Fall 11-6-2022

Interpretable Deep Learning for Discriminating Pneumonia from Lung Ultrasounds

Mohamed Abdel-Basset
The American University in Cairo (AUC)

Hossam Hawash

Khalid Abdulaziz Alnowibet

Ali Wagdy Mohamed
The American University in Cairo (AUC), aliwagdy@aucegypt.edu

Recommended Citation

Abdel-Basset, M.; Hawash, H.; Alnowibet, K.A.; Mohamed, A.W.; Sallam, K.M. Interpretable Deep Learning for Discriminating Pneumonia from Lung Ultrasounds. Mathematics 2022, 10, 4153. https://doi.org/10.3390/math10214153

mathematics

MDPI

Article

Interpretable Deep Learning for Discriminating Pneumonia from Lung Ultrasounds

Mohamed Abdel-Basset [1], Hossam Hawash [1], Khalid Abdulaziz Alnowibet [2], Ali Wagdy Mohamed [3,4,*] and Karam M. Sallam [5]

[1] Faculty of Computers and Informatics, Zagazig University, Zagazig 44519, Egypt
[2] Statistics and Operations Research Department, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia
[3] Operations Research Department, Faculty of Graduate Studies for Statistical Research, Cairo University, Giza 12613, Egypt
[4] Department of Mathematics and Actuarial Science School of Sciences Engineering, The American University in Cairo, Cairo 11835, Egypt
[5] School of IT and Systems, University of Canberra, Canberra, ACT 2601, Australia
* Correspondence: aliwagdy@staff.cu.edu.eg

Abstract: Lung ultrasound images have shown great promise to be an operative point-of-care test for the diagnosis of COVID-19 because of the ease of procedure with negligible individual protection equipment, together with relaxed disinfection. Deep learning (DL) is a robust tool for modeling infection patterns from medical images; however, the existing COVID-19 detection models are complex and thereby are hard to deploy in frequently used mobile platforms in point-of-care testing. Moreover, most of the COVID-19 detection models in the existing literature on DL are implemented as a black box, hence, they are hard to be interpreted or trusted by the healthcare community. This paper presents a novel interpretable DL framework discriminating COVID-19 infection from other cases of pneumonia and normal cases using ultrasound data of patients. In the proposed framework, novel transformer modules are introduced to model the pathological information from ultrasound frames using an improved window-based multi-head self-attention layer. A convolutional patching module is introduced to transform input frames into latent space rather than partitioning input into patches. A weighted pooling module is presented to score the embeddings of the disease representations obtained from the transformer modules to attend to information that is most valuable for the screening decision. Experimental analysis of the public three-class lung ultrasound dataset (PCUS dataset) demonstrates the discriminative power (Accuracy: 93.4%, F1-score: 93.1%, AUC: 97.5%) of the proposed solution overcoming the competing approaches while maintaining low complexity. The proposed model obtained very promising results in comparison with the rival models. More importantly, it gives explainable outputs therefore, it can serve as a candidate tool for empowering the sustainable diagnosis of COVID-19-like diseases in smart healthcare.

Keywords: explainable artificial intelligence; interpretable deep learning; convolutional networks; vision transformers; COVID-19; ultrasound image

MSC: 68T01; 68T05; 68T07; 68T09; 68T20; 68T30

Citation: Abdel-Basset, M.; Hawash, H.; Alnowibet, K.A.; Mohamed, A.W.; Sallam, K.M. Interpretable Deep Learning for Discriminating Pneumonia from Lung Ultrasounds. *Mathematics* **2022**, *10*, 4153. https://doi.org/10.3390/math10214153

Academic Editors: Adrian Sergiu Darabant and Diana-Laura Borza

Received: 6 October 2022
Accepted: 3 November 2022
Published: 6 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

By the beginning of 2020, human beings were conquered by the SARS-CoV-2 virus. That virus caused a pandemic infectious disease called COVID-19. The outbreak of COVID-19 has had a catastrophic impact on global health infrastructure, leading to millions of infected cases and thousands of deaths [1]. In clinical practice, real-time reverse transcriptase-polymerase chain reaction (RT-PCR) is used as a standard test to detect COVID-19 infection [2]. However, research studies demonstrated that RT-PCR suffers from

high false-positive rates because the clinical practice is strictly impacted by a variety of aspects such as appropriateness of specimen, phase of infection, specimen categories, and specimen conduct, containing sample acquisition time from the inception of COVID-19. Moreover, the fast proliferation of COVID-19 has led to a deficiency of RT-PCR test kits for the discovery of COVID-19. Another obvious shortcoming of this test involves direct interaction between medical staff and patients. This, in turn, made the medical staff more prone to infection, which caused a high death rate in the healthcare community. As a remedy to the above issues, doctors and scholars showed that medical images can be used as an effective tool to detect the presence of COVID-19. X-rays and computed tomography (CT) are broadly used in this respect [3]. Accordingly, the research moved from manual diagnosis of COVID-19 toward computer-aided diagnosis.

Ultrasound imaging is a non-invasive method that has already begun to replace X-rays in pulmonary disease. There has been a surge in demand for point-of-care ultrasound, which is supported by clinical facts and research findings. The advantages of point-of-care ultrasonography (POCUS) include its low cost, ease of mobility, and bedside convenience for patient safety. It is presently underutilized due to a lack of training options and an appreciation of the research supporting this approach. It was discovered that it enhances conventional diagnostic procedures and the technology is rapidly evolving. For wider use, it was also suggested that POCUS be introduced into urgent and core medicine curricula [4]. The latest epidemic of the COVID-19 pandemic forced the healthcare community to utilize ultrasound imaging in emergency departments [5–7]. Discoveries indicated that ultrasound imaging could help in both identifying COVID-19 cases and following up on their states during the hospitalization phase. Nevertheless, lung ultrasound offers just local pathological information on the status of the lung area. Thus, it is critical to thoroughly specify the necessary volume and distribution of lung regions to be scanned. This way, multi-institutional studies are designed to search for an ideal trade-off between a rapid and precise assessment to be performed [8]. For COVID-19 diagnosis, a twelve-area methodology was proposed to signify an optimum trade-off between precision, speed, and exam complication [9]. When it comes to evaluating patients' states from ultrasound data, both pathological information (i.e., pleural line, consolidations, opacities), and sonographic artifacts (i.e., B-lines and A-lines) come to be significant. Nevertheless, identifying this information and appropriately interpreting it demands extremely experienced doctors. Therefore, to date, lung ultrasound data are not broadly accepted, even though their capacity would be somewhat recommended, especially in the face of urgent requirements occurring in screening patients in the COVID-19 pandemic [10].

Deep learning (DL) as a subfield of machine learning (ML) was demonstrated as a key enabler for almost all medical image analysis tasks and computer-aided diagnosis systems. In this regard, convolutional neural networks (CNNs) are showing great capability for extracting valuable representations from medical images [11]. For COVID-19 diagnosis, CNNs are showing great promise in the detection and segmentation of infection and were demonstrated to be very robust tools for achieving many interesting responsibilities, particularly in the realm of image perception. Given an adequate amount of training samples and enough computing power, complex DL could even surpass doctors' performance on particular diagnosis tasks [12]. Research efforts are developing in the way of applying CNNs to ultrasound data. However, training data are difficult to obtain, which is a common issue in medical image analysis and makes it challenging to effectively train these complicated models. Unfortunately, the majority of DL solutions for COVID-19 diagnosis are designed as a black-box model, which means that the model just provides us with the final decision without justifying the reason behind it [13]. In other words, the doctors are unable to interpret the internal working methodology of the DL model, thereby they cannot trust the diagnosis results obtained from the DL model [14]. Generally speaking, the opaque nature of DL models constrains the ability to integrate them into real-world healthcare applications. These constraints come to be more serious in the case of critical or pandemic diseases.

