2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) | 979-8-3503-7800-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/METROXRAINE62247.2024.10796379

Vision Transformer Approaches for COVID-19 Pneumonia Assessment in Lung Ultrasound Images

Maria Chiara Fiorentino
Università Politecnica delle Marche
Ancona, Italy
m.c.fiorentino@staff.univpm.it

Riccardo Rosati
Università Politecnica delle Marche
Ancona, Italy
r.rosati@staff.univpm.it

Andrian Melnic
Università Politecnica delle Marche
Ancona, Italy
S1098384@studenti.univpm.it

Edoardo Conti
Università Politecnica delle Marche
Ancona, Italy
e.conti@staff.univpm.it

Primo Zingaretti
Università Politecnica delle Marche
Ancona, Italy
p.zingaretti@univpm.it

***Abstract*—In recent years, Convolutional Neural Networks (CNNs) have been at the forefront of advancements in medical imaging, particularly in the classification of disease severity from ultrasound images. However, CNNs often face challenges in capturing long-range dependencies within images, a crucial factor in accurately assessing conditions like COVID-19 pneumonia, where the evaluation of anatomical features and different artifacts in the lung ultrasound image (LUS) is vital. This research investigates the potential of Transformer-based models, known for their capability of capturing long-range dependencies, to classify the severity of COVID-19 pneumonia from LUS images. We explore and compare the performance of various architectures, including Swin Transformer Tiny (Swin-T), based solely on Multi-Head Self Attention (MHSA), and Bottleneck Transformer (BoTNet-50), an innovative hybrid architecture that integrates the MHSA mechanism into a convolutional backbone combining the strengths of both CNNs and Transformers. Our analysis, performed on the publicly available ICLUS dataset, reveals that BoTNet-50 outperforms ResNet-50 achieving an F1-Score of 0.6025, reflecting superior performance in LUS image classification. Swin-T, while initially underperforming when trained from scratch, saw considerable gains via transfer learning, ultimately reaching an F1-Score of 0.6513 and surpassing all ResNet-50 metrics. Moreover, Grad-CAM analysis highlights the remarkable sensitivity of Swin-T compared to ResNet-50 in detecting complex structures, leveraging its ability to discern broader contexts and complex spatial relationships in LUS images. This indicates that Transformer-based models hold significant promise for advancing the precision of COVID-19 pneumonia scoring assessment through LUS analysis.**

***Index Terms*—Deep Learning, COVID-19, Ultrasound, Vision Transformers**

I. Introduction

Emerging in late 2019, **COVID-19** is caused by the coronavirus **SARS-CoV-2** and rapidly escalated into a global pandemic, affecting millions worldwide. Primarily targeting the respiratory system, the disease manifests in a range of symptoms from mild cough with fever to severe pneumonia and more. Research into COVID-19 remains crucial today due to its widespread impact and the ongoing need for effective diagnostic and preventive measures. Early and accurate diagnosis is vital for timely intervention, with medical imaging playing a pivotal role in the early detection and monitoring of the disease's progression [1].

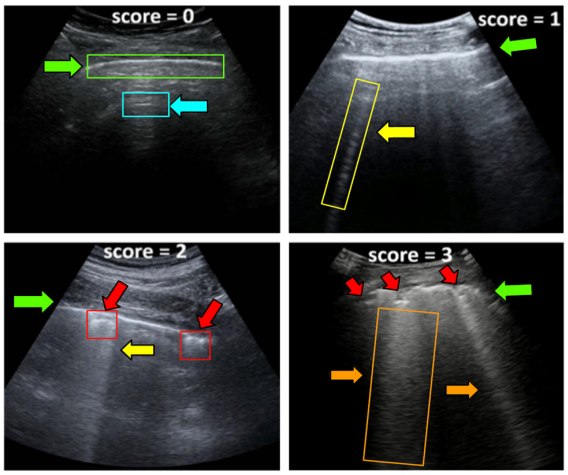


Fig. 1: Severity scores of [2], from healthy (score=0) to severe (score=3). The anatomical features are the pleural line (green) and subpleural consolidations (red). The artifacts are the A-lines (cyan), B-lines (yellow) and "white lung" (orange).

