DOI: 10.21917/ijivp.2023.0430

AUTOMATED SEVERITY SCORING OF COVID-19 CT SEQUENCES USING SPACE-TIME TRANSFORMERS

Mercy Ranjit[1], Gopinath Ganapathy[2], K. Vishnu Vardhan Reddy[3] and Sahana M. Prabhu[4]

[1,2]Department of Computer Science, Bharathidasan University, India
[3]Navodaya Medical College Hospital and Research Centre, India
[4]Robert Bosch Engineering and Business Solutions, Bengaluru, India

Abstract

Automatic diagnosis of Covid-19 lung complications from Computerized Tomography (CT) scans is an increasingly important research topic. In this rapidly developing area of Covid detection from medical image sequences, it is noted that most prior literature has focused on binary classification to detect diseased versus healthy cases from single X-ray or CT image. In this paper, we advance a step further by presenting a comprehensive framework for automated classification of the severity of lung infection (mild, moderate and severe) from CT sequences of confirmed Covid cases. We consider the sequence information for automation because in practice, the medical experts look at the CT sequence to score the severity of infection. We have collected a new lung CT sequence dataset at various stages of Covid infection from Indian patients. This dataset has been scored in terms of the severity of each lung lobe by experts in the field. We present a novel application of space-time transformers for CT sequences and achieve 93.3% accuracy for sequence level and 99% accuracy for patient-level, for multi- class classification of severity classes.

Keywords:

Artificial Intelligence, Computerized Tomography, Covid-19, Deep Learning, Image Classification, Vision Transformer

1. INTRODUCTION

Computerized Tomography (CT) is an advanced and expensive imaging modality that consists of a sequence of X-ray radiology images captured at different angles around the body. Covid-19 infection is characterized by respiratory issues affecting the lungs. Covid testing using images is becoming increasingly vital for lung infection symptoms. It is also used as a follow-up to Reverse transcription polymerase chain reaction (RT-PCR) testing. The CT scan images of lungs represent slices or different views of the lung tissues. Although there has been progress in detection of Covid versus normal from CT scans [1] [2], there is lack of research and labeled dataset for classification of different stages of Covid severity from CT sequences.

Lung scans of Covid patients exhibit white lesions of varying density, named as ground glass opacities (GGO) which appear hazy and less opaque in the initial stages. GGO's in lung tissues of infected patients appear white in the CT images whereas the lung tissue for a healthy person with free air flow in lungs appear black. These lesions are called consolidations if they appear more solid as opaque white patches and occlude the bronchial structures in the image. Even in early stages of severity, ground glass opacities (GGO) appear in the lung [3], therefore the presence of GGO is an important evidence of lung infection. As the severity increases, so does the GGO appearance in the air- filled portions of lung regions. Patients can exhibit: (i) only GGO, (ii) only consolidations, or (iii) combination of GGO as well as consolidations. Depending on the stage of the infection, the type of treatment also varies, so it is vital to correctly assess severity in an automatic manner.

A lung is composed of five lobes, two on the left and three on the right side: (i) Left Upper (LU), (ii) Left Lower (LL), (iii) Right Upper (RU), (iv) Right Middle (RM), (v) Right Lower (RL) lobes. The severity of the lung disease is typically labeled via visual inspection by experts. The infection is labeled on every lobe in a range from 0 to 5 (where 0 is healthy and 5 is most severe) and then take the sum to get the total score. Therefore, the total severity score (TSS) ranges from 0 to 25 overall for the five lobes of the lung. From our discussions with medical experts on the best approach to do multi-class classification, we have grouped the scores into three categories: mild (0<TSS<2), moderate (2<=TSS<8), and severe (8<=TSS<=25) [19]. The TSS scores for every lobe are determined from a sequence of CT images where that lobe is visible.

Among the classification methods for image sequences, most of the networks are based on convolutional neural networks (CNN), but the focus has shifted to transformers very recently [4]. State-of-the-Art (SOTA) performance was achieved in the field of natural language processing (NLP) using Transformer architecture for handling sequential text data [5]. There was a perceived gap in research to efficiently apply transformer architecture to handle image sequences. A comprehensive framework for applying transformers to images, viz. Vision Transformer (ViT) proposed in [4] was a breakthrough. Thereafter, there has been a plethora of applications for transformer architectures for both image and video data. The TimeSformer family of architectures introduced in [6], is an adaptation of visual transformers to video applications, by harnessing spatio-temporal feature learning from the sequence. The performance of TimeSformers has been compared to 3D convolutional neural networks in [6] and the former proved to be more efficient.

