IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 January 2023, accepted 14 March 2023, date of publication 22 March 2023, date of current version 27 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3260632

APPLIED RESEARCH

Region-of-Interest Aware 3D ResNet for Classification of COVID-19 Chest Computerised Tomography Scans

SHUOHAN XUE, (Student Member, IEEE), AND CHARITH ABHAYARATNE, (Member, IEEE)
Department of Electronic and Electrical Engineering, The University of Sheffield, S1 3JD Sheffield, U.K.
Corresponding author: Charith Abhayaratne (c.abhayaratne@sheffield.ac.uk)

ABSTRACT Coronavirus disease 2019, commonly known as COVID-19, is an extremely contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Computerised Tomography (CT) scans based diagnosis and progression analysis of COVID-19 have recently received academic interest. Most algorithms include two-stage analysis where a slice-level analysis is followed by the patient-level analysis. However, such an analysis requires labels for individual slices in the training data. In this paper, we propose a single-stage 3D approach that does not require slice-wise labels. Our proposed method comprises volumetric data pre-processing and 3D ResNet transfer learning. The pre-processing includes pulmonary segmentation to identify the regions of interest, volume resampling and a novel approach for extracting salient slices. This is followed by proposing a region-of-interest aware 3D ResNet for feature learning. The backbone networks utilised in this study include 3D ResNet-18, 3D ResNet-50 and 3D ResNet-101. Our proposed method employing 3D ResNet-101 has outperformed the existing methods by yielding an overall accuracy of 90%. The sensitivity for correctly predicting COVID-19, Community Acquired Pneumonia (CAP) and Normal class labels in the dataset is 88.2%, 96.4% and 96.1%, respectively.

INDEX TERMS COVID-19 diagnosis, transfer learning, 3D ResNet, CT scans, region-of-interest.

I. INTRODUCTION

Coronavirus disease 2019, commonly known as COVID-19, is an extremely contagious disease [1], [2] caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Since the first case was reported in December 2019 it has led to an ongoing pandemic claiming millions of lives, adding colossal pressure to healthcare systems globally [3] and a huge negative impact on the world economy. Despite the vaccination efforts in full force and many countries have cancelled public COVID rules, vanquishing COVID-19 and achieving herd immunity are still theoretically unreachable in the foreseeable future [4], [5] due to its possible reinfection rate [6], [7]. Recently, many countries have encountered COVID resurgence [8], [9]. Accordingly, there is a significant demand for rapid and accurate testing for the virus. The current gold standard for detecting the virus is the reverse-transcription polymerase chain reaction (RT-PCR) test. However, there are several limitations of the RT-PCR tests such as a false negative rate and long reaction time [10]. Hence, supplementary diagnosis approaches based on imaging and artificial intelligence (AI) technology could be utilised [11].

Deep learning-based AI technology has gained tremendous success in computer vision. Powered by the hardware advancements in Graphical Processing Units (GPUs), deep learning algorithms have attained state-of-the-art accuracy in learning deep and complex visual features since the ImageNet Large Scale Visual Recognition Challenge in 2012 [12]. Since then, deep learning has gained widespread popularity and achieved promising performance in medical imaging applications such as diagnostic classification, lesion and tumour segmentation or localisation [13], [14].

As imaging data became available, there has been an interest in applying deep learning to understand Computerized tomography (CT) scan images for diagnosis and further

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

analysis of COVID-19 infections. More details and a review of these works are presented in Section II. CT scans are regarded as one of the most conventional medical imaging techniques vital in identifying viral pneumonia effectively [15]. A CT scan generates cross-sectional slices of the bones, blood vessels and soft tissues inside the human body to visualise the pathogenesis of the infection and thereby aid the diagnosis of the disease. Despite the fact that CT images are vastly digitised and they have led many researchers to publish work in this field, CT images require laborious studying and labelling by experienced radiologists. This has resulted in a lack of CT datasets with high-quality labels since the correlated annotation work can be very expensive [16]. Therefore, developing deep learning models for raw CT image analysis can be a promising supplementary diagnostic approach that not only improves the prognostic accuracy but also reduces the workload of radiologists.

Conventional CT-scans classification methods primarily employ a two stage-approach: slice classification followed by the final overall classification. However, training a slice classifier requires slice-wise labels of the CT data. As mentioned above, labelling slices could be manually tedious and time-consuming work. For this reason, high-quality labelled slice-wise data are not always available publicly. This has a huge negative impact when new data sets emerge and the trained models are needed urgently as in the case of the COVID-19 pandemic. Hence, our study considers the scenario where slice-wise annotated training data are not available. In this paper, we propose a single-stage full 3D convolution neural network (CNN) approach that does not require slice-wise annotations for lung CT scan classification. Preliminary results of our work were presented in [17], which proposed a 3D CNN-based approach to solve this problem for the first time in the literature.

