COVID-19 Classification Using Pre-trained Models and Disease Severity Score Masks

Ebrahim A. Nehary, Sreeraman Rajan, and Carlos Rossa
Department of Systems and Computer Engineering
Carleton University, Ottawa, Canada
ebrahimali@cmail.carleton.ca; sreeramanr@sce.carleton.ca; rossa@sce.carleton.ca

***Abstract*—The significance of early detection of COVID-19 has been widely acknowledged as a means of reducing its spread and mortality rates among patients. Deep learning techniques for COVID-19 classification based on ultrasound (US) data have been extensively employed. However, detecting COVID-19 based on US images continues to be challenging primarily due to limited datasets with noisy and low-resolution images. This study investigates methods to enhance classification performance by incorporating disease severity score masks while training pre-trained models enhanced with self-attention mechanisms. The disease severity scores range from 0 for healthy lung tissue to 1 for initial signs of abnormality, and 2 and 3 for advanced pathological artifacts. These masks and their corresponding US images are employed as inputs to pre-trained models for feature extraction. Subsequently, features extracted from the masks are utilized to recalibrate features obtained from US images using self-attention mechanisms. The proposed method achieves classification accuracy of 95.4, 90.4, 95%, 83%, and 92% when using pre-trained models VGG16, NASNet-Mobile, MobileNet_V2, ResNet50, and Xception, respectively. Further, all pre-trained models yield a low standard deviation of less than 5%. The results demonstrate that incorporating disease severity masks improves the classification performance, thus offering promising techniques for enhancing COVID-19 detection using ultrasound imaging.**

***Index Terms*—COVID-19 , US images, pre-trained models, classification, severity score, attention**

I. Introduction

The COVID-19 virus has affected more than half a billion people by mid-2022, resulting in 6 million deaths globally [1]. This virus affects the respiratory system, with varying severity of symptoms ranging from minor, such as cough and fever, to major such as organ failure and even death [2]–[5]. The virus is considered a threat to humans because it can spread directly through contact with an affected person or indirectly through airborne or droplet transmission. [6], [7]. Consequently, the spread of the virus caused a worldwide shutdown after being proclaimed a pandemic by the World Health Organization.

Detection of COVID-19 mainly relies on the reverse transcription-polymerase chain reaction (RT-PCR) test; however, this test has been reported to have low sensitivity [5]. Therefore, medical imaging modalities with high sensitivity and accuracy such as computed tomography (CT) [8], [9] and ultrasound (US) [10], [11] were proposed to detect COVID-19. Ultrasound is considered a better option than CT because it is non-ionizing, portable, easy to disinfect (to prevent the spread of COVID-19 from patient to another subject or to healthcare workers such as physicians and nurses), and less expensive [5], [12]. However, detection and classification of COVID-19 using US images are still challenging due to limited available datasets with noisy and low resolution of US images.

In literature, various deep-learning models, mostly utilizing pre-trained models, have been proposed for COVID-19 detection using ultrasound (US) images. Pre-trained models are preferred due to the limited availability of US image datasets. These models aim to classify COVID-19 into three classes, namely, lungs affected by the COVID-19 virus, lungs affected by bacterial pneumonia, and healthy lungs (normal). For instance, pre-trained VGG16 was proposed in [12]–[14], while VGG19, InceptionV3, Xception, and ResNet50 were presented in [15] for COVID-19 detection. Additionally, VGG16 and ResNet18 were proposed in [16], ResNet50 was employed in [17], and ResNet50, DenseNet121, Inception_ResNet_V2, and Inception_V3 were introduced in [18] to extract features from input US images and then classify the condition of the subject lung. Further pre-trained methods are presented in [10], [19]. It is worth noting that most of these pre-trained models are trained using ImageNet dataset [20], except pre-trained using in [18] they are trained using RadImageNet [21]. Although pre-trained models (trained using ImageNet or RadImageNet) are used for COVID-19 detection to tackle the limited US dataset, the classification results need further improvement using a more advanced classification method. Therefore, pre-trained models are also used to extract features from US images in this work and refine these features using features that are extracted from disease severity score masks.