1.1. Research Gaps

When it comes to the diagnosis of COVID-19 from ultrasound images using DL, many research gaps are encountered, making it challenging to do well in combating COVID-19 either in or after the outbreak. By investigating the recent literature, this study considers the following open gaps:

- **Efficiency:** In clinical practice, the task of detecting COVID-19 from lung ultrasound data (either images or videos) necessitates high experience from doctors. Similarly, DL should be able to effectively model the disease representations from the ultrasound data in such a way that enables discriminating between the manifestations of COVID-19 and other kinds of pneumonia with the lowest possible error rate. The healthcare system does not tolerate any errors, especially in the diagnosis of pandemic diseases, making the accuracy of the model an essential requirement [15];
- **Complexity:** Deep models are demonstrated as robust tools for learning inherent features and diagnostic cues from medical image analysis. These models usually have a complicated building structure so that they can model different representations from large-sized and multi-dimensional training datasets. This complexity, in turn, necessitates the model training to be performed on powerful and computationally efficient machines, making it challenging to effectively detect COVID-19 from lung ultrasound data. Another challenge to be considered in this respect is that the complex models are unable to do well when trained on limited training data, which is a typical scenario when dealing with COVID-19 lung data [16,17];
- **Opaqueness:** As stated above, the deep models are characteristically composed of multiple building blocks and layers with several nonlinear interconnected interactions. Even if one is to inspect all these layers and describe their relations, it is unfeasible to fully comprehend how the neural network came to its decision. Therefore, DL is often considered a 'black box'. To properly understand how the model made its choice, one would need to examine all of these constituting layers and define their relationships. This is practically infeasible, leading the community to declare the DL model as a "black box" model. There is growing concern that these black boxes might exhibit some unobserved bias in making their decisions [13,14,18]. This can have far-reaching repercussions, especially in medical applications. When it comes to COVID-19 detection, the same problem exists; however, the dangerousness of the disease even after the outbreak means that doctors have no chance to trust such opaque models. Marginal mistakes in the diagnostic decision may cause catastrophic consequences. Explainable artificial intelligence (XAI) is an evolving subfield to provide a good explanation if it gives insight into how a neural network came to its decision and/or can make the decision understandable.

1.2. Contributions

In response to the above challenges and gaps, this work presents a novel DL solution that affords and is interpretable and efficient in the detection of COVID-19 from ultrasound data. The main contributions of this work are pointed out as follows:

- First, a lightweight convolutional transformer network for flexible and robust modeling of COVID-19 from ultrasound data is designed, while dissipating the nightmare of "data-hungry" models by being able to effectively learn from scratch and attain high screening performance on small-size data.
- Second, two parallel transformer modules are designed with window-based and shifted-window multi-head self-attention layers, respectively, aiming to improve the representational power of the model, while maintaining a few numbers of parameters.
- Third, the convolutional patching module is integrated to empower image tokenization to sustain local spatial representations by encoding relationships between patches.
- Fourth, a weighted pooling module is presented to get rid of the necessity for class tokens and by scoring the sequential embeddings of the disease representations captured by the transformer modules to better relate information across the input frames.

- A gradient activation mapping integrated after-weighted pooling to empower the proposed model visually explains its decided class for a given input frame, which is achieved by highlighting the contribution of different biomarkers in the ultrasound frame.
- Finally, an experimental evaluation of the public lung ultrasound dataset demonstrates the ability of the proposed solution to precisely screen COVID-19, while generating a visual explanation for the generated decisions.

1.3. Organization

The remaining part of this work is systematically organized as follows. Section 2 discusses the literature studies relevant to COVID-19 detection. Section 3 argues the methodology of the proposed solution. In addition, the experimental setting of this study is discussed in Section 4. Then, the results, analysis, and findings are given in Section 5. Finally, Section 6 concludes this work.

2. Related Work

DL was demonstrated to be effective in a variety of imaging tasks spanning semantic segmentation, object detection, etc. Inspired by these achievements, more recently, DL has been progressively applied in medical applications, such as pneumonia detection, localization, and segmentation from chest X-rays. This in turn shows that DL can be used to help and automate preliminary diagnosis, which is of enormous importance to the medical profession.

2.1. Deep Learning for COVID-19 Screening

The literature contains a lot of studies for the screening of COVID-19 from different modalities of medical images, among them the lung ultrasound gains the least research attention despite its demonstrated promise for screening and follow-up diagnosis. For example, Born et al. [19] proposed a convolutional model, called POCOVID-net, which is dedicated to identifying COVID-19, healthy, and bacterial pneumonia from lung ultrasound frames and videos. The POCOVID-net was accompanied by a class activation map (CAM) as an interpretability technique for localizing the spatiotemporal pulmonary manifestations, which are regarded as valuable for human-in-the-loop circumstances in medical studies. Similarly, Diaz-Escobar et al. [20] applied many pre-trained models (i.e., VGG19, InceptionV3, Xception, and ResNet50) to finetune them on lung ultrasound frames to detect COVID-19 and pneumonia patients. Awasthi et al. [21] presented a lightweight convolutional model, termed Mini-COVIDNet [21], for detecting COVID-19 from ultrasound images, where the model can be deployed and used in resource-constrained applications making it ideal for a point-of-care situation. The Mini-COVIDNet was trained to optimize focal loss function to lessen the impact of class imbalance. Moreover, Frank et al. [22] proposed a DL framework that combines domain knowledge into deep networks by feeding anatomical representations and ultrasound artifacts as an extra channel comprising vertical and pleural artifact masks in addition to original lung ultrasonic frames. They claimed that the direct inclusion of this domain knowledge enables the deep networks to achieve different diagnosis tasks using ultrasonic imagery quickly and efficiently. Additionally, the framework was enabled to learn from both convex as well as linear probes and it shows good performance on the COVID-19 severity assessment task, as well as the semantic segmentation model. In addition, Muhammad et al. [17] proposed screening COVID-19 from ultrasound images using a lightweight convolutional model of consisting of five major building convolutional blocks with a small number of trainable parameters. Then, the feature maps from each block are fused to generate a representation vector to be fed into a fully connected layer (FCL), where the final decision is made.

Moreover, Marco et al. [23] presented a DL system for screening COVID-19 from ultrasound data using a pre-trained and residual convolutional model, which is trained (using transfer learning and data augmentation techniques) to quantify the severity of infection as well. In another approach, Xue et al. [24] developed a multimodal approach for assessing the severity of COVID-19 from two types of modality data (ultrasound data and clinical information), where a dual-level supervised multiple-instance learning was applied to combine the zone-associated features and patient-associated representations from heterogeneous training data, hence resulting in discriminatory features. The model aligned the two modalities using a contrastive learning module while maintaining the discriminatory representations of each of them. Furthermore, Roy et al. [15] proposed a spatial transformer network, that concurrently forecasts the degree of severity of COVID-19 in ultrasound frames and localizes pathological artifacts in a weakly supervised manner. The authors also developed a lightweight technique for the efficient aggregation of scores of frames at the video level.

2.2. Explainable Medical Image Analysis

With the increased acceptance of DL solutions in medical image diagnosis, explainability becomes an inevitable requirement to effectively use these solutions in real-world healthcare. To this end, many studies have recently emphasized developing explainable models for COVID-19 screening. For example, Wu et al. [25] proposed a multi-task DL framework that jointly classifies COVID-19 and segment infections from a CT scan with the main aim of using the segmentation results to provide an explanation for the screening decisions. In a similar way, Wang et al. [26] proposed a joint learning framework, called DeepSC-COVID, for the screening and segmentation of COVID-19 lesions from 3D CT scans. In particular, the DeepSC-COVID model is composed of three sub-networks, namely the cross-task feature sub-network, segmentation subnetwork for segmenting 3D COVID-19 lesions, and classification subnetwork for identifying COVID-19, pneumonia, and non-pneumonia cases. The latter one contains a multi-layer visualization method to produce evidential masks that include tiny and imprecise lesions for making the task screening of COVID-19 explainable. During the training of DeepSC-COVID, a task-aware loss was developed based on our visualization method for effective collaboration between classification and segmentation. Though the integration of segmentation and classification tasks in the single model help provide an explainable diagnosis, it makes the model very complex and has a very large number of parameters. In addition, Shi et al. [27] presented an explainable attention transfer model network to automatically screen COVID-19 from chest X-ray and CT scans, which consisted of a teacher model and a student model. The former models the global representation and uses a deformable attention module to distill the infection lesions to intensify the reaction to lesions and restrain noise in unrelated areas with an extended reception field. Next, an image fusion unit was proposed to integrate attention knowledge transmitted from teacher to student with the necessary representations in the original input. The student model was designed to concentrate on sporadically formed lesion areas to learn discriminatory features. In [28], the Gradient-weighted CAM (Grad-CAM) algorithm was employed for debugging the convolutional models to provide explainability of its classification decision in chest X-rays.