The aim of this research is to quantify the severity of Covid-19 infection from Lung CT Sequences. This will help identify more serious patients and prioritize treatment based on severity scores. Rather than taking a segmentation approach which is more computationally expensive, we prefer solving the severity problem as a classification task. Our approach mimics the medical expert's procedure of severity scoring, by observing the sequences of lung lobes rather than relying only on stand-alone images. We explore the effectiveness of Transformer networks when applied to CT sequential image data. Transformer architecture has been proven effective for text sequences and for videos. This paper presents a novel application of TimeSformer for structure-varying medical sequences, which has not been previously addressed. We have collected a new CT dataset from an Indian hospital, taken from 102 patients. This dataset contains

7797 CT images from 102 patients. 7090 overlapping image sequences were constructed from these images with a sequence length of eight. Each lobe of the lungs in the CT sequences has been labeled for severity score by experts in the field.

2. RELATED WORKS

For computer vision tasks, attention has been applied in conjunction with convolutional neural networks (CNN) or used to replace some layers while keeping their overall structure in place. Wholly replacing CNNs with purely attention-based transformer architecture and applying it directly to image patches for classification was presented recently in [4].

Following the successful performance of transformers first in natural language processing (NLP) and then for image classification tasks, researchers applied transformers as a building block for video processing in [6]. Intuitively, this extension seems straightforward as a video is a sequence of images. However, Vision Transformers from [4] cannot be directly applied since we need to consider not only space but also time. Rather than processing the frames as isolated images, we need to incorporate attention that accounts for the variation between consecutive frames.

A convolution-free approach to video classification, viz. TimeSformer, built exclusively on self-attention over space and time, was presented in [6]. It adapts the standard Transformer architecture to video by enabling spatio-temporal feature learning directly from a sequence of frame-level patches. The experimental study in [6] compares five different self-attention schemes and concludes that divided attention, where temporal attention and spatial attention are separately applied within each block, leads to the best video classification accuracy.

Typical effects of Covid-19 in lung images include visible lesions with ground-glass opacities (GGO). Imaging interpretations are vital for diagnosis and monitoring of disease progression and the evaluation of treatment [1]. The clinical guidelines for radiologists to interpret CT scans for manually diagnosing Covid-19 were detailed in [3].

Image datasets in [13] are small sets of CT scans and X-rays, collected primarily from China, USA, Italy, and Japan. There is a lack of labeled CT scan data particularly for severity level of Covid-19 patients from India. A recent survey in [15] states that radiologists can diagnose Covid-19 more accurately with CT scans, rather than other medical imaging modalities such as X-ray and Ultrasound scans.

An approach for detection and severity scoring for disease monitoring using X-ray images for Covid-19 patients was presented in [7]. A simple CNN architecture with stochastic pooling was used for chest CT-based Covid-19 diagnosis in [17]. A weakly-supervised deep learning method for detecting Covid-19 infection from CT images was proposed in [8]. Localization of ground-glass opacities (GGO) using class-activation maps (CAM) was presented in [9]. A deep 3D-CNN architecture consisting of 121-layers, named De-COVID19-Net was presented in [10], which uses 3D convolution to synthesize spatial information of the CT image. An overview of artificial intelligence methods, particularly deep learning for the detection of Covid-19 from medical imaging data was provided in [11] and [13]. These image-based methods used deep learning algorithms, specifically CNNs for Covid-19 detection and most of them operated at an image-level.

Severity quantification of COVID- 19 on X-Ray images (not CT scans) using Vision Transformer's attention as the backbone for a segmentation pipeline was introduced in [16]. These attention-based mechanisms were also at image level.

Automatic segmentation of infected regions of lung infection from CT scan was proposed in [2] and [14] using InfNet architecture and a multi encoder-decoder network called ConvSegNet respectively. A pipeline of commonly used convolutional neural network (CNN) architectures for classification and segmentation, namely, ResNet-50 and U-net, were used in [12]. These approaches require pixel-wise segmentation data which involves a lot of labeling effort. There is a need for multi-class classification-based detection and localization of the disease which can lead to faster diagnosis as it involves less computational complexity and data labeling efforts. Our work takes a classification approach and uses clinically relevant Total Severity Score (TSS) as the basis for evaluation.