Deep-learning models have also been proposed to predict the severity score of COVID-19 infection based on lung ultrasound (LUS) images [22]. The severity score measures the progression of COVID-19 infection. Score 0 represents a healthy lung, indicated by the presence of a continuous pleural line. The first sign of alteration in the pleural line (A-line) is marked as an abnormality and given a score of 1. Advanced lung infection with consolidation, whether small or large, is indicated by score 2. Finally, the presence of a hyperechogenic area (white lung) under the pleural surface and B-line is assigned a score of 3. Different deep-learning models have been proposed to predict the disease severity

979-8-3503-5173-6/24/$31.00 ©2024 IEEE

score [22]–[26]. For example, spatial transformer networks and frame-based segmentation have been proposed for score prediction and estimation of the segmentation mask indicating pathological artifacts [22]. However, a protocol for assessing COVID-19 severity to score lung ultrasound dataset has yet to be standardized [27]. Therefore, it is hard to train deep learning for disease severity scores and compare results across various datasets, even if such datasets are available.

However, exploiting severity scoring may help enhance the classification of COVID-19 into the lungs affected by COVID-19, lungs affected by bacterial pneumonia, and healthy lungs. A healthy lung should have a severity score of 0, while COVID-19 or bacterial pneumonia are expected to have other scores (1, 2, and 3) due to their consolidation, B-line, and other pathological artifacts. Fortunately, a publicly available pre-trained segmentation model trained by [22] can predict the severity score mask, as shown in Figure 1 (further details in Section II). Therefore, this work investigates the idea of using the predicted score mask features to calibrate the features extracted from the associated US image to enhance the classification performance of pre-trained models. The features are extracted from the score mask and US image using various pre-trained models, namely, VGG16, NASNet-Mobile, MobileNet_V2, ResNet50, and Xception. The attention is carried out using multihead self-attention. It is also worth mentioning that the CNN model has shared weights to reduce its complexity and keep it lightweight.

II. Methodology

As mentioned earlier, this work investigates leveraging the potential of a score mask to improve the classification of ultrasound (US) images (frames) into three classes: COVID-19, bacterial pneumonia, and normal (healthy lung). Since there is no available severity score, the publicly available pre-trained model by [22] is employed to obtain the severity score mask for each US image. Frames (images) are extracted from the US video clip by choosing non-adjacent frames, as it has been shown in [18] that the non-adjacent frame selection is superior to the common method (constant frame selection) and comparable to random frame selection. Then, each frame is fed as input to the pre-trained models provided by [22] to obtain the severity score. Each US frame (image) has four severity score masks, as shown in Figure 1. The COVID-19 frame depicted in Figure 1(a) and its corresponding score masks 0, 1, 2, and 3 are shown in (b), (c), (d), and (e) respectively. Since the severity score is based on the appearance of pathology artifacts in the US image, this COVID-19 case in figure 1 should have pathology artifacts (such as B-line) in scores 1 to 3. As expected, more artifacts (B-lines) are visible in score masks 1, 2, and 3. Similarly, bacterial pneumonia is depicted in Figure 1(f) and its corresponding score masks 0, 1, 2, and 3 are shown in (g), (h), (i), and (j) respectively. Consolidation artifacts can be clearly seen in Figure 1(j). In contrast, the normal (healthy lung) US image is shown in the last row of Figure 1 (k). The normal class is characterized by the presence of A-lines, which appear as horizontal lines in the US image. As expected, only A-lines should appear in the scoring mask 0 (see Figure 1(l)). Such masks may aid in recalibrating US image features using multi-head self-attention and subsequently enhancing COVID-19 classification.

The chosen frame containing the US image and its associated disease severity score are used as the inputs to the pre-trained model with shareable weights as illustrated in Figure 2. A pre-trained CNN model is utilized to extract features from the US images and their corresponding score masks. Any CNN model from the pool of pre-trained models such as VGG16, NASNet-Mobile, MobileNet_V2, ResNet50, and Xception) may be considered to extract abstract features. Notably, the CNN model is designed to be shareable, ensuring the proposed model remains lightweight and computationally efficient. Next, the extracted features from the US image undergo self-attention with features extracted from each scored mask individually. The multi-head self-attention mechanism, as proposed in [28], is employed here to recalibrate the features extracted from the US image. Subsequently, the recalibrated features are concatenated and fed into two fully connected layers (FC) for final classification. The first fully connected layer comprises 128 neurons with the Rectified Linear Exponential Unit (RLEU) activation function. In comparison, the second fully connected layer consists of three neurons (corresponding to the number of classes) with the softmax activation function. The pre-trained model's layers are frozen except for the last block. The proposed model is trained using 100 epochs, focal loss function (a generalization of the cross-entropy loss function), Adam optimization, and an early stopping method based on validation loss.