3. Research Methodology

This section introduces and discusses the proposed framework for screening COVID-19 from lung ultrasound data. To obtain a better interpretation of the proposed framework, an illustration of its structural design is presented in Figure 1. As observed, the proposed framework consists of six main building modules and we are going to dive into the details of each of them in the next subsections.

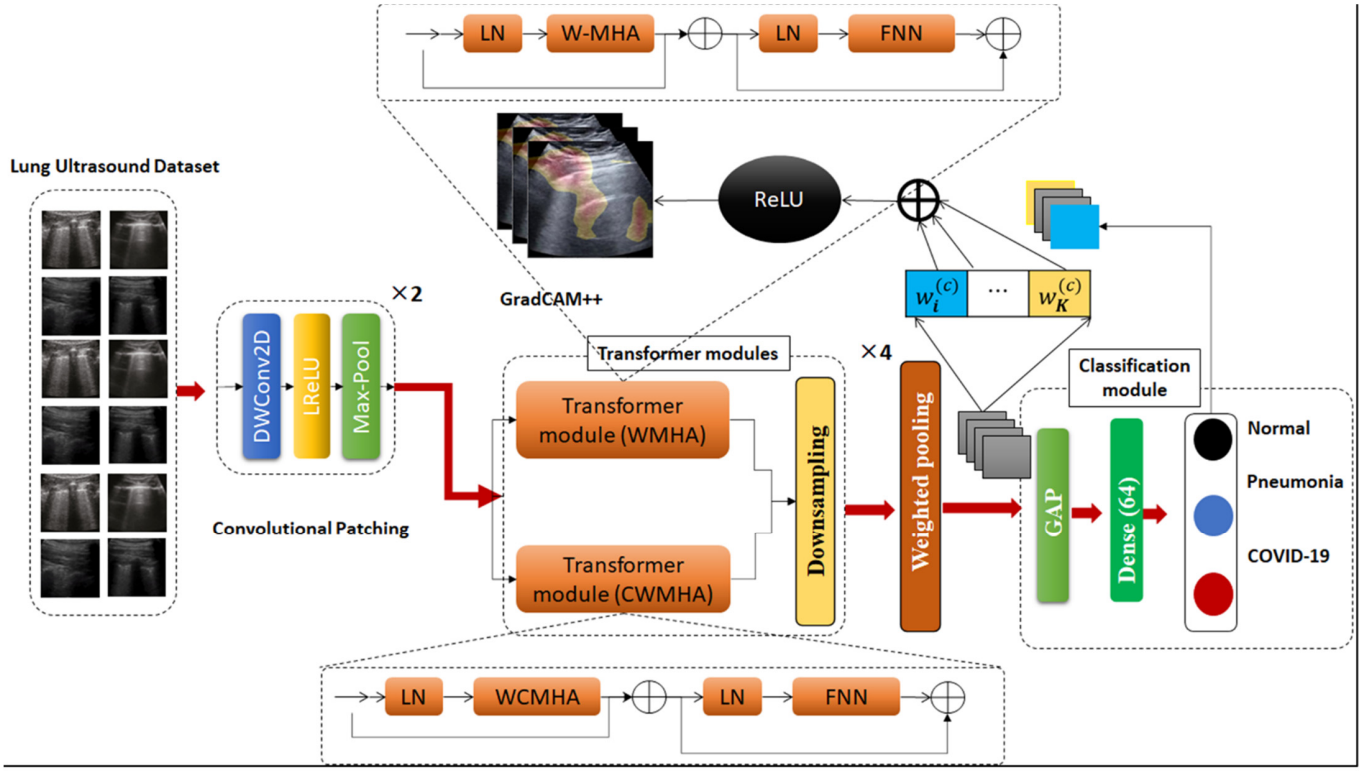

Figure 1. Illustration of the proposed explainable convolutional transformer network for screening pneumonia from lung ultrasound data.

3.1. Convolutional Patching

The conventual language Transformer was designed to accept input in form of a one-dimensional sequence of token embeddings. This seems inappropriate when dealing with 2D ultrasound frames, therefore the input frame $x \in \mathbb{R}^{H\times W\times C}$ is reshaped into a sequence of compressed 2D patches $x_p \in \mathbb{R}^{N\times((p\times p)\cdot C)}$, whereby $H \times W$ represents the spatial dimensions of the original frame, C represents the number of channels, $p \times p$ denotes the spatial dimensions of each patch, and N denotes the number of patches (it is calculated as follows $N = \frac{HW}{P^2}$) , denoting the length of the input sequence of Transformer modules. Each of these modules utilizes a fixed size latent vector D, hence, the generated frame patches are mapped to d dimensions through learnable linear projection, as shown in Equation (1).

$$z_0 = \left[x_{class}, x_p^1 E,\ x_p^2 E,\ \cdots x_p^N E\right] + E_{pos}, E \in \mathbb{R}^{N\times((p\times p)\cdot C)}, E_{pos} \in \mathbb{R}^{(N+1)\times(d)} \quad (1)$$

where the generated output z_0 of this projection can be referred to as patch or token embeddings. The patch embeddings were supplied by one-dimensional position embeddings $E_{pos} \in \mathbb{R}^{(N+1)\times(D)}$ to preserve spatial pathological information in ultrasound frames.

To establish an inductive bias in the proposed model, the standard image patching, as well as token embedding, are replaced with a straightforward convolutional module. The design of a convolutional module consists of a depth-wise convolutional layer ($DWConv2D$) activated by LeakyReLU ($LReLU$) function and followed by a max-pooling layer. In doing so, the above formula can be formulated as:

$$z_0 = MaxPool(LReLU(DWConv2D(x))) \quad (2)$$

where the $DWConv2D$ layer contains a number of d filters, equal to the embedding dimension in Equation (1). Two convolutional modules are stacked to generate convolutional patching. This convolutional patching makes the model more flexible and simpler than the

standard vision transformer. In particular, the convolutional modules are introduced to embed the input ultrasound frames into a latent representation, which is more efficient for modeling pathological information in subsequent layers. Variations in the size of image size do not have an effect on the number of parameters but impact the length of the sequence and consequently the required computing. Even though the standard patching operation necessitates that the dimensions of input frames be dividable by the size of the patch. The convolutional modules enabled the model to accept input of various sizes of data with no requirement for clipping or padding as it alleviates the necessity of uniformly partitioning an image into patches. Another advantage of the proposed model lies in the fact that convolution and max pool layer could be overlapping, enabling the sequence length to become increased but conversely, improving screening performance by infusing inductive bias. Obtaining these convolutional patches enables the model to hold the local spatial pathological information eliminating the need for positional embedding as it achieves a very decent performance.

3.2. Transformer Modules

The project sequence of embeddings z_0 are then passed to a stacked transformer module. Each of these modules consists of alternating layers of multiheaded self-attention (MHA) and Feed Forward Network (FNN) blocks.

The traditional self-attention (SA) [29] attention layer is commonly used to calculate the attention score for each head on a global receptive field, leading to quadratic computations in terms of the number of tokens, which makes it inappropriate for ultrasound frames/videos that require a huge set of tokens for modeling pathological information. To address that, window-based or local SA is adopted to calculate SA for local windows, where windows are disposed to uniformly divide the image in a non-overlapping way. Given that the window includes $M \times M$ non-overlapping patches, SA is computed locally within the window.

