacz, F.M. Alshehri, S.K. Alduraibi, K. Alshamrani, M.A. Alkhalik Basha, A. Alduraibi, M.K. Saeed, S. Rahman, "A Novel Method for COVID-19 Diagnosis Using Artificial Intelligence in Chest X-ray Images." *Healthcare* (Basel). 2021 Apr 29;9(5):522. doi: 10.3390/healthcare9050522. PMID: 33946809; PMCID: PMC8145061.
[18] M. E. H. Chowdhury *et al.*, "Can AI Help in Screening Viral and COVID-19 Pneumonia?,"*IEEE Access*, vol. 8, pp. 132665–132676, 2020, doi: 10.1109/ACCESS.2020.3010287.
[19] T. Rahman *et al.*, "Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection using Chest X-rays Images", *Comput. Biol. Med.*, p. 104319, Mar. 2021, doi: 10.1016/j.compbiomed.2021.104319.
[20] A. M. Freeman and J. Townes R. Leigh, "Viral Pneumonia", *Encycl. Respir. Med. Four-Volume Set*, pp. 456–466, Jul. 2021.
[21] Y. LeCun *et al.*, "Backpropagation Applied to Handwritten Zip Code Recognition", *Neural Computation*, vol. 1, no.4, pp. 541–551, Dec. 1989, doi: 10.1162/neco.1989.1.4.541.
[22] A. Vaswani *et al.*, "Attention Is All You Need", 2017, arXiv:1706.03762.
[23] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", 2020, arXiv:2010.11929.

[24]C. Szegedy *et al*., "Going Deeper With Convolutions", *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.
[25]F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions", 2016, arXiv:1610.02357.
[26]K. He *et al.*, "Deep Residual Learning for Image Recognition", 2015, arXiv:1512.03385.
[27]J. Devlin *et al.*, "BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding", 2019, arXiv:1810.04805v2.
[28]D. Mitchell, A. Netravali, "Reconstruction Filters in Computer-Graphics", *Proceedings of the 15th annual conference on computer graphics and interactive techniques (SIGGRAPH '88)*, pp. 221–228, Jun. 1988, doi:10.1145/378456.378514.
[29]A. Hassani *et al*,, "Escaping the Big Data Paradigm with Compact Transformers", 2021, arXiv:2104.05704.

SERGIO CANNATA (SM) was born in Italy in 1985. He earned a Bachelor Degree in Electronic Engineering at Università degli Studi di Catania, Italy, in 2015, and a Master's Degree in Mechatronic Engineering in 2017 at Politecnico di Torino, Italy. After working as a research assistant within the Department of Control and Computer Engineering at Politecnico di Torino, in 2019 he joined the PhD program in Electrical, Electronics and Communication Engineering within the Department of Electronics and Communication Engineering at the same institution, in which he is currently carrying his research activity about smart sensors for IoT and deep learning for biomedical applications.



ANNUNZIATA PAVIGLIANITI (M) received the Master degree (summa cum laude) in Electronic Engineering from Università degli Studi Mediterranea of Reggio Calabria, in 2018 and earned a Ph.D. in Metrology at the Politecnico of Turin on the topic "Deep learning algorithms applied in signal processing" in 2022. Her current research interests include neural networks, data analysis, pattern recognition, automatic defect detection systems, signal processing and biomedical applications. She is an IEEE member and has been active as treasurer of Politecnico di Torino IEEE Women in Engineering Affinity Group.

EROS PASERO (M) is Professor of Electronic System Engineering in the Department of Electronics and Telecommunications and the head of the Neuronica Lab at the Politecnico of Turin. He is now the President of SIREN, the Italian Society for Neural Networks and the General Chairman of WIRN2017, the international Italian workshop for artificial neural networks. Prof. Pasero interests lie in Artificial Neural Networks and Electronic Sensors. Hardware neurons and synapses are studied for neuromorphic approaches; neural software applications are applied to real life proof of concepts. Innovative wired and wireless sensors are also developed for biomedical, environmental and automotive applications. Data coming from sensors are post processed by means of artificial neural networks. He is an IEEE Member and has been selected as a Distinguished Lecturer for the IEEE Instrumentation&Measurement Society for the 2021-2024 program as Telemedicine and Artificial Intelligence expert.

GIANSALVO CIRRINCIONE (F) received the M.S. degree in electrical engineering from the Politecnico di Torino, Italy, in 1991, and the Ph.D. degree (with the congratulations of the jury) from the Laboratoire d'Informatique et Signaux de l'Institut National Polytechnique de Grenoble, Grenoble, France, in 1998. He was a Postdoctoral Scholar with the Department of Signals, Identification, System Theory and Automation (SISTA), Leuven University, Leuven, Belgium, in 1999. Since 2000, he has been an Assistant Professor with the Department of Electrical Engineering, University of Picardie Jules Verne, Amiens, France. He is currently an Adjunct Associate Professor with The University of the South Pacific. His current research interests include neural networks, data analysis, computer vision, brain models, and system identification.

MAURIZIO CIRRINCIONE (F) received the Laurea degree in electrical engineering from the Polytechnic University of Turin, Turin, Italy, in 1991 and the Ph.D. degree in electrical engineering from the University of Palermo, Palermo, Italy, in 1996.,From 1996 to 2005, he was a Researcher with the Section of Palermo, Institute for Studies on Intelligent Systems for Automation (ISSIA) - National Research Council (CNR), Palermo, Italy. In 2005, he joined the University of Technology of Belfort-Montbeliard, Belfort, France, as a Full Professor. He is currently the Head of the "School of Engineering and Physics," University of the South Pacific, Suva, Fiji. His current research interests include neural networks for modeling and control, system identification, intelligent control, power electronics, renewable energy systems, and electrical machines and drives. Dr. Cirrincione was awarded the 1997 "E.R.Caianiello" prize for the best Italian Ph.D. thesis on neural networks.