We present a novel approach of determining the severity class from CT sequences of the lung lobes instead of single images as in previous works in concurrence with the approach taken by medical experts. Space-Time attention-based classification of CT sequences has also not been addressed in any of the earlier works.

3. PROPOSED METHOD

We first present the detailed description of the dataset and present sample CT scan slices and tables of severity scores of patients in each mild, moderate, and severe category.

3.1 DATASET DESCRIPTION

We have collected a new CT dataset from 102 patients, consisting of 7090 sequences, constructed from 7797 images, and labeled for lobe-level severity scores by medical experts. Note that all the CT Sequences are from Covid-19 positive patients only with varying levels of severity and kindly provided for research purposes by Navodaya Medical Institute, Raichur, India.

Table.1. Multi-class classification of Total Severity Score (TSS).

Severity category	Total Severity Score
Mild	0<TSS<2
Moderate	2<=TSS<8
Severe	8<=TSS<=25

Table.2. Sample Patients with severe Covid-19 infection

RUL	RML	RLL	LUL	LLL	TSS
5	5	4	4	4	22
4	4	5	4	5	22
4	4	5	4	5	22
3	5	5	5	5	23
5	5	5	4	5	24

The severity category is based on the total severity score (TSS) which is computed from the five lobes, the severity of each lobe

ranging from 1 to 5 as shown in Table.1. To give a split of TSS scores of samples, we have presented the data of five patients in the category of Severe and Moderate classes in Table.2 and Table.3, respectively. The TSS scores are the sum of lobe-level scores and categorized into three classes as in Table.1 by consulting with doctors and in reference to [19].

Table.3. Sample Patients with moderate Covid-19 infection

RUL	RML	RLL	LUL	LLL	TSS
1	1	3	1	1	7
2	1	2	1	1	7
1	1	3	1	1	7
1	1	2	1	2	7
1	1	2	1	2	7

The number of images in a CT lung sequence depends on the height of the patient and typically varies in the order of 120 to 180 images per sequence. The order of appearance of the lung lobes in the sequence is not consistent across all patients as some are top-bottom and others are in bottom-top direction. Therefore, the slices in which a particular lobe is visible for a patient (e.g., the left upper lobe) will not match with the slices for which the same lobe appears for another patient. To combat the varying sequence length problem and to provide coverage for the whole CT sequence for a patient, the input dataset was prepared by creating overlapping sequences with a sequence length of 8. Considering a lung sequence of length 120, the first 8 slices form one sequence, slices 2 to 9 forms the next, 3 to 10 the next, etc. This pre-processing handles the problem of varying CT lengths for different patients and ensures that the model learns from all the lung regions. Shorter sequence length was also chosen to accommodate the changes in the biological structure of lungs which occur every few slices. This will avoid the unnecessary attention computations across slices from different lobes. As the model learns from the sequences of all the lung lobes with overlapping sequences, it gains the ability to score for each lobe separately as well. It is not strictly restrictive towards the starting and ending positions of the sequence. This method also efficiently keeps the sequence length under check to deal with the time and space complexities of the space-time attention computations in transformers.

The CT scans of patients belonging to each of the categories are presented next, showing various views of the lung portions. Fig. 1 presents a cross-section of CT scans pertaining to Severe cases, while Fig. 2 and Fig. 3 show samples of Moderate and Mild cases, respectively. It can be observed that Severe cases in Fig. 1 have more solid consolidations when compared to Figures 2 and 3, which exhibit GGOs.

Each slice of the CT sequence captures a different cross-section of the lung and the GGOs and consolidations appear more prominent in the lower portions of the lungs which is typically found to be most infected across patients. For arriving at a severity score diagnosis, the doctor manually does a visual examination of all the slices of the CT scans to assess the severity at every lobe of the lung. This is a time-consuming process as the doctor must examine sequence of 120 to 180 images per patient and will surely benefit from automation.