Among various diagnostic methods, Computed Tomography (CT) has played a crucial role in diagnosing and monitoring the disease's progression [1]. CT's high diagnostic sensitivity allows for early detection of pulmonary abnormalities, even before clinical symptoms manifest. However, challenges like equipment availability, infection risks and exposure to ionizing radiation pose constraints on its widespread use [1].

Point-of-care **Lung Ultrasound (LUS)** has emerged as a valuable alternative for managing COVID-19 pneumonia [1]. Unlike CT scans, LUS imaging entails no radiation exposure, minimizes patient contact, and eliminates the need for patient relocation, reducing the risk of viral transmission. While less sensitive than thoracic CT, LUS effectively identifies characteristic alterations in the pleura and subpleural regions associated with pneumonia [1].

979-8-3503-7800-9/24/$31.00 ©2024 IEEE

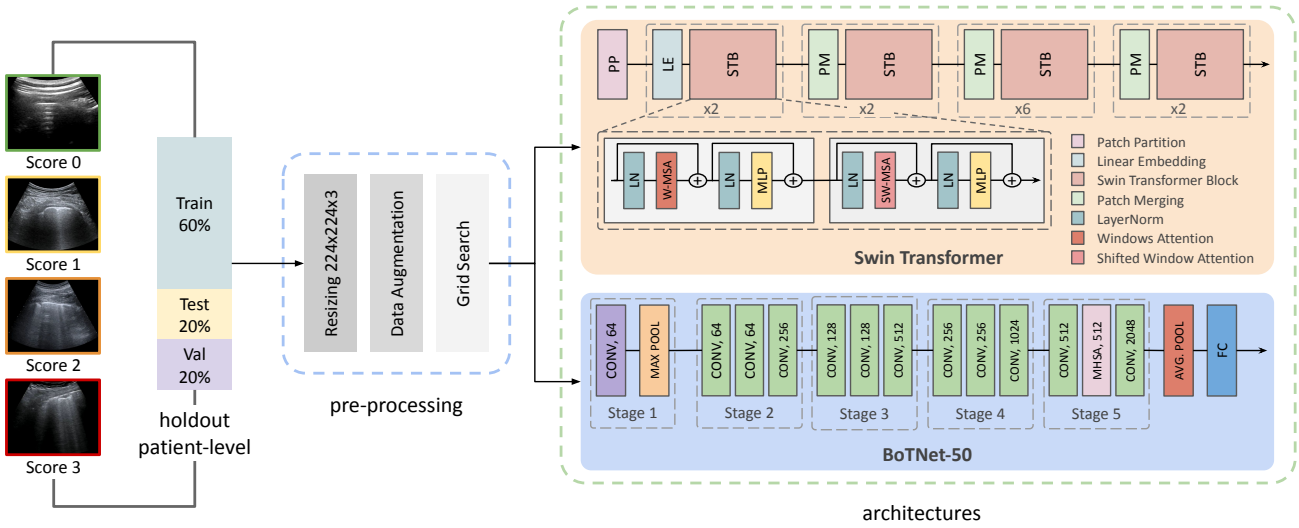


Fig. 2: The proposed workflow involves splitting the ICLUS dataset with patient holdout, followed by the preprocessing phase and testing two different architectures (Swin Transformer and BoTNet) to classify frames into the four severity classes.

The inherent characteristics of lung ultrasounds, marked by the presence of artifacts like A-lines and B-lines, areas of consolidation and the occurrence of pleural alterations (Figure 1), present significant challenges. These images often lack clear demarcations and exhibit a high degree of intra- and inter-patient variability, complicating the standardization of analysis. Thus, ultrasound interpretation requires substantial operator expertise, leading to potential human error. Furthermore, quantifying pleural alterations often relies on subjective and qualitative scales, limiting result comparability and reproducibility.