For each instance in patch embedding $z \in \mathbb{R}^{N \times D}$, a weighted summation is calculated for all values v in the sequence. Then, the attention scores are calculated according to pairwise correspondence between two embedding elements and the corresponding query q_i and key k_j representations.

$$[q, k, v] = zU_{qkv},\ U_{qkv} \in \mathbb{R}^{M^2 \times d_h} \tag{3}$$

$$a = softmax\left(\frac{qk^T}{\sqrt{d_h}} + B\right),\ a \in \mathbb{R}^{N \times N} \tag{4}$$

$$H(z) = a \cdot v \tag{5}$$

where H represents the attention head; $q, k, v\ \in \mathbb{R}^{M^2 \times d_h}$ represent the query, key, and value matrices; d_h denotes the query/key dimension, M^2 represents the number of window patches and $B \in \mathbb{R}^{M^2 \times M^2}$ represents the relative position bias [16].

The MHA is an expansion of the above calculations by calculating the SA for multiple heads concurrently and then concatenating the output of each of them, and later projecting this concatenation in FFN.

$$MHA(z) = concat\left[H^0, H^1, \cdots, H^{h-1}\right] \tag{6}$$

Layer norm (LN) is applied at the beginning of each module, while the residual connection is applied to the MHA and FNN layers in each module.

By calculating window-based SA, the computational complexity of a global MHA and a window-based MHA (WMHA) over an image containing $h \times w$ patches are formulated as follows:

$$\Omega(MHA) = 4hwC^2 + 2(h \times w)^2 C \tag{7}$$

$$\Omega(WMHA) = 4hwC^2 + 2M^2(h \times w)C \tag{8}$$

where the first formula is quadratic with respect o the number of patches $h \times w$, and the other formula is linear when M is constant. Therefore, the global SA calculation is mostly unreasonable for a large number of patches, while the local SA is usable. However, local SA does not have connections among windows, limiting its representation power [16]. As a remedy, the network requires the application of cross-window connectivity while retaining the effective calculation of non-overlapping windows. This is achieved by designing two parallel transformer modules, one uses WMHA and the other use cross-window-based MHA (CWMHA). Mathematically speaking, the flow of information in the first transformer modules could be formulated as follows:

$$\widetilde{z}_l^I = WMHA\Big(LayerNormalize\Big(z_{l-1}^I\Big)\Big) + z_{l-1}^I, \qquad l = 1 \cdots L \tag{9}$$

$$z_l^I = FFN\Big(LayerNormalize\Big(\widetilde{z}_{l-1}^I\Big)\Big) + \widetilde{z}_{l-1}^I, \qquad l = 1 \cdots L \tag{10}$$

In a similar way, the flow of information in the other transformer module is calculated as below:

$$\widetilde{z}_l^{II} = CWMHA\Big(LayerNormalize\Big(z_{l-1}^{II}\Big)\Big) + z_{l-1}^{II}, \qquad l = 1 \cdots L \tag{11}$$

$$z_l^I = FFN\Big(LayerNormalize\Big(\widetilde{z}_{l-1}^{II}\Big)\Big) + \widetilde{z}_{l-1}^{II}, \qquad l = 1 \cdots L \tag{12}$$

Like language models [30], a trainable embedding was prepended to the sequence of embedded patches ($z_0 = x_{class}$), whose status at the output of the transformer modules (z_L) serves as the frame/video representation y, as shown in Equation (4).

$$y = LayerNormalize(z_L) \tag{13}$$

During either fine-tuning or pre-training, the classification module accompanies the z_L. The classification module is implemented with FNN composed of a single hidden layer.

3.3. Down-Sampling Module

In order to generate a hierarchical representation, the number of tokens is reduced by down-sampling modules, introduced to reduce the number of tokens as the depth of the network increases. In the earlier patch-merging module, the features of each set of $g \times g$ neighboring patches are concatenated and then passed to a convolutional 1×1 layer. This way, the number of tokens is decreased by a multiple of $g \times g$ (i.e., down-sampling), and the dimension of the output is turned into $2C$. The same process is applied after each transformed module and by the end, the stacked modules jointly generate a hierarchical representation with the same dimensions as the feature map generated from standard convolutional networks.

3.4. Weighted Pooling

To encode the sequential outcomes into a singular class index, rather than applying a class token (as commonly performed in common transformer networks), the proposed model presents a weighted pooling layer. Simply, the outputs of the transformer modules are pooled over the whole sequence of data because they include appropriate representation across various sections of the ultrasound frames. Mathematically speaking, the sequential pooling operation can be declared as the mapping function $\mathcal{M} : \mathbb{R}^{b \times n \times d} \rightarrow \mathbb{R}^{b \times d}$ given as:

$$x_L = \mathrm{f}(x_0) \in \mathbb{R}^{b \times n \times d} \tag{14}$$

where x_L represents the feature maps generated from L-th of the transformer module and b, n, d denote the size of the mini-batch, length of the sequence, and embedding

dimension. Hence, x_L maps are passed to a linear layer with $SoftMax$ activation to generate the following:

$$x'_L = softmax\left(g(x_L)^T\right) \in \mathbb{R}^{b\times 1\times n} \tag{15}$$

Following this, the obtained probability scores are used to calculate the pooled output as follows:

$$z = x'_L \times x_L = softmax\left(g(x_L)^T\right) \times x_L \tag{16}$$

The design of weighted pooling enables the model to score the sequential embeddings of latent representations generated from transformer modules and robustly relate data throughout the input data. This behavior is similar to the process of attention to sequential data. This pooling layer can be implemented in either trainable or non-trainable manner; however, the latter case is more efficient for the reason that every embedded patch includes a different quantity of entropy. Accordingly, the network is capable of assigning higher scores to input patches that comprise more pathological information valuable for the screening of pneumonia or of COVID-19. Furthermore, the weighted pooling enables the model to improve by using information from heterogeneous sources.

3.5. Classification Module

Given the output of the weighted pooling, the model calculates the final screening decision in the classification module consisting of two fully connected layers, the first with 64 neurons and containing three units corresponding to three classes, i.e., COVID-19, pneumonia, and normal. The last layer is normally activated with SoftMax activation. The parameters of the model-optimized categorical focal loss (SFL) function [31] help lessen the impact of class imbalance on the final classification performance.

$$CFL = -\sum_{i=1}^{n=3} \alpha_i (i - p_i)^{\gamma}\, log\, p_i \tag{17}$$

The hyperparameter γ enables fine-tuning of the weight of various samples. If $\gamma = 0$, this signifies the categorical cross-entropy. Given a higher value of γ, a small set of simply categorized ultrasound frames participate in calculating the training loss, while frames belonging to the minority class are assigned a higher weight. α_i represents a balance factor.

3.6. Explainability Module

When it comes to explaining the DL model, CAM is a popular approach for generating an activation map for every input sample signifying pixel-wise donation to the decision, or disease type in our case. The discriminative ability of CAM stems from the fact that it generates class-aware activation maps offering more analysis at the class level. However, CAM suffers from primary limitations that it adds to the Softmax layer, hence, performing re-training, which might cause the performance to degrade [32]. Grad-CAM++ [18] is integrated to calculate activation maps with no modification to the DL model. They also require a weight matrix to bring together feature maps. This could be accomplished by initially estimating the gradient of a given class with respect to each feature map and then applying global average pooling on the derivatives to obtain a weight matrix. This way, Grad-CAM++ prevents adding up additional layers, thereby eliminating performance degradation and re-training problems. The calculation of weights $w_k^{(c)}$ in Grad-CAM++ can be formulated as follows:

$$w_k^{(c)} = \sum_{i=1}^{H}\sum_{j=1}^{W} \alpha_k^{(c)}(i,\, j).ReLU\left(\frac{\partial Y^{(c)}}{\partial A_k(i,\, j)}\right), \tag{18}$$

where $Y^{(c)}$ represent the model estimated probability for class c immediately prior to the SoftMax layer and $\alpha_k^{(c)}(i,\ j)$ represent weighting factors for class-specific pixel-wise gradients calculated as follows:

$$\alpha_k^{(c)}(i,\ j) = \frac{1}{\sum_{i,j} \frac{\partial Y^{(c)}}{\partial A_k(i,\ j)}} = \frac{\frac{\partial^2 Y^{(c)}}{(\partial A_k(i,\ j))^2}}{2.\frac{\partial^2 Y^{(c)}}{(\partial A_k(i,\ j))^2} + \sum_{a,b} A_k(a,b).\frac{\partial^3 Y^{(c)}}{(\partial A_k(i,\ j))^3}} \tag{19}$$

where (i,j) and (a,b) represent the iterators over the same activation map A_k and were applied to evade disorientation. The final saliency map can be calculated as follows:

$$L^{(c)}_{Grad-CAM++}(x,\ y) = ReLU\left(\sum_k w_k^{(c)} A_k(x,y)\right) \tag{20}$$

where $A_k(x,y)$ is the activation of node k in the intended network layer at the location (x,y).