The performance of the proposed method is evaluated by comparing the true label with the predicted label, and then the confusion matrix is generated. Sensitivity, precision, F1_score, and accuracy are used to assess the performance of the proposed method, and they are given as follows.

$$Sensitivity = \frac{TP}{TP + FN}\,, \tag{1}$$

$$Precision = \frac{TP}{TP + FP}\,, \tag{2}$$

$$F1\_score = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision}\,, \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}\,, \tag{4}$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively.

III. Results

The proposed method is evaluated using US COVID-19 images in [12], [29]. The dataset comprises 202 videos and 59 images captured using convex or linear ultrasound probes from 216 patients. This dataset exhibits high level of heterogeneity as it is acquired from various sources; therefore, only a subset of the dataset is utilized in this study so that the dataset used is homogeneous. The data is obtained from Northumbria (a

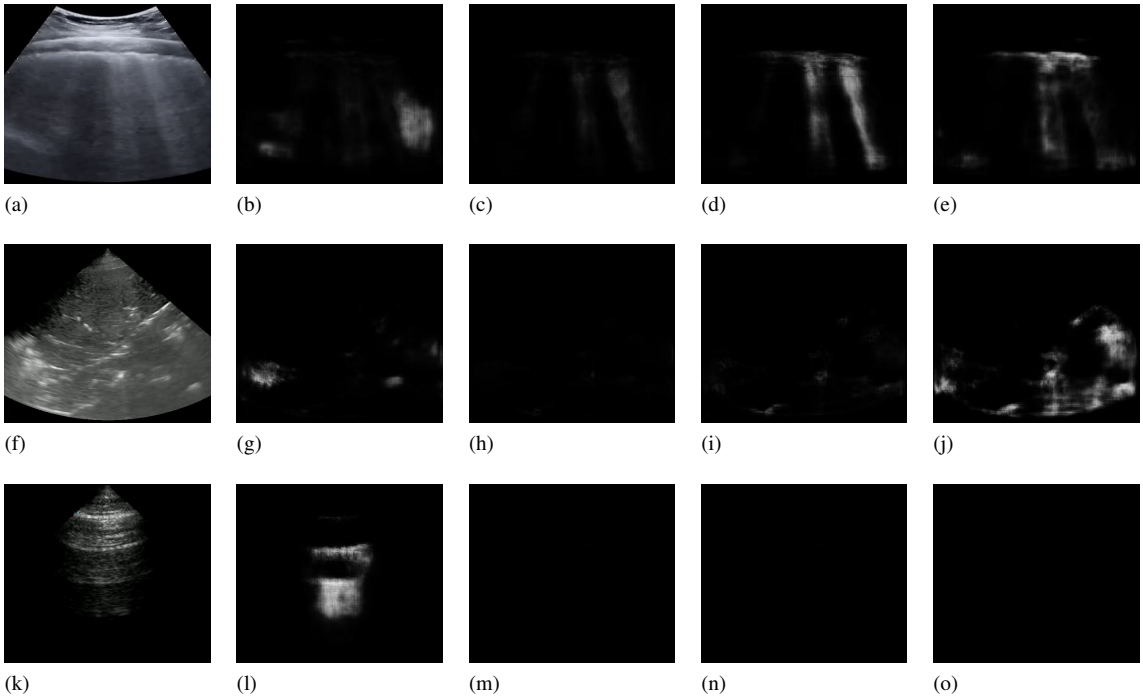

Fig. 1: Samples of US images depicting COVID-19, bacterial pneumonia, and healthy lungs are provided, each with their associated disease severity scores. COVID-19 US images are shown in (a), with corresponding score masks 0, 1, 2, and 3 depicted in (b), (c), (d), and (e), respectively. Bacterial pneumonia US images are shown in (f), with corresponding score masks 0, 1, 2, and 3 shown in (g), (h), (i), and (j), respectively. Normal (healthy lung) US images are presented in (k), with corresponding score masks 0, 1, 2, and 3 illustrated in (l), (m), (n), and (o), respectively.