To assist with ultrasounds diagnosis, in recent years several Deep Learning methodologies have been explored. Among these, **Convolutional Neural Networks (CNNs)** have shown effectiveness in classifying, segmenting, and localizing ultrasound images by learning relevant visual features. In this field, [3] and [4] have developed state-of-the-art CNN-based approaches for LUS image classification. The former integrated a Spatial Transformer Network (STN) with an ordinal technique, while the latter introduced a framework incorporating anatomical domain knowledge during training.

While CNNs excel at identifying localized patterns and geometric structures, they may struggle with capturing complex spatial relationships and global features in ultrasound images, a limitation that becomes apparent in the presence of noise or less defined, small, or widespread pathological structures. The models based on the **Vision Transformer (ViT)**, introduced by [5], emerge as a promising alternative to CNNs, relying on a groundbreaking mechanism to analyze the entire image context and capture long-range relationships within it. ViT models have shown notable success in medical imaging, ranging from pneumonia diagnosis via chest radiography [6], to stroke classification on brain CT images [7] and automated brain tumor classification in MRI scans [8].

Regarding ultrasound images, different studies have employed ViT models, such as BabyNet introduced by [9], a hybrid architecture similar to the Bottleneck Transformer [10], designed for video-level analysis of fetal ultrasounds. Recently, [11] explored the potential of ViT models for classifying breast ultrasound images, comparing them with CNN networks. The study showed that pretrained ViT models can achieve comparable or even superior performance to CNNs, highlighting the potential of ViT models in ultrasound image classification tasks when pretrained on large datasets.

Despite the encouraging application of ViTs across medical imaging techniques, their utilization in LUS images is currently nonexistent. Further, Transformer-based models have substantial potential to address conventional limitations of CNNs in grasping complex spatial features by effectively capturing long-range relationships and overcoming challenges posed by less defined structures in LUS images. With the aim of filling this gap, this work explores the potential of Transformer methodologies, specifically the Swin Tiny and BotNet architectures, by testing which approach provides superior performance in LUS image classification. Selected for their effectiveness in diverse image classification tasks [10] [12] and potential adaptability to the challenges of LUS domain complexities, these models were chosen to evaluate whether a pure Transformer approach could outperform CNN-based technologies or if a hybrid model incorporating convolutions still remains essential.

II. Materials & Methods

In this section, we first introduce the dataset of LUS images utilized for training and evaluation. Then, we investigate the applicability and effectiveness of two proposed architectures based on Transformer methodologies. These include Swin Tiny [12], a transformer-exclusive approach and BoTNet [10], a hybrid model integrating transformers with a CNN.

A. Dataset

The **I**talian **C**ovid-19 **Lu**ng **U**ltra**s**ound **D**ata**B**ase (**ICLUS-DB**) [3] is a collection of 277 lung ultrasound videos from 35 patients, resulting in a total of 58924 frames. These videos were acquired from various clinical centers across Italy using a variety of US scanners equipped with both convex and linear probes, depending on the clinical requirements. The dataset includes videos from patients confirmed positive for COVID-19 (49%), suspected cases (11%) and healthy individuals (40%). Each frame in the dataset has been annotated to indicate the severity of COVID-19-related lung abnormalities scored on a scale ranging from 0 to 3, with a total of 45560 frames from convex probes and 13364 frames from linear probes. The distribution of scores across the dataset is as follows: 5684 frames scored as 3 (10%), 18,972 frames scored as 2 (32%), 14,295 frames scored as 1 (24%) and 19,973 frames scored as 0 (34%).

B. Pure MHSA Transformer: Swin Tiny

The **Swin Transformer** (**Swin-T**, **S**hifted **Win**dows **T**ransformer) introduced by [12] represents a significant advancement in the field of Vision Transformers, distinguished by its hierarchical approach to image processing. The architecture addresses some key limitations of the original ViT by introducing a more efficient and scalable hierarchical approach based on **Multi-Head Self Attention (MHSA)** mechanism computed on sliding windows. We actually opted for Swin Transformer over ViT primarily due to its efficient computation. While both models share similar structures, Swin-T operates within small, non-overlapping patch windows instead of compute attention globally across all patches.