4. Experimental Design

This section defines the design settings of proof-of-concept experiments in terms of implementation setup, evaluation metrics, and the dataset adopted for training and evaluations.

4.1. Implementation Setup

To set up the experimentations in this work, a TensorFlow 2.6 running over Python 3.8 virtual environment is employed for implementing the deep models. All experiments are performed on a Dell workstation armed with RAM (256 GB) and CPU (Intel ® Xe®(R) CPU E5-2670 0@ 2.60 GHz). The training of that models was accelerated by NVIDIA Quadro graphical processing unit (GPU). All the experiments are performed using five-fold cross-validations strategies.

4.2. Evaluation Metrics

For evaluating the detection performance of the proposed method and the competing ones, a set of popular multi-class classification metrics calculated as a function of false positive (FB), false negative (FN), true negative (TN), and true positive (TP) samples are opted for and defined as follows:

$$Accuracy\ (A) = \frac{TP + TN}{TP + TN + FP + FN} \times 100, \tag{21}$$

$$Sentivity\ (Se) = \frac{TP}{TP + FN} \times 100, \tag{22}$$

$$Specificity\ (Sp) = \frac{TN}{TN + FP} \times 100, \tag{23}$$

$$F1 - score\ (F1) = \frac{2TP}{2TP + FP + FN} \times 100, \tag{24}$$

Beyond the above metrics, the Area Under the Curve (AUC) is adopted to assess the detection capability of the model.

4.3. Data Description

To train and evaluate the proposed method, a public and open-source LUS dataset is used, which is known as the POCUS dataset. The dataset consists of image and video samples belonging to three classes of infection, namely COVID-19, viral pneumonia, and bacterial pneumonia, in addition to samples from healthy individuals. The dataset contains a total of 261 recordings (202 videos + 59 images) captured from a total of 216 patients with either linear or convex probes. The distribution of samples across different classes is given

in more detail in Table 1. Linear probes have high frequency, leading to a superior resolution that enables improved investigation of irregularities near the pleural line [22]. However, the linear probe penetrates the lung tissue less than the convex probe, which could make it hard to tell the difference between B-line artifacts (a major lung artifact) and hidden tissue. The images and videos in the POCUS dataset were aggregated from a variety of sources, such as clinical data obtained from academic ultrasound courses, hospitals, scientific publications, public medical repositories, and health-tech corporations. The complete details of different sources of data in the POCUS dataset can be found in reference [19]. The COVID-19 cases were confirmed by RT-PCR. The dataset is supplemented by a comprehensive metadata file encapsulating the anonymized patient identifier, source URL, source identifier, sex, age, symptoms, pathological manifestations, video frame rate, image resolution, and the total of frames per video. The length and type of videos are a varied (160 ± 144 frames) dataset, whereas they have a frame rate of 25 ± 10 Hz. Outstandingly, all samples in the dataset were reported to be revised and confirmed by one medical expert with more than 10 years of clinical experience and an academic instructor. Figure 2 shows some examples of 2D ultrasound frames for COVID-19, pneumonia, and healthy patients.

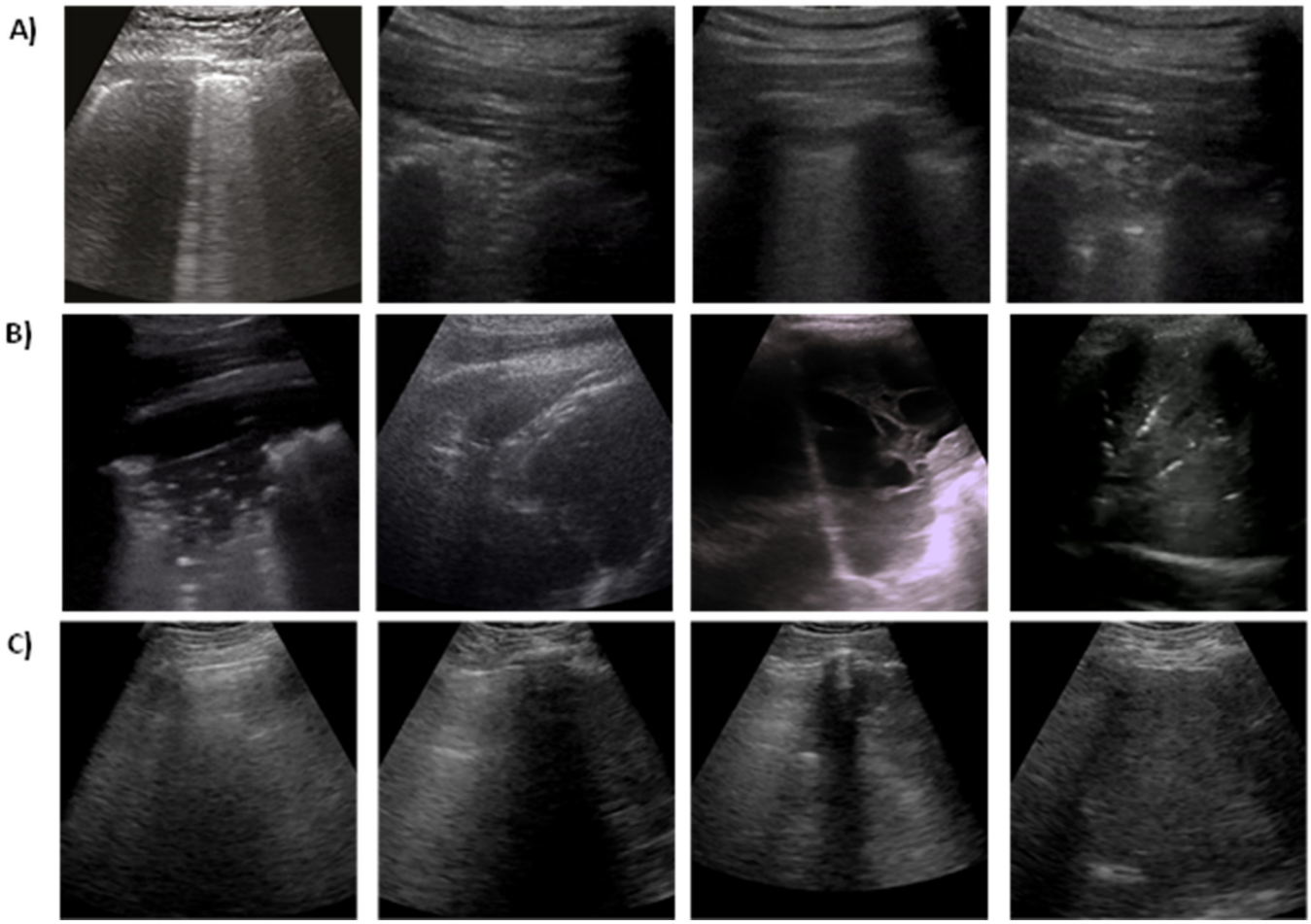

Figure 2. Examples of lung ultrasound images for different pulmonary diseases: (**A**) COVID-19, (**B**) Pneumonia, (**C**) normal case.

Table 1. The class distribution of the POCUS dataset.