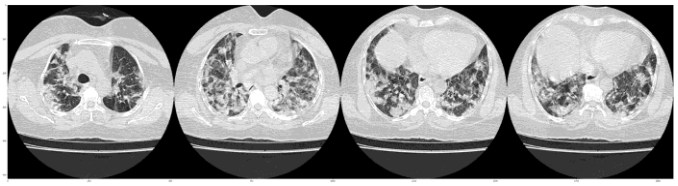

Fig.1. Clinical Type – Severe (8<=TSS<=25)

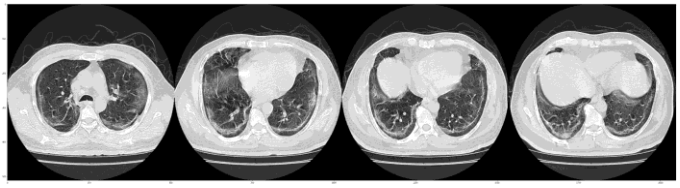

Fig.2. Clinical Type – Moderate (2<=TSS<8)

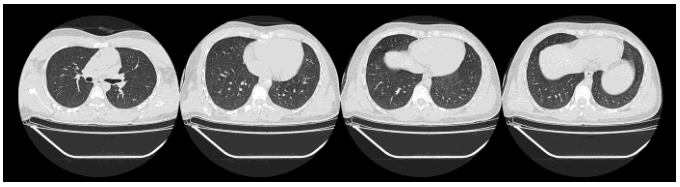

Fig.3. Clinical Type – Mild (0<TSS<2)

3.2 SPACE-TIME TRANSFORMER

The TimeSformers algorithm, which is based on space-time transformer architecture, is adapted from [6] for our application and described in detail. But we shall first give a brief overview of the working of transformer architecture. Self-attention is the fundamental operation of transformer, and it is a sequence-to-sequence weighted average operation. The mathematical operation is the dot product of vectors to which we apply softmax to map the values to [0,1] to compute the self-attention weights. To apply self-attention, we first compute Query, Key, and Value vectors represent the following operations on the input vector:

Query: Input is compared to every other vector to establish the weights for its own output. ii) Key: Input is compared to every other vector to establish the weights for the output of the j^{th} vector. iii) Value: Input is used as part of the weighted sum to compute each output vector once the weights have been established.

Multi-head self-attention is a small number of copies of the self-attention mechanism applied in parallel, each with their own key, value, and query transformation. The transformer building block applies, in sequence: a self-attention layer, layer normalization, a feed forward layer and another layer normalization. Residual connections are also added after normalization. ViT (Vision Transformer) architecture divides the image into patches and then computes self-attention between the patches of the same image.

TimeSformers captures the attention for spatial context within the same image as well as across slices. Among the five variants of TimeSformers proposed in [6], the architecture that has achieved the best results is Divided Space-Time Attention. Given a slice and one of its patches as a query (q), it first computes the spatial attention over the rest of patches, termed as keys, (k). Later, the temporal attention is computed in the same patch of the

query, in all the corresponding patches across slices. This process fetches the most attentive value (v) patch for the query patch q across patches in the same location across slices and across all patches in the same slice.

The Divided Space-Time Attention architecture (shown in Fig. 4) independently applies temporal attention and spatial attention for each slice one after the other. Fig. 5 shows the space neighborhood and the time neighborhood for the query patch wherein blue patch is query, orange patches are space neighbors and green patches are time neighbors. White patches are ignored for self-attention computations.

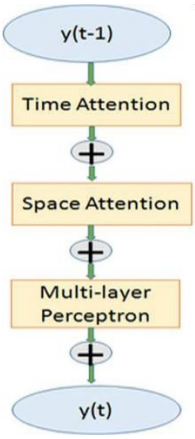

Fig.4. Space Time Transformers

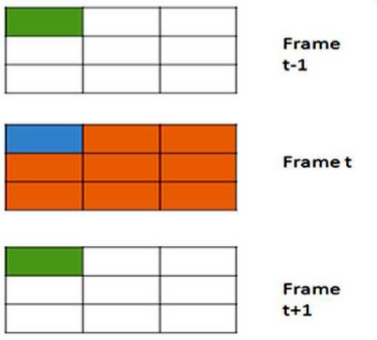

Fig.5. Patch neighborhood for space-time model

Each CT sequence has 8 slices, and each slice is divided into flat patches $p(s,t)$ of size P^2C where P is the patch-size and C is the channel size, which is 16 and 3, respectively. This results in an input dimension of 768 per patch. The number of patches S in a slice is given by $(H * W)/P^2$ where P is the patch size and H, and W are the height and width of the slice. The experiments used an input resolution of (224,224) resulting in 196 patches per slice. Each patch $p(s,t)$ is mapped to an embedding vector $z_0(s,t)$, which is then given as input to the transformer. This is computed using a learnable embedding matrix E of dimension (P^2C, D) which for our experiment was (768, 128). A trainable positional embedding e is added to the concatenated sequence of projections to add a spatial representation of each patch within the sequence. The computation of the transformer input is as given below:

$$\mathbf{z_0}(s,t) = E.\mathbf{p}(s,t) + \mathbf{e}(s,t) \quad (1)$$

where $\mathbf{p}(s,t)$ with $s = 1, 2...S$ represents the flattened patches in a slice (S=196 for our experiment). The CT slices are represented by t=1, 2...T, where T=8. **e** represents the position embedding that encodes the position of each patch. The positional embedding **e** is of dimension $(S + 1, D)$ where unity corresponds to the classification token which is added along with the patches. The input to the transformer is of dimension $(T, S + 1, D) = (8, 196 + 1, 128)$.

The input $\mathbf{z_0}(s,t)$ is projected to query, key and value matrices using the below formula to help calculate the self-attention values. LN stands for layer normalization, l is the index over the attention blocks L=1, a is the index over the attention heads A=4, q is the query matrix, k represents the key matrix and v represents the value matrix.

$$q_{(s,t)}^{(l,a)} = W_Q^{(l,a)} LN\left(z_{(s,t)}^{(l-1)}\right) \quad (2)$$

$$k_{(s,t)}^{(l,a)} = W_K^{(l,a)} LN\left(z_{(s,t)}^{(l-1)}\right) \quad (3)$$

$$v_{(s,t)}^{(l,a)} = W_V^{(l,a)} LN\left(z_{(s,t)}^{(l-1)}\right) \quad (4)$$

The attention weights over the slices $1 \ldots T$ are calculated using the below formula:

$$\alpha_{(s,t)}^{(l,a)slices} = soft\max\left(\frac{q_{(s,t)}^{(l,a)^T}}{\sqrt{D_h}\left[k_{(0,0)}^{(l,a)}\left\{k_{(s,t')}^{(l,a)}\right\}_{t'=1,\ldots,T}\right]}\right) \quad (5)$$

D_h=32 represents the latent dimension which is D/A (128/4) for our experiment. The self-attention values for the slice-based attention weights from Eq.(5) and value vector v are calculated as below:

$$t_{(s,t)}^{(l,a)slices} = \alpha_{(s,t),(0,0)}^{(l,a)slices} v_{(0,0)}^{(l,a)} + \sum_{t'=1}^{T} \alpha_{(s,t),(s,t')}^{(l,a)slices} v_{(s,t')}^{(l,a)} \quad (6)$$

Then, the concatenation of these vectors from all attention heads is projected as in Eq.(7), the residual connection from the encoding of previous layer is also added.

$$z_{(s,t)}'^{(l)slices} = w\begin{bmatrix} t_{(s,t)}^{(l,1)} \\ \vdots \\ t_{(s,t)}^{(l,A)} \end{bmatrix} + z_{(s,t)}^{(l-1)} \quad (7)$$

The encoding based on slice attentions from Eq.(7) is then fed back for spatial attention weight computation using a new set of query, key, and value matrices as below:

$$\alpha_{(s,t)}^{(l,a)space} = soft\max\left(\frac{q_{(p,t)}^{(l,a)^T}}{\sqrt{D_h}\left[k_{(0,0)}^{(l,a)}\left\{k_{(s',t)}^{(l,a)}\right\}_{s'=1,\ldots,S}\right]}\right) \quad (8)$$

The spatial attention values are calculated using the attention weights from Eq.(8) and the value matrices projected using $z_{(s,t)}'^{(l)slices}$ from Eq.(7) as below:

$$s_{(s,t)}^{(l,a)space} = \alpha_{(s,t),(0,0)}^{(l,a)space} v_{(0,0)}^{(l,a)} + \sum_{s'=1}^{S} \alpha_{(s,t),(s',t)}^{(l,a)slices} v_{(s',t)}^{(l,a)} \quad (9)$$