Eventually, we employed the Tiny version of the model, namely **Swin Tiny**, proposed in [12] of which a comprehensive overview is shown in Figure 2 within the entire framework's flow.

The RGB input image is initially processed through the **Patch Partition** module, that divides the images into a non-overlapping grid of patches or spatial tokens. The number of patches is calculated based on the dimensions of the input image and the chosen patch size. Each patch measures 4×4 pixels, yielding a feature size of $4 \times 4 \times 3 = 48$ for each patch, where 3 are the image channels. The patches are linearly transformed in order to obtain tokens of dimension $C = 48$, which are then provided as input to a pair of Swin Transformer Blocks completing Stage 1 [12]. To achieve a hierarchical representation, Swin Transformer reduces the number of tokens by the **Patch Merging** layers placed at the beginning of each of the last three stages out of the total four. The first Patch Merging concatenates the features of each 2×2 group of adjacent patches, thereby reducing the number of tokens by a factor of four and doubling the feature size to 8×8. This process is repeated in the last three stages. As the input progresses through the stages, the spatial dimension of the tokens is reduced, while the feature dimension increases. This sequence leads to the formation of a feature pyramid, where each level represents the input at a lower spatial resolution but with larger feature dimensions.

stage	output	ResNet-50	BoTNet-50
c1	112×112	$7 \times 7, 64$, stride 2 3×3 max pool, stride 2	$7 \times 7, 64$, stride 2 3×3 max pool, stride 2
c2	56×56	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
c3	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
c4	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
c5	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ \textbf{MHSA}, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
# parameters		25.5×10^6	20.8×10^6

TABLE I: Comparative table between ResNet-50 and BoTNet-50 with input dimensions of 224×224.

The **Swin Transformer Block** (STB) encapsulates the MHSA mechanism. To allow the learning of long range dependencies, at every stage in Swin-T architecture, there are two consecutive STBs except in Stage 3, where there are 6 STBs in tandem. The first block applies a regular windowed configuration (W-MHSA) while the subsequent block adopts a windowing configuration that is shifted from that of the preceding layer (SW-MHSA, Shifted W-MHSA), in which the windows are displaced by $(\frac{M}{2}, \frac{M}{2})$ pixels. The window size is set to 7×7 patches $(M = 7)$.

C. Hybrid CNN and Self Attention model: BoTNet-50

The **Bottleneck Transformer Network (BoTNet)** introduced by [10] integrates a self-attention mechanism with a ResNet architecture [13]. This is achieved by replacing traditional Bottleneck blocks with Bottleneck Transformer (BoT) blocks, where the 3×3 convolution is substituted with a MHSA module [10], as shown in Figure 2. In BoTNet, unlike the Swin-T, the MHSA is applied globally, mirroring the traditional ViT. Consequently, the computational complexity scales quadratically with input size. However, placing BoT blocks in the network's deeper layers, where feature maps have reduced spatial dimensions, ensures efficient integration of MHSA. For training and testing on the ICLUS dataset, BoTNet-50, a version based on ResNet-50, was chosen to allow a direct comparison with ResNet-50 and Swin Tiny, as these three architectures have a comparable number of parameters, consistent with the state-of-the-art ResNet-based architectures. A more in-depth layer configuration of BoTNet-50 is shown in Table I.