		COVID-19	Bacterial Pneumonia	Viral Pneumonia	Healthy	Total
CONVEX	Videos	64	49	3	66	182
	Images	18	20	/	15	53
LINEAR	Videos	6	2	3	9	20
	Images	4	2	/	/	6
Total		92	73	6	90	261

4.4. Data Preparation

As the usual step in developing a DL solution, data need to be pre-processed before going to the training stage. In this regard, the convex ultrasound probes are used for training models in all experimentations. Owing to the small number of samples belonging to the viral pneumonia class (3 convex videos), the data are pre-processed by eliminating the data of that class and the training is performed using only the other three classes. Moreover, all convex ultrasound samples (179 videos and 53 images) were physically pre-processed by dividing the videos into separate images at a 3 Hz frame rate (i.e., maximum of 30 frames per video) resulting in a dataset comprising a total of 1204 COVID-19 images, 704 bacterial pneumonia images, and 1326 images of normal cases. Moreover, the generated image samples are cropped into a quadratic window, eliminating artifacts, ration bars, and text, and then they are resized into 224 × 224 pixels. Different from hold-out testing data, all the presented stated results in this work are attained from five-fold cross-validation stratified by the count of examples in each class. The image samples were divided at the patient level; therefore, it is guaranteed that the frames of one video are existing only per one-fold and that the number of videos per class is almost the same for all folds. All models were trained to classify images as COVID-19, pneumonia, and non-infected. Furthermore, the data were augmented by applying image flipping, rotations $\left[-10^{\circ}, 10^{\circ}\right]$ and translations (up to 10%, which in turn differentiate the data and help avoid overfitting).

5. Results and Analysis

5.1. Comparative Analysis

To evaluate the competitiveness of the proposed model, detection performance is fairly compared against the cutting-edge COVID-19 models, namely CNN [17], POCOVID-Net [19], Mini-COVIDNet [21], Residual Net [23], and Inception [20]. To assure that comparison is very fair, the results of the competing methods are reproduced in the same experimental environment, settings, and training data. During the experiments, an overfitting issue was observed in the case of Mini-COVIDNet, therefore we had to apply some regularization methods to assure the results remain real and representative. Table 2 shows the results obtained from comparative experiments by calculating the mean and standard deviation (std) over different validation folds. Moreover, the comparison also includes the number of parameters of each model to give an intuitive understanding of the model's complexity. It is notable that the pre-trained Residual Net [23] shows the lowest screening performance (accuracy: 71.5%) despite its large number of parameters. Moreover, POCOVID-Net [19] and Mini-COVIDNet [21] show relative improvement in COVID-19 screening performance with an accuracy of 82.1% and 82.7%, respectively. However, they exhibit a large number of parameters. Notably, CNN [17] achieved the most competitive performance across all evaluation metrics, while maintaining a small number of parameters compared with the above methods. More significantly, the proposed model demonstrated robust detection performance across all performance metrics (accuracy: 93.4% F1-score: 93.1%, AUC: 97.5%), overcoming the competing methods with large margins. As observed, the number of parameters of the proposed model is surprisingly smaller than all the competing methods, which can be attributed to the elegant design of building blocks,

i.e., convolutional patching and window-based attention. The lightweight nature of the proposed model makes it a time- and space-efficient solution that can be easily integrated into the real-world healthcare system.

Table 2. Comparison of the 5-Fold cross-validation performance (mean ± std) of the proposed methods and competing models for COVID-19 screening from a frame-based lung ultrasound dataset. The Precision, Recall, F1-score, and AUC metrics are reported for each class.

METHOD	ACCURACY (%)	NO. PARAMS	CLASS	PRECISION (%)	RECALL (%)	F1-SCORE (%)	AUC (%)
CNN [17]	90.3 ± 5.13	389,540	COVID-19	93.8 ± 4.75	91.9 ± 2.00	92.8 ±2.8	96.8 ± 3.68
			Pneumonia	95.1 ± 2.90	96.2 ± 1.73	95.6 ± 2.2	97.2 ± 5.59
			Normal	80.1 ± 5.87	75.6 ± 5.2	77.8 ± 5.5	83.1 ± 10.5
POCOVID-NET [19]	82.1 ± 11.6	14,747,971	COVID-19	84.6 ± 6.80	88.1 ± 10.8	86.3 ± 8.3	95.3 ± 1.19
			Pneumonia	93.9 ± 4.20	91.5 ± 2.10	92.7 ± 2.8	97.3 ± 1.57
			Normal	56.2 ± 8.22	51.9 ± 2.90	54.0 ± 4.3	68.2 ± 2.74
MINI-COVIDNET [21]	82.7 ± 10.2	3,361,091	COVID-19	81.9 ± 3.92	91.8 ± 9.65	86.6 ± 5.6	95.3 ± 6.54
			Pneumonia	82.4 ± 4.56	90.3 ± 5.32	86.2 ± 4.9	95.7 ± 9.13
			Normal	62.3 ± 9.50	44.7 ± 11.5	52.1 ± 10.4	63.1 ± 7.63
RESIDUAL NET [23]	71.5 ± 13.5	23,851,011	COVID-19	76.4 ± 8.21	85.2 ± 6.26	80.6 ± 7.1	89.2 ± 5.88
			Pneumonia	89.7 ± 9.76	63.1 ± 9.08	74.1 ± 9.4	82.5 ± 4.82
			Normal	33.2 ± 11.7	37.3 ± 10.7	35.1 ± 11.2	55.6 ± 8.62
INCEPTION [20]	83.3 ± 7.71	20,867,625	COVID-19	**85.6** ± 4.79	80.9 ± 8.27	83.2 ± 6.1	93.1 ± 7.01
			Pneumonia	**80.2** ± 3.43	81.2 ± 3.10	80.7 ± 3.3	91.9 ± 7.33
			Normal	**67.3** ± 15.4	60.7 ± 7.71	63.8 ± 10.3	72.6 ± 4.51
PROPOSED	93.4 ± 3.46	290,891	COVID-19	95.8 ± 2.58	94.5 ± 1.22	95.1 ± 1.7	98.1 ± 2.30
			Pneumonia	95.1 ± 3.84	94.8 ± 1.34	94.9 ± 2.0	98.8 ± 1.68
			Normal	91.2 ± 3.73	87.7 ± 3.43	89.4 ± 3.6	95.7 ± 1.33

5.2. Statistical Analysis

To further investigate whether the achieved results are statistically significant from those obtained from the competing DL methods, the Friedman omnibus test is applied as a common, popular way of contrasting the performance given by ML models. Hence, the Friedman test is applied to the stratified five-fold cross-validation results and the calculated p-values are presented in Table 3. The Python library SciPy is used to implement the Friedman test with threshold $\sigma = 0.05$. It could be noted that all p-values are less than the threshold, implying the rejection of the null hypothesis. This means that the results of the proposed model statistically differ from those of the competing methods across different metrics.

Table 3. The statistical results obtained from the Friedman test with a significance threshold $\sigma = 0.05$.

Method	Accuracy	F1-score	AUC
Proposed vs. CNN [17]	6.25×10^{-3}	5.22×10^{-8}	5.03×10^{-3}
Proposed vs. POCOVID-Net [19]	7.40×10^{-5}	6.98×10^{-3}	9.90×10^{-4}
Proposed vs. Mini-COVIDNet [21]	5.96×10^{-3}	5.96×10^{-6}	7.63×10^{-3}
Proposed vs. Residual Net [23]	2.06×10^{-4}	4.76×10^{-3}	8.89×10^{-7}
Proposed vs. inception [20]	9.84×10^{-6}	5.50×10^{-5}	2.54×10^{-5}

5.3. Ablation Analysis

To deep dive into the building blocks of the proposed model, a group of ablation experiments is performed to analyze the role and the contribution of each building block to the total screening performance. The results of the ablation experiments are presented in Table 4. In these experiments, a shallow version of the standard vision transformer is used as a baseline model and achieved 89.6% accuracy, 89.2% F1-score, and 94.8% AUC. Moving forward, the inclusion of the transformer module with WMHA is shown to be beneficial for improving the COVID-19 screening performance (accuracy: 90.7%, F1-score: 90.8%, AUC: 95.1%). In the same way, the inclusion of the transformer module with SWMHA is observed to lead to similar improvements. More interestingly the parallel integration of the above modules in our model significantly improved the classification performance across all metrics (accuracy: 92.5%, F1-score: 91.9%, AUC: 97.1%). This explains the importance of both features in separate windows and cross-windows being essential to improving the representation power of the model. Finally, the integration of the proposed weighted pooling module is obviously improving the classification performance.