The concatenation of the space attention values from Eq.(9) is projected using the below equation and passed through a feed forward network with residual connections from slice encodings.

$$z'^{(l)spaces}_{(s,t)} = W \begin{bmatrix} s^{(l,1)}_{(s,t)} \\ \vdots \\ s^{(l,A)}_{(s,t)} \end{bmatrix} + z'^{(l)slices}_{(s,t)} \quad (10)$$

$$z^{l}_{(s,t)} = MLP\left(LN\left(z'^{(l)spaces}_{(s,t)}\right)\right) + z'^{(l)spaces}_{(s,t)} \quad (11)$$

The encoding $z^{l}_{(s,t)}$ from the last layer corresponding to the classification token is passed through a linear layer with three output nodes to determine the severity of covid-19 infection.

4. EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETUP

We have used TimeSformer for sequence-wise classification with ViT backbone. We performed three-fold cross-validation which are split patient-wise during training using stratified split on the classes. Note that there are several sequences per patient, and while taking three-fold cross validation, we have done stratified split such that patients are not duplicated across training and testing. As the number of patients are limited, too large value of K will result in iterations that are not different, we selected K=3 given it provided representative coverage of the distribution in all three folds. The former can harness the sequence information to provide lobe-level severity if sequences from a lobe are presented to the model thereby allowing for lobe-wise scoring. We also present the patient-level accuracy from the sequence-wise for TimeSformer by using the maximum predictions for each patient. The experiments were performed using Standard_NC6s_v3 version of Azure GPU virtual machines with v100 GPU.

The dataset configuration details are presented in Table 4. The images were normalized and resized before giving it as the input to the network. Each input slice is resized to 256*256, center-cropped to standard input size of 224*224 and then divided into uniformly spaced $16 * 16$ non-overlapping patches. The slices were normalized using the mean and standard deviation of all the training sequence image channel-wise. Since the CT scans of each patient are irregular in the length of sequences (number of slices varying), we have split it into smaller, overlapping sequences of uniform sequence length. This also allows us to score on any set of sequences from the patient.

Table.4. Dataset Configuration

Configuration	Value
Images	7797
CT Sequences	7090
Image Size	224 * 224
Normalization	mean= [0.485, 0.456, 0.406] std= [0.229, 0.224, 0.225]

The details about the TimeSformer sequence configuration and training settings are presented in Table 5. Since the CT sequences were not uniform in length, and to avoid truncating thereby wasting data slices, we took overlapping sub-sequences of 8 slices. Patch size of 16x16 pixels for ViT was used in accordance with [4]. We have used the Cosine learning rate scheduler with a learning rate of 1e-4, and Gamma parameter of 0.6 and trained for 10 epochs with the batch size of 16.

Table.5. TimeSformer Configuration

Configuration	Value
No of Attention Heads	4
No. of Attention Blocks	1
Dimension of encoding vectors	128
Dimension of encoding vectors of attention heads	32
Patch Size (ViT)	16*16
Sequence Length	8
Attention Dropout	0.1
Multi-Layer Perceptron Dropout	0.1

4.2 RESULTS

We present detailed classification report for sequence transformers in Table.6. For multi-class classification, weighted average is a performance measure taking the number of samples in the three categories [18].

Table.6. Classification Report for TimeSformer (Sequence-wise)

	Precision	Recall	F1-Score	Support
Mild	0.96	0.88	0.92	1538
Moderate	0.83	0.97	0.90	1869
Severe	0.98	0.93	0.96	3683
Accuracy			0.933	7090
Macro-Avg	0.93	0.93	0.93	7090
Weighted-Avg	0.94	0.93	0.93	7090

Confusion matrix for sequence-level predictions of TimeSformer is presented in Table.7, which indicates good performance as the off-diagonal elements are very less. The misclassifications (off-diagonal elements) correspond to the CT scans of the upper-most lung-lobes. The expert doctors observed that portion of the lung to be least susceptible to Covid-19 infection when compared to the lower lobes, which means there may not be much distinction between the three classes.