D. Training strategy and performance metrics

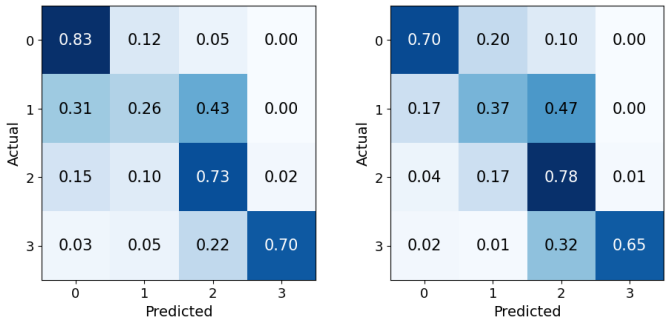
We split the ICLUS dataset into train, test and validation sets, allocating 60%, 20% and 20% of the patients respectively. The division is not made at frames or videos level but rather

	from Scratch			Pretrained		
	ResNet-50	**BoTNet-50**	**Swin Tiny**	**ResNet-50**	**BoTNet-50**	**Swin Tiny**
AUROC	0.7884	0.8057	0.7055	0.8017	-	**0.8311**
Accuracy	0.6013	0.5974	0.4649	0.6388	-	**0.6594**
F1-Score	0.5896	0.6025	0.4682	0.6280	-	**0.6513**
# parameters	25.6 M	20.8 M	27.5 M	25.6 M	-	27.5 M

TABLE III: Performance comparison of models trained from scratch (left) and pretrained on ImageNet-1K (right).

	from Scratch			Pretrained	
	ResNet-50	**BoTNet-50**	**Swin-T**	**ResNet-50**	**Swin-T**
Batch Size	64	64	64	64	64
Initial LR	0.0007	0.0039	0.0001	0.0001	0.0002
Dropout	0.1	0.2	0.3	0.1	0.1
Weight Decay	0.01	0.01	0.01	0.01	0.001

TABLE II: Best performing hyperparameters found via optimization strategy employing Grid Search.



(a) ResNet-50 (b) Swin Tiny

Fig. 3: Confusion matrices of ResNet-50 and Swin Tiny pretrained on ImageNet-1k.

at the patient level in order to ensure that videos from the same patient do not end up in different sets, which would lead to *information leakage*. We also ensured that each set maintains the same distribution of classes as the original dataset. During training, frames are uniformly resized to a standard resolution of $224 \times 224 \times 3$ pixels. Following resizing, frames undergo conversion and normalization using mean and standard deviation values pre-computed from the training set images. To promote better generalization, we adopted on-line data augmentation with the following transformations: rotation ($\pm$ 20°), scaling (max 50%), translation ($\pm$ 15%), horizontal flip (p=50%), contrast and brightness distortion ($\pm$ 30%).

In the training pipeline, an optimization strategy employing *Grid Search* is used to find the optimal values of critical hyperparameters such as Learning Rate (LR), Dropout and Weight Decay (Table II). Learning Rate values are sampled from a log-uniform distribution within the range of 10^{-4} to 0.01. The Batch Size is fixed at 64 for all models. At the end of the optimization process, the configuration that performs best on the validation set is selected for an extended training period up to a maximum of 100 epochs with early stopping. As loss function we use *Categorical Cross Entropy* with a weighting system to address dataset's unbalanced class distributions. The models are trained using various optimization algorithms, including *Stochastic Gradient Descent (SGD)* with a momentum of 0.9, *Cosine Annealing* with warm restarts [14] and *AdamW*.

Positional encoding for the MHSA modules utilizes a relative positional bias [10], [15], calculated from the relative distance between positions in input sequences, a method better suited for vision tasks [10].

The methods employed in this study involve both training from scratch and transfer learning techniques. The performance of the models are measured in terms of AUROC (Area Under ROC), Accuracy and F1-Score, with the latter serving as the primary metric due to the dataset's class imbalance.

The experiments were performed using PyTorch on a cluster equipped with a NVIDIA GeForce RTX 2080 Ti dedicated GPU. The code is publicly available as an online repository [1].

III. Results & Discussion

The comparative analysis was conducted using an experimental setup where ResNet-50 (baseline), BotNet-50 and Swin Tiny models were trained both from scratch and using a transfer learning procedure. Table III presents the results of the winning configurations from the Grid Search process.