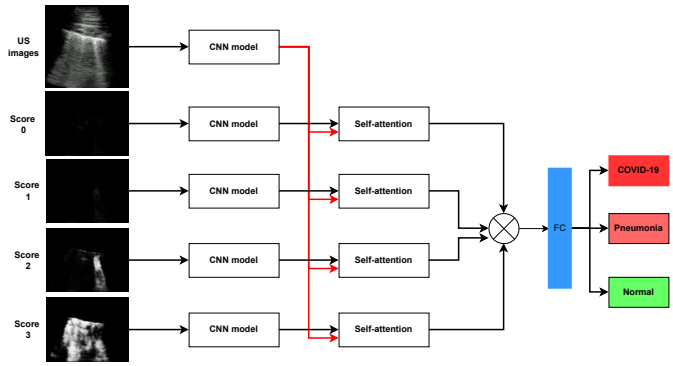

Fig. 2: The overall classification model takes the US image and its associated disease severity score masks as input. The severity score ranges from 0 to 3, where 0 indicates a healthy lung, 1 indicates the first sign of abnormality, and 2 and 3 indicate an advanced pathological state. FC and $\otimes$, pneumonia denote the fully connected layer, concatenation, and Bacterial pneumonia, respectively.

healthcare center serving a large population in the northeast of the United Kingdom) and Neuruppin (a medical school in Neuruppin, Germany). This data acquisition employs GE Healthcare US equipment and follows the BLUE protocol [30]. Furthermore, the presence of COVID-19 has been confirmed through RT-PCR testing, while bacterial pneumonia has been confirmed using thoracic X-ray and CT scans. For this paper, the data is divided into five-fold cross-validation based on the video clip to avoid data leakage. The results provided are averaged over five repetitions and presented. Performance metrics used in this paper are averaged sensitivity, precision, F1-score, and accuracy, along with their standard deviations.

Table I illustrates the classification performance using various CNN models (pre-trained models: VGG16, NASNet-Mobile, MobileNet_V2, ResNet50, and Xception) along with all severity score masks. The performance metrics are all above 90%, except when utilizing the pre-trained ResNet50 model, which may be attributed to limited data for training this particular model. The pre-trained VGG16 model demonstrates the highest classification performance with an accuracy of 95.4%. It exhibits the lowest standard deviation across all classification metrics, except for precision, where the lowest standard deviation is observed with the pre-trained Xception model at 3%. In summary, these preliminary results, employing score masks, pre-trained models, and multi-head self-attention, indicate an improvement in classification performance. Further experimentation is required, including training with more advanced deep learning models such as vision transformers and potentially collecting a larger COVID-19 US dataset.

The use of the severity score mask contributes to improving the classification results. To investigate the effectiveness of employing this severity score mask, experiments are conducted

TABLE I: The classification results when using various CNN models and masks of disease severity scores (0, 1, 2, and 3). The bold values indicate the best mean (M) and standard deviation (S) results.

CNN model		Sensitivity	Precision	F1_score	Accuracy
VGG16	M	**95.6**	**95.6**	**95.6**	**95.4**
	S	**3.3**	3.2	**3.2**	**3.4**
NASNet-Mobile	M	90.6	91.2	90.4	90.4
	S	4.3	3.5	4.3	4.5
Mobilenet_V2	M	95.0	95.2	95.0	95.0
	S	3.5	3.0	3.5	3.5
ResNet50	M	83.6	83.6	83.2	83.0
	S	4.4	4.4	4.5	4.6
Xception	M	91.6	92.6	92.2	92.0
	S	4.0	**3.0**	3.7	4.1