Table.7. Confusion Matrix for TimeSformer (Sequence-wise)

	Mild	Moderate	Severe
Mild	1342	116	80
Moderate	45	1814	10
Severe	2	256	3425

Patient-wise accuracy was computed by taking the maximum predictions of all sequences for each patient. The average accuracy, precision and recall for patient-wise classification from all 3 folds of cross-validation are presented in Table 8. It is observed that all three performance indicators are above 95% for sequence transformers. TimeSformer provides additional context using sequence information to predict patient wise severity scores, rather than relying on a single image for a prediction. It also offers advantages for cases where lobe level differences are present for

the same patient and we need to calculate the severity level for different lobes rather than a patient-level severity class.

Table.8. Patient-level average metrics for 3-fold cross-validation

Metric	Value
Accuracy	0.99
Precision	0.97
Recall	0.99

4.3 EXPLAINABILITY OF PREDICTIONS

Explainability analysis is useful for more comprehensive diagnosis and trustworthiness of the predictions. The prediction of severe cases of Covid-19 from the ViT backbone is explained using XRAI [20], a region-based saliency method. This approach over-segments the image and evaluates the importance of each segment using Integrated Gradients based attributions. It then coalesces smaller regions into larger segments based on the attribution scores.

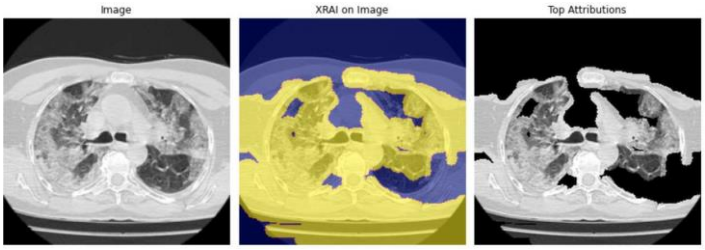

Fig.6. XRAI Attributions of a severe case of covid - CT slice from upper lung region

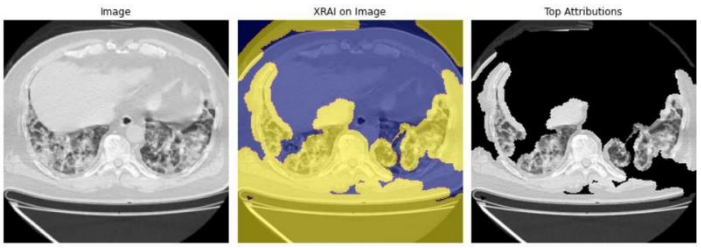

Fig.7. XRAI Attributions of a severe case of covid - CT slice from middle lung region

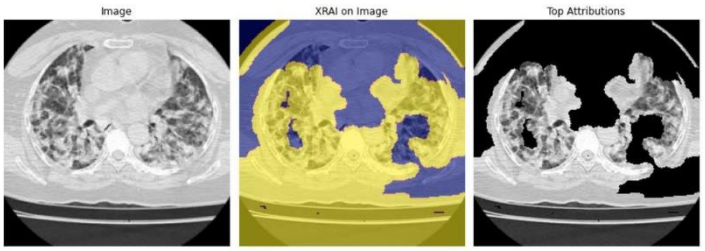

Fig.8. XRAI Attributions of a severe case of covid - CT slice from lower lung region

The Fig.6-Fig.8 shows the XRAI-based attributions generated from CT slices of severe Covid predictions pertaining to the upper, middle, and lower lung regions of different Covid patients.

The model has indeed observed the infected areas of the lungs with GGO and consolidations while predicting the severe class of infection. This paves the way for localization of the regions to pinpoint diseased tissues, without the need for explicit segmentation.

5. CONCLUSION

In this paper, we have presented space-time transformer-based attention mechanism for automatic severity classification of lung infection in Covid-19 patients from CT sequences. We have collected a new CT dataset from 102 patients in India, consisting of 7797 images, which are grouped into 7090 overlapping sequences of slice length 8. Each lung lobe has been visually examined and labeled by experts in the field. We achieved 93.3% accuracy for multi-class classification of severity scores at sequence level. Maximum predictions for sequences of each patient yields a patient-level accuracy of 99.0%. Lobe-level severity can also be inferred using this work by presenting the sequences from the different lung lobes to the model. This research can be easily extended to other medical scenarios with sequential image data.

ACKNOWLEDGMENT

We thank the Navodaya Medical Institute, Raichur, India for kindly providing the CT dataset for this study. The included studies have been approved by the ethics committee of Navodaya Medical Institute, Raichur, India.