Our study includes an efficient method to identify the region-of-interest (RoI) to use as the input to the 3D convolution neural network (CNN) and modify the CNN to support the use of RoI in all layers of the network. We then utilise 3D Gradient-weighted Class Activation Mapping (Grad-CAM) to add explainability to the prediction by displaying the class activation regions in the 3D CT scan volume. Furthermore, we also present a comprehensive ablation study to investigate the contributions of various methodological parameters to the model performance. In this paper, we demonstrate our work by considering the 3D ResNet models [18] as the 3D CNN. The main contributions of our work are as follows:

- Proposal of a framework for CT scan classification without slice-wise annotations;
- Proposal of an algorithm for extraction of salient slices to support the 3D RoI;
- proposal of RoI-aware 3D ResNet architecture for supporting the slice-wise RoI in the deep learning network; and
- Evaluation of the proposed method for depths of 3D ResNet models (18, 50 and 101) and various ranges of RoI.

Although in general there have been many published papers using deep learning for CT image analysis, very few papers proposed reproducible methods that follow the best practice for deep learning models and justify their applicability in real-world in-vivo scenarios [19]. To address the latter, our study proposes a tailor-made pre-processing technique instead of using external segmentation resources. Following reproducible research principles, we describe the design steps in detail, provide all experimental settings and make the code and the trained model available publicly[1]

The rest of the paper is organised as follows: Section II reviews the latest related work that uses deep learning for fast COVID-19 detection. Section III elaborates on our proposed single-stage fully-automatic framework. Section IV presents the experimental results and evaluates the overall model performance. Finally, Section V presents the conclusion and future research prospects.

II. RELATED WORK

This section reviews the related work on the deep learning-based classification of chest scans for detecting COVID-19. These methods are reviewed based on two criteria: the data type and the model (deep learning network) type.

The data types are mainly either 2D or 3D. In general, 2D data can either be chest X-Ray or CT scan images that contain several individual planar images without spatial information corresponding to a volume. 3D data are typically CT scans. They consist of a sequence of slice images with corresponding spatial information to form a 3D volume. There are mainly two drawbacks of utilising 3D CT data. Firstly, data with high-quality slice-wise annotations are limited as they require laborious efforts from experienced radiologists. Secondly, chest CT scans can often include tissues that are irrelevant to diagnosis, hence requiring computationally expensive pre-processing steps, such as, segmentation.

Commonly used model types can be categorised into 3 groups: 2D models; 2D models followed by 1D models (2D+ models); and 3D models. The 2D models merely learn the planar features of 2D data whilst the latter two model types can learn the volumetric features of 3D data. Table 1 provides a summary and intuitive comparison of the related work in terms of 4 aspects: the data modality type; the use of slice-wise annotation; the inclusion of a pre-processing segmentation step; and the model type used.

Following the above description of data types and model types, the remainder of this section reviews the related work under three catgories: 1) 2D models on 2D data; 2) 2D+ models on 3D data; and 3) 3D models on 3D data.

A. 2D MODELS ON 2D DATA

Recent works that used 2D models on 2D data showed satisfactory classification accuracy [20], [21]. COVNet [20],

[1] https://github.com/lestrance/RoI-Aware-ResNet. Our method is applicable to any CT volumetric dataset without requiring slice labels so that it can be deployed subsequently to use without further retraining or with full retraining or with transfer learning the model using other CT datasets.

TABLE 1. An Overview of Related Work (compared in terms of the data modality type, the usage of slice-wise annotation, the inclusion of a segmentation step and the model type).

Paper reference	Data Type	Slice-wise Annotation used?	Data Pre-processing	Model Type
COVNet [20]	2D	×	None	2D
Oh *et al.* [21]	2D	×	None	2D
Wu *et al.* [22]	3D	×	None	2D+
Xu *et al.* [23]	3D	√	Segmentation	2D+
AI-Corona [24]	3D	√	None	2D+
Purohit *et al.* [25]	3D	√	None	2D+
Hu *et al.* [26]	3D	√	None	2D+
COVID-FACT [27]	3D	√	None	2D+
Chaudhary *et al.* [28]	3D	√	Segmentation	2D+
Garg *et al.* [29]	3D	√	None	2D+
Li *et al.* [30]	3D	√	Segmentation	2D+
CNR-IEMN [31]	3D	√	Segmentation	2D+
Yang *et al.* [32]	3D	√	Segmentation	3D
Ours (Present work)	3D	×	Segmentation	3D

a fully-automatic framework for COVID-19 diagnosis from 2D chest CT images, used 2D ResNet-50 [33] as the backbone network for feature extraction. COVNet used 4352 chest CT scans from 3322 patients aiming at distinguishing COVID-19 from other pneumonia. The results showed 90% sensitivity and 96% specificity for COVID-19 detection.