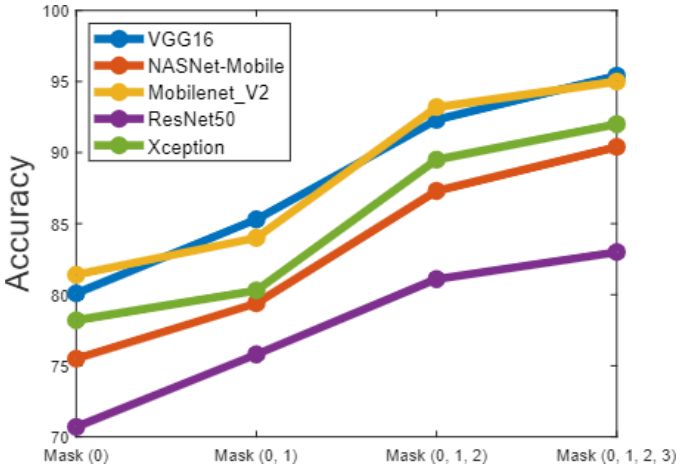

Fig. 3: The accuracy of various CNN models when using US images with mask representation of score 0, mask representation of scores 0 and 1, mask representation of scores 0, 1, and 2, and mask representation of scores 0, 1, 2, and 4.

using different configurations: one mask representing score 0, two masks representing scores 0 and 1, three masks representing scores 0, 1, and 2, and all score masks as depicted in Figure 2. The accuracy results for all configurations using pre-trained models are shown in Figure 3. Figure 3 illustrates that the classification performance improves by including more severity score masks, with the best result obtained when using all severity score masks. For instance, VGG16 provides accuracies of 80%, 85%, 92%, and 95% when using masks for scores 0, 0 and 1, 0, 1, and 2, and 0, 1, 2, and 3, respectively. This demonstrates an improvement from 80% to 95% when utilizing all severity score masks. Furthermore, pre-trained VGG16 and MobileNet_V2 have almost similar classification performance, as depicted in Figure 3. These two pre-trained models provide the best classification results among all the models considered in this paper. These results indicate that feature extraction from the severity score masks enhances the classification performance.

IV. Conclusions

Detection of COVID-19 using ultrasound (US) images remains challenging for several reasons, including the noise and low resolution inherent in US images, as well as the limited availability of datasets. This study aims to enhance classification performance by incorporating severity scores corresponding to US images, utilizing pre-trained models to address dataset limitations, and employing self-attention mechanisms to recalibrate features extracted from US images based on individual severity masks. The improvements in classification performance demonstrate the effectiveness of the proposed method, with all pre-trained models achieving an accuracy of over 90%, except for the ResNet50 model. However, a limitation of this study is the reliance on severity score masks from a pre-existing method, as the corresponding severity score data is private. Therefore, further experimentation is contingent upon releasing this dataset to the public domain or acquiring a similar scoring dataset. Nevertheless, despite this limitation, the promising results suggest that incorporating severity masks enhances classification performance and may facilitate the effective detection of COVID-19 using US images.

Future work may include pre-trained models optimized using RadImagenet instead of Imagenet. Also, independent models for severity and classification can be attempted when appropriate curated dataset becomes publicly available.

V. Acknowledgments

This work was financially supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] "Centers for disease control and prevention (CDC)." Available at https://coronavirus.jhu.edu/map.html. Accessed: 2024-05-08.

[2] T. Chen *et al.*, "Clinical characteristics of 113 deceased patients with coronavirus disease 2019: Retrospective study," *bmj*, vol. 368, 2020.

[3] N. El-Rashidy *et al.*, "Comprehensive survey of using machine learning in the COVID-19 pandemic," *Diagnostics*, vol. 11, no. 7, p. 1155, 2021.

[4] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, and P.-R. Hsueh, "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges," *International Journal of Antimicrobial Agents*, vol. 55, no. 3, p. 105924, 2020.

[5] C. McDermott, M. Łącki, B. Sainsbury, J. Henry, M. Filippov, and C. Rossa, "Sonographic diagnosis of COVID-19: A review of image processing for lung ultrasound," *Frontiers in Big Data*, vol. 4, p. 612561, 2021.

[6] P. Resende *et al.*, "The ongoing evolution of variants of concern and interest of SARS-CoV-2 in Brazil revealed by convergent indels in the amino (N)-terminal domain of the spike protein," *Virus Evolution*, vol. 7, no. 2, p. veab069, 2021.

[7] E. Volz *et al.*, "Transmission of SARS-CoV-2 Lineage B. 1.1. 7 in England: Insights from linking epidemiological and genetic data," *MedRxiv*, pp. 2020–12, 2021.