REFERENCES

[1] Di Dong, Zhenchao Tang, Shuo Wang, Hui Hui, Lixin Gong, Yao Lu and Zhong Xue, "The Role of Imaging in the Detection and Management of COVID-19: A Review", *IEEE Reviews in Biomedical Engineering*, Vol. 14, pp. 16-29, 2020.

[2] Deng Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen and Ling Shao, "Inf-Net: Automatic Covid-19 Lung Infection Segmentation from CT Images", *IEEE Transactions on Medical Imaging*, Vol. 39, No. 8, pp. 2626-2637, 2020.

[3] T.C. Kwee and R.M. Kwee, "Chest CT in COVID-19: What the Radiologist Needs to Know", *Radiographics*, Vol. 40, No. 7, pp. 1848-1865, 2020.

[4] A. Dosovitskiy, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", *Proceedings of International Conference on Learning Representations*, pp. 1-21, 2021.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you Need", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998-6008, 2017.

[6] G. Girdhar, H. Wang and L. Torresani, "Is Space-Time Attention all you need for Video Understanding?", *Proceedings of International Conference on Machine Learning*, pp. 1007-1017, 2021.

[7] M. Frid-Adar, R. Amer, O. Gozes, J. Nassar and H. Greenspan, "COVID-19 in CXR: From Detection and Severity Scoring to Patient Disease Monitoring," *IEEE*

Journal of Biomedical and Health Informatics, Vol. 25, No. 6, pp. 1892-1903, 2021.

[8] S. Hu, W. Menpes-Smith and H. Ye, "Weakly Supervised Deep Learning for Covid-19 Infection Detection and Classification from CT Images", *IEEE Access*, Vol. 8, pp. 118869-118883, 2020.

[9] H. Alshazly, C. Linse, E. Barth and T. Marinetz, "Explainable Covid-19 Detection using Chest CT Scans and Deep Learning", *Sensors*, Vol. 21, No. 2, pp. 455-464, 2021.

[10] L. Meng, Y. Zha and J. Tian, "A Deep Learning Prognosis Model Help Alert for COVID-19 Patients at High-Risk of Death: A Multi-Center Study", *IEEE Journal of Biomedical and Health Informatics*, Vol. 24, No. 12, pp. 3576-3584, 2020.

[11] M. Jamshidi, "Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment", *IEEE Access*, Vol. 8, pp. 109581-109595, 2020

[12] O. Gozes, M. Frid-Adar, N. Sagie, H. Zhang, W. Ji and H. Greenspan, "Coronavirus Detection and Analysis on Chest CT with Deep Learning", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2020.

[13] M.N. Islam, T.T. Inan, S. Rafi, S.S. Akter, I.H. Sarker and A.N. Islam, "A Systematic Review on the Use of AI and ML for Fighting the COVID-19 Pandemic," *IEEE Transactions on Artificial Intelligence*, Vol. 2, No. 2, pp. 139-156, 2021.

[14] T. Mahmud and S.Y. Kung, "Cov-SegNet: A Multi Encoder-Decoder Architecture for Improved Lesion Segmentation of COVID-19 Chest CT Scans", *IEEE Transactions on Artificial Intelligence*, Vol. 2, No. 3, pp. 283-297, 2021.

[15] S. Latif, A. Razi, A. Weller and J. Crowcroft, "Leveraging Data Science to Combat Covid-19: A Comprehensive Review", *IEEE Transactions on Artificial Intelligence*, Vol. 2, No. 2, pp. 139-156, 2021.

[16] G. Kim, "Severity Quantification and Lesion Localization of COVID-19 on CXR using Vision Transformer", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2021.

[17] Y.D. Zhang, "A Five-Layer Deep Convolutional Neural Network with Stochastic Pooling for Chest CT based Covid-19 Diagnosis", *Machine Vision and Applications*, Vol. 32, No. 1, pp. 1-13, 2021.

[18] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks", *Information Processing and Management*, Vol. 45, No. 4, pp. 427-437, 2009.

[19] A. Kumar, "COVID19 Severity Scoring from CT: Primer for Radiologists", Available at https://medium.com/predible/covid19-severity-scoring-from-ct-primer-for-radiologists-930536dfade5, Accessed at 2023.

[20] A. Kapishnikov and M. Terry, "XRAI: Better Attributions Through Regions", *Proceedings of International Conference on Computer Vision*, pp. 4942-4951, 2019.