A. Analysis of models trained from scratch

Starting from the analysis of the results achieved training the models from scratch, ResNet-50 demonstrated solid overall performance with an F1-score of 0.5896. These results are aligned with those found in [3] and [4]. BotNet-50, which incorporates a self-attention mechanism into the last stage of ResNet-50, surpassed the latter in F1-score, achieving 0.6025, and demonstrated an improvement in AUROC, reaching 0.8057 suggesting a slight better discrimination capability. Instead Swin Tiny, which relies exclusively on the self-attention mechanism without incorporating convolutional layers, did not achieve comparable performance levels.

One possible reason for this discrepancy could be the absence in Transformers of certain inductive biases found in CNNs, which capture local spatial relationships and image pattern hierarchies. Convolutions introduce architectural biases

[1] https://github.com/andrian-melnic/lung-ultrasound-self-attention

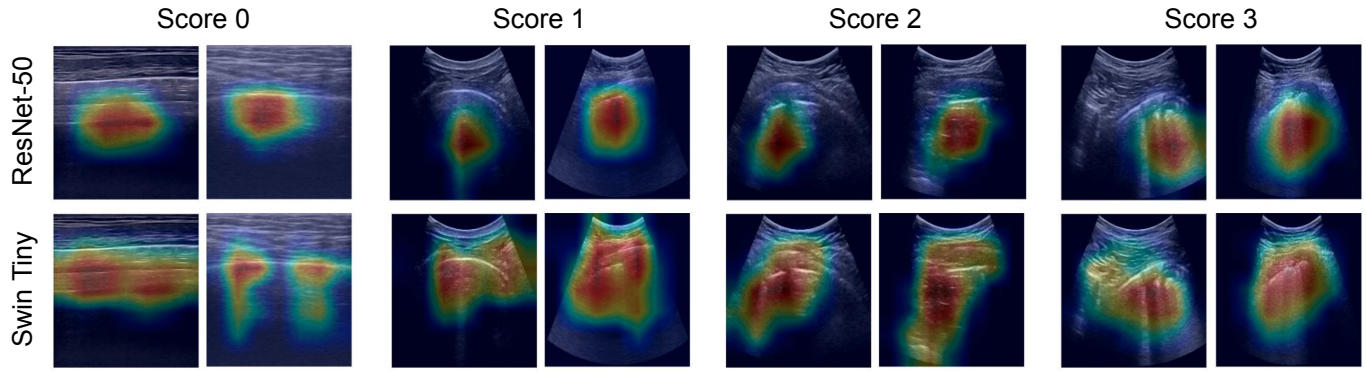


Fig. 4: GradCAM activation maps for pretrained ResNet-50 and Swin Tiny models, with two samples provided for each score, highlighting Swin Tiny's ability to capture long-range diagnostic features.

like translation equivariance, guiding the model to learn specific local features [5]. These biases can benefit the training with limited data, facilitating faster learning and improved generalization. However, in contexts with large amount data, these biases may limit flexibility [16]. Swin partially recovers some biases through its hierarchical approach in applying the MHSA [12], but our results suggest insufficiency in clinical datasets. Literature supports this, indicating that Transformers relying solely on self-attention may require larger datasets and longer training periods to effectively learn complex relationships without CNN biases [5], [11].

Frozen Layers	F1-Score	Accuracy	AUROC	Parameters
None	0.6513	0.6594	0.8311	27.5 M
Attention (MHSA)	0.6425	0.6454	0.8241	18.9 M
Feed Forward	**0.6503**	**0.6497**	**0.8283**	**10.2 M**
Swin Block 1-4	0.6387	0.6433	0.8268	26.3 M

TABLE IV: Ablation study on Swin-T Layer Freezing, investigating the impact on the number of trainable parameters and its consequent effects on performance metrics.