[8] L. Sun *et al.*, "Adaptive feature selection guided deep forest for COVID-19 classification with chest CT," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2798–2805, 2020.

[9] Z. Wang, Q. Liu, and Q. Dou, "Contrastive cross-site learning with redesigned net for COVID-19 CT classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2806–2813, 2020.

[10] G. Muhammad and M. Hossain, "COVID-19 and non-COVID-19 classification using multi-layers fusion from lung ultrasound images," *Information Fusion*, vol. 72, pp. 80–88, 2021.

[11] N. Awasthi, A. Dayal, L. Cenkeramaddi, and P. Yalavarthy, "Mini-COVIDNet: Efficient lightweight deep neural network for ultrasound based point-of-care detection of COVID-19," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 6, pp. 2023–2037, 2021.

[12] J. Born *et al.*, "Accelerating detection of lung pathologies with explainable ultrasound image analysis," *Applied Sciences*, vol. 11, no. 2, p. 672, 2021.

[13] J. Born *et al.*, "POCOVID-Net: automatic detection of COVID-19 from a new lung ultrasound imaging dataset (POCUS)," *arXiv preprint arXiv:2004.12084*, 2020.

[14] E. Nehary, S. Rajan, and C. Rossa, "Lung ultrasound image classification using deep learning and histogram of oriented gradients features for COVID-19 detection," in *IEEE Sensors Applications Symposium (SAS)*, pp. 1–6, 2023.

[15] J. Diaz-Escobar *et al.*, "Deep-learning based detection of COVID-19 using lung ultrasound imagery," *Plos One*, vol. 16, no. 8, p. e0255886, 2021.

[16] J. Roberts and T. Tsiligkaridis, "Ultrasound diagnosis of COVID-19: Robustness and explainability," *arXiv preprint arXiv:2012.01145*, 2020.

[17] H. Che *et al.*, "Multi-feature multi-scale CNN-derived COVID-19 classification from lung ultrasound data," in *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2618–2621, 2021.

[18] E. Nehary, S. Rajan, and C. Rossa, "Comparison of COVID-19 classification via imagenet-based and radimagenet-based transfer learning models with random frame selection," in *IEEE Sensors Applications Symposium (SAS)*, pp. 1–6, 2023.

[19] J. Wang *et al.*, "Review of machine learning in lung ultrasound in COVID-19 pandemic," *Journal of Imaging*, vol. 8, no. 3, p. 65, 2022.

[20] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[21] X. Mei *et al.*, "RadImageNet: An open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artificial Intelligence*, vol. 4, no. 5, p. e210315, 2022.

[22] S. Roy *et al.*, "Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound," *IEEE transactions on Medical Imaging*, vol. 39, no. 8, pp. 2676–2687, 2020.

[23] O. Frank *et al.*, "Integrating domain knowledge into deep networks for lung ultrasound with applications to COVID-19," *IEEE Transactions on Medical Imaging*, vol. 41, no. 3, pp. 571–581, 2021.

[24] U. Khan *et al.*, "Deep learning-based classification of reduced lung ultrasound data from COVID-19 patients," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 5, pp. 1661–1669, 2022.

[25] L. Carrer *et al.*, "Automatic pleural line extraction and COVID-19 scoring from lung ultrasound data," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 11, pp. 2207–2217, 2020.

[26] A. G. Dastider, F. Sadik, and S. A. Fattah, "An integrated autoencoder-based hybrid CNN-LSTM model for COVID-19 severity prediction from lung ultrasound," *Computers in Biology and Medicine*, vol. 132, p. 104296, 2021.

[27] A. Lombardi, , *et al.*, "Ultrasound during the COVID-19 pandemic: A global approach," *Journal of Clinical Medicine*, vol. 12, no. 3, p. 1057, 2023.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[29] J. Born *et al.*, "L2 accelerating COVID-19 differential diagnosis with explainable ultrasound image analysis: an AI tool," *Thorax*, vol. 76, no. Suppl 1, pp. A230–A231, 2021.

[30] D. A. Lichtenstein, "BLUE-protocol and FALLS-protocol: Two applications of lung ultrasound in the critically ill," *Chest*, vol. 147, no. 6, pp. 1659–1670, 2015.