Table 4. Ablation experiments of the proposed model under 5-fold cross-validation.

METHOD	ACCURACY	F1-SCORE	AUC
BASELINE	88.1 ± 2.96	87.3 ± 4.99	94.2 ± 2.45
+CONVOLUTIONAL PATCHING	89.6 ± 1.74	89.2 ± 3.09	94.8 ± 2.68
+WMHA	90.7 ± 3.73	90.8 ± 2.16	95.1 ± 1.08
+SWMHA	90.9 ± 2.35	90.3 ± 2.24	95.6 ± 1.52
+ (WMHA||SWMHA)	92.5 ± 3.16	91.9 ± 1.93	97.1 ± 2.39
+WIGHTED POOLING	93.4 ± 3.46	93.1 ± 2.43	97.5 ± 1.77

5.4. Explainability Analysis

To understand and interpret the screening decision obtained from the proposed model, the class-related saliency maps for some COVID-19 and pneumonia cases are presented in Figure 3A,B, correspondingly.

It could be seen that the GRAD-CAM++ shades the significant lung areas in the ultrasound frame, contributing to making the prediction of the output classes. It is also observable that the model activates diverse zones in the input frame corresponding to different biomarkers to be learned and considered during the screening. These diverse zones adopted for screening are learned intrinsically in the model. Moreover, it is notable that the regions of activations vary from one frame to another, even though they both belong to the same class of infection, hence, these activations can be further enhanced by using more ultrasound data from the same class. For the pneumonia frame, one may observe the presence of pleural consolidations, which is a common biomarker for that disease [33,34]. On the other hand, COVID-19 was demonstrated to show abnormal pleural lines together with upright artifacts in lung ultrasound frames/videos. Figure 3A highlights the location of the lung infection lesions where it is obvious that our model is considering the area in close proximity to the pleural lines for screening the COVID-19 class. This assessment also establishes that the proposed solution could promptly identify the cases suffering from considerable lung abnormalities displayed as B-lines, which enables improved screening of COVID-19 patients.

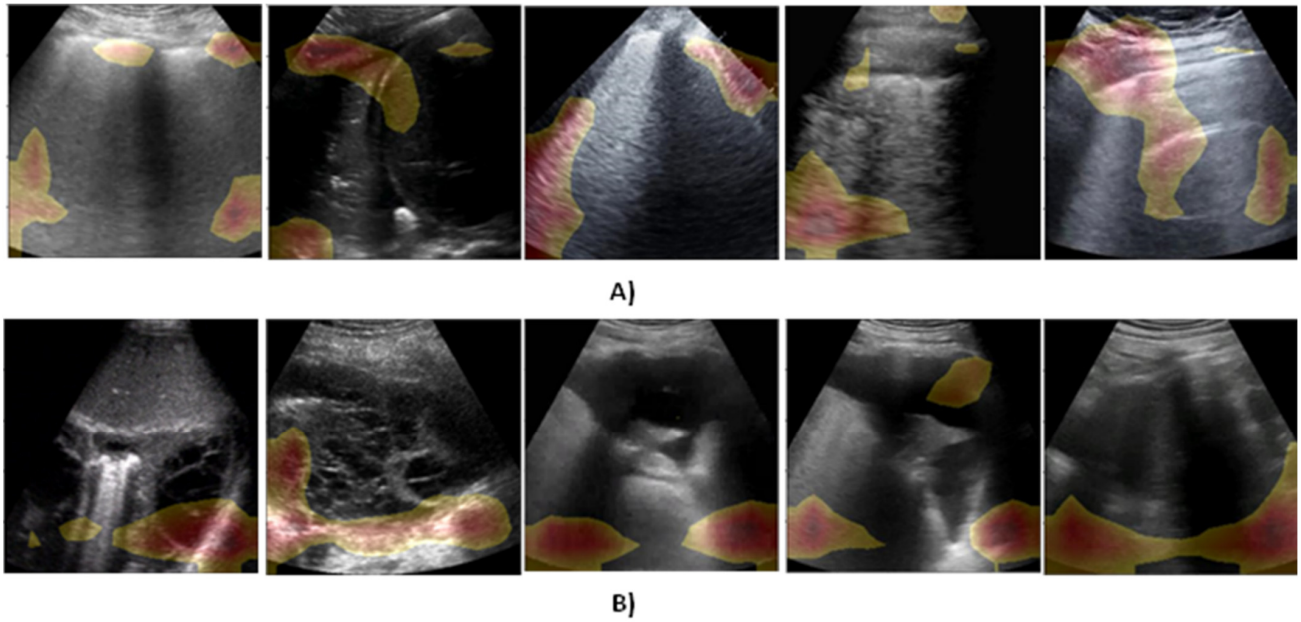

Figure 3. Illustration of visual explanation maps generated by Grad-CAM++ for some of the samples of lung ultrasound frame images after visualization. (**A**) COVID-19-infected lung, (**B**) Pneumonia-infected lung.

6. Conclusions and Future Work

This work presents a lightweight and explainable convolutional transformer model for the efficient screening of COVID-19 from ultrasound data. The representation power of the model is further improved by convolutional patching and parallel window-based transformation modules. The findings demonstrate that the proposed solution considerably improves COVID-19 detection performance with high information compactness, meaning that it achieves both efficiency (high detection accuracy) and effectiveness (a small number of trainable parameters). Beyond and above this, the classification decisions obtained from the proposed solution can be visually explained so they can be easily interpreted and trusted by medical staff. These competitive advantages of the proposed solution render it a candidate for improving the quality of ultrasonic diagnosis in smart healthcare systems during and after the pandemic.

This work can be extended in three ways in the future. First, in coping with the sustainable development strategy in Egyptian Vision 2030, the proposed solution will be extended to be provided as a sustainable diagnosis service/system that can be collaboratively trained using ultrasound data from different Egyptian hospitals. By doing so, the Egyptian Ministry of Healthcare will have great management of COVID-19-like pandemics with automated, efficient, interpretable, and effective tools. Second, the proposed solution will be extended to take advantage of 5G and B5G communication to deliver the patients' data and corresponding diagnosis decisions in real time. This direction will specifically focus on the responsiveness of our system as an essential requirement of the sustainability of the Egyptian healthcare system. Third, the proposed solution will be extended to learn from different modalities of data to improve the quality and functionality of COVID-19 diagnosis in the healthcare system.

Author Contributions: Conceptualization, M.A.-B., H.H. and K.M.S.; methodology, M.A.-B., H.H. and K.M.S.; software, M.A.-B. and H.H.; validation, M.A.-B., H.H., A.W.M., K.A.A. and K.M.S.; formal analysis, M.A.-B., H.H., K.A.A., A.W.M. and K.M.S.; investigation, M.A.-B., H.H., A.W.M. and K.M.S.; data curation, M.A.-B., H.H., A.W.M. and K.M.S.; writing—original draft preparation, M.A.-B. and H.H.; writing—review and editing M.A.-B., H.H., K.A.A., A.W.M. and K.M.S.; visualization, M.A.-B., H.H. and K.M.S.; supervision, M.A.-B.; funding acquisition, K.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Researchers Supporting Program at King Saud University, (RSP-2021/305).

Acknowledgments: The authors present their appreciation to King Saud University for funding the publication of this research through the Researchers Supporting Program (RSP-2021/305), King Saud University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare that there are no conflict of interest in the research.