B. Analysis of pretrained models

Moving on to the results about transfer learning procedure, all the models were pretrained on ImageNet-1K. Unfortunately, BotNet-50 is not included in the comparison due to the unavailability of a public pretrained model. Therefore, in this case the analysis focuses on comparing ResNet-50, as baseline reference, with Swin Tiny. The optimization algorithm *AdamW* was preferred over SGD as it was used for training these architectures on the ImageNet dataset. As expected, both models have significantly benefited from transfer learning approach, showing improvements in both Accuracy and F1-Score. For ResNet-50, the scores are respectively 63.88% and 62.80%. The process had greater effects with Swin Tiny, which achieved 65.94% and 65.13% respectively, surpassing ResNet-50 and highlighting the effectiveness of transfer learning even for self-attention based models. As shown in Figure 3, Swin Tiny tends to erroneously classify healthy samples as classes 1 or 2. This inclination could lead to an increase in false positives, which, in the clinical context of COVID-19, would necessitate greater caution and further assessments. There is also a 5% decline in performance for the minority class 3. It is plausible that the minority status of class 3 had a bigger influence on Swin compared to ResNet-50 due to the it's lack of inductive biases. Hence, it may require more substantial data augmentation or further tuning of the learning strategy to address these shortcomings effectively.

Overall, transfer learning allowed better classification performance for Swin Tiny. Despite its convolution-free nature and slightly lower performance for the minority class 3, Swin outperformed ResNet-50 across all metrics as shown in Table III. It is also highlighted that the usage of a pretrained MHSA model on larger datasets can mitigate the lack of inductive biases, enabling them to learn from more restricted datasets while maintaining high generalization capabilities.

C. Qualitative analysis using GradCAM method

This section compares heatmaps from the ResNet-50 and Swin Tiny models trained on the ICLUS dataset, implementing the GradCAM method as described in [17]. This technique highlights the areas of input images that influence the model's predictions, providing insights into how each model interprets lung ultrasound images. Heatmaps were generated on frames not used in the training process. Two examples are presented for each class of the problem in Figure 4. ResNet-50 tends to focus on local textures and contours, primarily identifying artifacts such as B-lines and white lung, while often overlooking the pleural region. In contrast, Swin Tiny exhibits a more holistic approach, detecting complex structures while also evaluating the continuity of the pleural line. Directly comparing the GradCAMs of the two architectures, Swin Tiny appears to have greater sensitivity in detecting complex structures compared to ResNet-50. Eventually, Swin's heatmaps cover broader contexts, which can be beneficial for identifying severe conditions where spatial relationships are crucial.

D. Ablation study on Layer Freezing in Swin-T

Inspired by [18], we explored the impact of freezing specific Swin layers on model performance. Training involved selectively freezing either the Feed Forward or MHSA layers. Additionally, similar to ResNet, we froze the first four less deep Swin blocks. As shown by Table IV, freezing different layers has effect on classification performance. With no layers frozen, the model trains all 27.5 million parameters, achieving an F1-Score of 0.6513. By freezing only the MHSA layers, there is a less then 1% decrease in F1-score performance, with about 8 million fewer parameters to update. However, freezing the feed-forward layers results in performance comparable to the reference model, while significantly reducing training complexity. This suggests that the effective representations of context and relationships between parts of the image, generated by the MHSA modules, may be more crucial for performance than the capacity of the feed-forward layers to process the attention maps themselves. On the other hand, freezing the initial Swin blocks offers no significant advantages.

IV. Conclusions

This study investigates the potential of Transformer methodologies, specifically the Swin Tiny and BotNet architectures, in delivering superior performance in LUS image classification, filling the gap of their utilization in this domain reliant on CNN-based methods. BoTNet-50, by integrating MHSA within ResNet-50, outperforms the baseline ResNet-50 in terms of performance and efficiency. Swin Tiny struggles without pretraining, emphasizing the significance of transfer learning. Results from the application of transfer learning confirmed its effectiveness in adapting pretrained model weights to specialized datasets, leading to notable improvements for Swin Tiny in both accuracy and F1-Score. Ultimately, this approach enabled Swin Tiny to outperform ResNet-50 across all metrics. GradCAM analysis reveals Swin's superior sensitivity in global features detection and ablation studies revealed that freezing Swin feedforward layers maintains performance with reduced complexity. This milestone marks the first successful integration of Transformer-based architectures in LUS classification. Our findings suggest that a pure MHSA Transformer can outperform CNNs if trained on large datasets or with crucial pretraining, offering a new approach to LUS analysis challenges. As a future work, research could focus on developing more effective methods for handling data limitations, exploring new Vision Transformer models and adopting cross-validation methods to enhance results strength, which could significantly advance LUS image classification.