References

1. WHO. WHO Coronavirus Disease. 2020. Available online: WHO.int (accessed on 30 September 2022).
2. Garg, A.; Ghoshal, U.; Patel, S.S.; Singh, D.V.; Arya, A.K.; Vasanth, S.; Pandey, A.; Srivastava, N. Evaluation of seven commercial RT-PCR kits for COVID-19 testing in pooled clinical specimens. *J. Med Virol.* **2020**, *93*, 2281–2286. [CrossRef] [PubMed]
3. Bernheim, A.; Mei, X.; Huang, M.; Yang, Y.; Fayad, Z.A.; Zhang, N.; Diao, K.; Lin, B.; Zhu, X.; Li, K.; et al. Chest CT Findings in Coronavirus Disease-19 (COVID-19): Relationship to Duration of Infection. *Radiology* **2020**, *295*, 200463. [CrossRef] [PubMed]
4. Mischi, M.; Bell, M.A.L.; Van Sloun, R.J.G.; Eldar, Y.C. Deep Learning in Medical Ultrasound—From Image Formation to Image Analysis. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **2020**, *67*, 2477–2480. [CrossRef]
5. Bansal, G.; Chamola, V.; Narang, P.; Kumar, S.; Raman, S. Deep3DSCan: Deep residual network and morphological descriptor based framework forlung cancer classification and 3D segmentation. *IET Image Process.* **2020**, *14*, 1240–1247. [CrossRef]
6. Rohmetra, H.; Raghunath, N.; Narang, P.; Chamola, V.; Guizani, M.; Lakkaniga, N.R. AI-enabled remote monitoring of vital signs for COVID-19: Methods, prospects and challenges. *Computing* **2021**, 1–27. [CrossRef]
7. Chamola, V.; Hassija, V.; Gupta, V.; Guizani, M. A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact. *IEEE Access* **2020**, *8*, 90225–90265. [CrossRef]
8. Zhang, Y.; He, X.; Tian, Z.; Jeong, J.J.; Lei, Y.; Wang, T.; Zeng, Q.; Jani, A.B.; Curran, W.J.; Patel, P.; et al. Multi-Needle Detection in 3D Ultrasound Images Using Unsupervised Order-Graph Regularized Sparse Dictionary Learning. *IEEE Trans. Med Imaging* **2020**, *39*, 2302–2315. [CrossRef]
9. Mento, F.; Perrone, T.; Fiengo, A.; Tursi, F.; Macioce, V.N.; Smargiassi, A.; Inchingolo, R.; Demi, L. Limiting the areas inspected by lung ultrasound leads to an underestimation of COVID-19 patients' condition. *Intensive Care Med.* **2021**, *47*, 811–812. [CrossRef]
10. McElyea, C.; Do, C.; Killu, K. Lung ultrasound artifacts in COVID-19 patients. *J. Ultrasound* **2020**, *25*, 333–338. [CrossRef]
11. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]
12. Xie, X.; Niu, J.; Liu, X.; Chen, Z.; Tang, S.; Yu, S. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Med. Image Anal.* **2021**, *69*, 101985. [CrossRef] [PubMed]
13. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans. Neural Networks Learn. Syst.* **2020**, *32*, 4793–4813. [CrossRef] [PubMed]
14. Ahmed, I.; Jeon, G.; Piccialli, F. From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where. *IEEE Trans. Ind. Inform.* **2022**. [CrossRef]
15. Roy, S.; Menapace, W.; Oei, S.; Luijten, B.; Fini, E.; Saltori, C.; Huijben, I.; Chennakeshava, N.; Mento, F.; Sentelli, A.; et al. Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound. *IEEE Trans. Med. Imaging* **2020**, *39*, 2676–2687. [CrossRef]
16. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, Montreal, BC, Canada, 11–17 October 2021; pp. 9992–10002. [CrossRef]
17. Muhammad, G.; Hossain, M.S. COVID-19 and Non-COVID-19 Classification using Multi-layers Fusion From Lung Ultrasound Images. *Inf. Fusion* **2021**, *72*, 80–88. [CrossRef] [PubMed]
18. Chattopadhyay, A.; Sarkar, A.; Howlader, P. Grad-CAM ++ : Improved Visual Explanations for Deep Convolutional Networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847. [CrossRef]
19. Born, J.; Wiedemann, N.; Cossio, M.; Buhre, C.; Brändle, G.; Leidermann, K.; Goulet, J.; Aujayeb, A.; Moor, M.; Rieck, B.; et al. Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis. *Appl. Sci.* **2021**, *11*, 672. [CrossRef]
20. Diaz-Escobar, J.; Ordóñez-Guillén, N.E.; Villarreal-Reyes, S.; Galaviz-Mosqueda, A.; Kober, V.; Rivera-Rodriguez, R.; Rizk, J.E.L. Deep-learning based detection of COVID-19 using lung ultrasound imagery. *PLoS ONE* **2021**, *16*, e0255886. [CrossRef]
21. Awasthi, N.; Dayal, A.; Cenkeramaddi, L.R.; Yalavarthy, P.K. Mini-COVIDNet: Efficient Lightweight Deep Neural Network for Ultrasound Based Point-of-Care Detection of COVID-19. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **2021**, *68*, 2023–2037. [CrossRef]
22. Frank, O.; Schipper, N.; Vaturi, M.; Soldati, G.; Smargiassi, A.; Inchingolo, R.; Torri, E.; Perrone, T.; Mento, F.; Demi, L.; et al. Integrating Domain Knowledge Into Deep Networks for Lung Ultrasound With Applications to COVID-19. *IEEE Trans. Med. Imaging* **2021**, *41*, 571–581. [CrossRef]
23. La Salvia, M.; Secco, G.; Torti, E.; Florimbi, G.; Guido, L.; Lago, P.; Salinaro, F.; Perlini, S.; Leporati, F. Deep learning and lung ultrasound for Covid-19 pneumonia detection and severity classification. *Comput. Biol. Med.* **2021**, *136*, 104742. [CrossRef]

24. Xue, W.; Cao, C.; Liu, J.; Duan, Y.; Cao, H.; Wang, J.; Tao, X.; Chen, Z.; Wu, M.; Zhang, J.; et al. Modality alignment contrastive learning for severity assessment of COVID-19 from lung ultrasound and clinical information. *Med Image Anal.* **2021**, *69*, 101975. [CrossRef] [PubMed]
25. Wu, Y.-H.; Gao, S.-H.; Mei, J.; Xu, J.; Fan, D.-P.; Zhang, R.-G.; Cheng, M.-M. JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 3113–3126. [CrossRef]
26. Wang, X.; Jiang, L.; Li, L.; Xu, M.; Deng, X.; Dai, L.; Xu, X.; Li, T.; Guo, Y.; Wang, Z.; et al. Joint Learning of 3D Lesion Segmentation and Classification for Explainable COVID-19 Diagnosis. *IEEE Trans. Med. Imaging* **2021**, *40*, 2463–2476. [CrossRef] [PubMed]
27. Shi, W.; Tong, L.; Zhu, Y.; Wang, M.D. COVID-19 Automatic Diagnosis with Radiographic Imaging: Explainable Attention Transfer Deep Neural Networks. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 2376–2387. [CrossRef] [PubMed]
28. Brunese, L.; Mercaldo, F.; Reginelli, A.; Santone, A. Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays. *Comput. Methods Programs Biomed.* **2020**, *196*, 105608. [CrossRef]
29. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *2017*, 5999–6009.
30. Ganesh, P.; Chen, Y.; Lou, X.; Khan, M.A.; Yang, Y.; Sajjad, H.; Nakov, P.; Chen, D.; Winslett, M. Compressing large-scale transformer-based models: A case study on bert. *Trans. Assoc. Comput. Linguistics* **2021**, *9*, 1061–1080. [CrossRef]
31. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
32. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM : Visual Explanations from Deep Networks. *Int. J. Comput. Vis.* **2019**.
33. Boccatonda, A.; Ianniello, E.; D'Ardes, D.; Cocco, G.; Giostra, F.; Borghi, C.; Schiavone, C. Can Lung Ultrasound be Used to Screen for Pulmonary Embolism in Patients with SARS-CoV-2 Pneumonia? *Eur. J. Case Rep. Intern. Med.* **2020**. [CrossRef]
34. Monteiro, R.A.D.A.; Duarte-Neto, A.N.; da Silva, L.F.F.; de Oliveira, E.P.; Nascimento, E.C.T.D.; Mauad, T.; Saldiva, P.H.D.N.; Dolhnikoff, M. Ultrasound assessment of pulmonary fibroproliferative changes in severe COVID-19: A quantitative correlation study with histopathological findings. *Intensive Care Med.* **2021**, *47*, 199–207. [CrossRef] [PubMed]