References

[1] Antonello D'Andrea, Giovanna Di Giannuario, Gemma Marrazzo, Lucia Riegler, Donato Mele, Massimiliano Rizzo, Marco Campana, Alessia Gimelli, Georgette Khoury, and Antonella Moreo. L'imaging integrato nel percorso del paziente con COVID-19: dalla diagnosi, al monitoraggio clinico, alla prognosi. *Giornale Italiano di Cardiologia n. 5*, 21, 2020.

[2] Gino Soldati, Andrea Smargiassi, Riccardo Inchingolo, Danilo Buonsenso, Tiziano Perrone, Domenica Federica Briganti, Stefano Perlini, Elena Torri, Alberto Mariani, Elisa Eleonora Mossolani, Francesco Tursi, Federico Mento, and Libertario Demi. Proposal for international standardization of the use of lung ultrasound for patients with covid-19. *Journal of Ultrasound in Medicine*, 39, 2020.

[3] Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris Huijben, Nishith Chennakeshava, Federico Mento, Alessandro Sentelli, Emanuele Peschiera, Riccardo Trevisan, Giovanni Maschietto, Elena Torri, Riccardo Inchingolo, Andrea Smargiassi, Gino Soldati, Paolo Rota, Andrea Passerini, Ruud J.G. Van Sloun, Elisa Ricci, and Libertario Demi. Deep learning for classification and localization of covid-19 markers in point-of-care lung ultrasound. *IEEE Transactions on Medical Imaging*, 39, 2020.

[4] Oz Frank, Nir Schipper, Mordehay Vaturi, Gino Soldati, Andrea Smargiassi, Riccardo Inchingolo, Elena Torri, Tiziano Perrone, Federico Mento, Libertario Demi, Meirav Galun, and Yonina C. Eldar. Integrating domain knowledge into deep networks for lung ultrasound with applications to covid-19. *IEEE Transactions on Medical Imaging*, 41, 2022.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR 2021 - 9th International Conference on Learning Representations*, 2021.

[6] Lumin Xing, Wenjian Liu, Xiaoliang Liu, and Xin Li. An enhanced vision transformer model in digital twins powered internet of medical things for pneumonia diagnosis, 2023.

[7] Oğuzhan Katar, Ozal Yildirim, and Yeşim Eroğlu. Vision transformer model for efficient stroke detection in neuroimaging. *2023 4th International Informatics and Software Engineering Conference (IISEC)*, pages 1–6, 2023.

[8] S. Tummala, Seifedine Kadry, Syed Ahmad Chan Bukhari, and Hafiz Tayyab Rauf. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Current Oncology*, 29:7498 – 7511, 2022.

[9] Szymon Płotka, Michal K. Grzeszczyk, Robert Brawura-Biskupski-Samaha, Paweł Gutaj, Michał Lipa, Tomasz Trzciński, and Arkadiusz Sitek. Babynet: Residual transformer module for birth weight prediction on fetal ultrasound video. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13434 LNCS, 2022.

[10] Aravind Srinivas, Tsung Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021.

[11] Behnaz Gheflati and Hassan Rivaz. Vision transformers for classification of breast ultrasound images. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2022-July, 2022.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, 2016.

[14] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

[15] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. pages 464–468, 2018.

[16] Yunsung Lee, Gyuseong Lee, Kwang seok Ryoo, Hyojun Go, Jihye Park, and Seung Wook Kim. Towards flexible inductive bias via progressive reparameterization scheduling. pages 706–720, 2022.

[17] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019.

[18] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers, 2022.