In [21], Oh et al. introduced a deep transfer learning approach with limited X-Ray data. This work firstly feeds raw data into a segmentation network to extract the lung contour followed by patch-wise training based pre-trained ResNet-18 model [33]. The final prediction is determined by fusing patch decisions. The results showed an overall accuracy of 88.9% and a sensitivity of 92.5%.

B. 2D+ MODELS ON 3D DATA

In 2D+ models for 3D data, firstly a 2D convolutional neural network (CNN) is trained to classify 2D slices of the 3D CT volume followed by combining the slice-wise predictions to yield patient-wise predictions [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. The latter is usually achieved by employing another 1D Recurrent Neural Network (RNN), such as, Long Short term Memory (LSTM). A review of the recent 2D+ models for 3D data is presented in this section.

Wu et al. proposed a deep learning-based multi-view fusion approach [22]. They used 2D ResNet-50 [33] as the backbone network to extract planar features from 3 angles of the CT volume. The dataset was collected from two cooperative hospitals in China including 368 COVID cases and 127 CAP cases. Experiments mainly compared the model performance between the single-view model and the multi-view fusion model, the results illustrated that the multi-view fusion model yielded better accuracy (76%) and covid sensitivity (81.9%). Likewise, another work employed a segmentation network to extract annotated infected regions from the pulmonary CT followed by inputting the segmented regions to a pre-trained 2D ResNet-18 to categorise into 3 classes: COVID-19; Influenza; and Healthy [23]. This method achieved an overall accuracy of 86.7%.

Moreover, authors in [24] developed a deep learning framework called AI-Corona, which employed many CNNs, such as DenseNet, ResNet, Xception and EfficientNet as the backbone network. A total of 2124 CT slices were used to train the model and the experimental results yielded an accuracy of 96.4% and COVID sensitivity of 92.4%. Purohit et al. proposed a LeNet-based deep learning model which employed a multi-image representation approach to augment the data [25]. The augmentation method utilised an image sharpening process to enhance edge features. Experimental results showed that the model trained with augmented images achieved an accuracy of 95.38% for 3D CT scans.

Authors in [26] proposed a weakly supervised 2D+ model that can reduce the requirements of annotated data. Noticeably, 450 3D chest CT volumes were utilised for training and 60 annotated 3D CT volumes were utilised for lung segmentation [34]. A multi-scale learning scheme was used to localise the lesions of the infection. An overall accuracy of 87.4% was achieved. Similarly, in COVID-FACT [27], a 2-stage fully-automated framework comprising a pre-trained U-Net called R231CovidWeb [35] for segmentation and a Capsule network (CapsNet [36]) for capturing the spatial information. COVID-FACT achieved an accuracy of 90.82%, a sensitivity of 94.55%, a specificity of 86.04% and an AUC score of 0.97.

Another 2D+ model attained an overall accuracy of 90% by utilising a two-stage binary classification approach [28]. The first classifier was trained to distinguish the normal samples from the infected with either COVID-19 or CAP. The second classifier differentiates the COVID-19 cases from the CAP cases. The backbone networks utilised in stages one and two are DenseNet [37] and EfficientNet [38], respectively. Another 2D+ model presented a two-stage learning strategy inspired by CapsNet [29]. In this method, firstly a slice classifier was trained using ResNet-50 as the backbone network based on two datasets, COVIDx-CT and COVID-CT-MD followed by a BiLTSM [39] network for final classification. Their model achieved an overall accuracy of 88.89%.

Moreover, an ensemble learning approach was proposed in [30]. This method first used a FixMatch semi-supervised approach to train a slice-level classifier followed by the AdaBoost algorithm [40] to train a sequence classifier to obtain the final prediction. CNR-IEMN [31], another 2D+ model, proposed a two-stage learning scheme, that includes a multi-task learning strategy using 4 trained CNNs for slice-level classification followed by the XGboost [41] classifier to get the final diagnosis. This method achieved a very good COVID-19 sensitivity of 91.4%.

C. 3D MODELS ON 3D DATA

The approaches using 2D+ models for 3D data demonstrated good patient-level prediction results. But they require

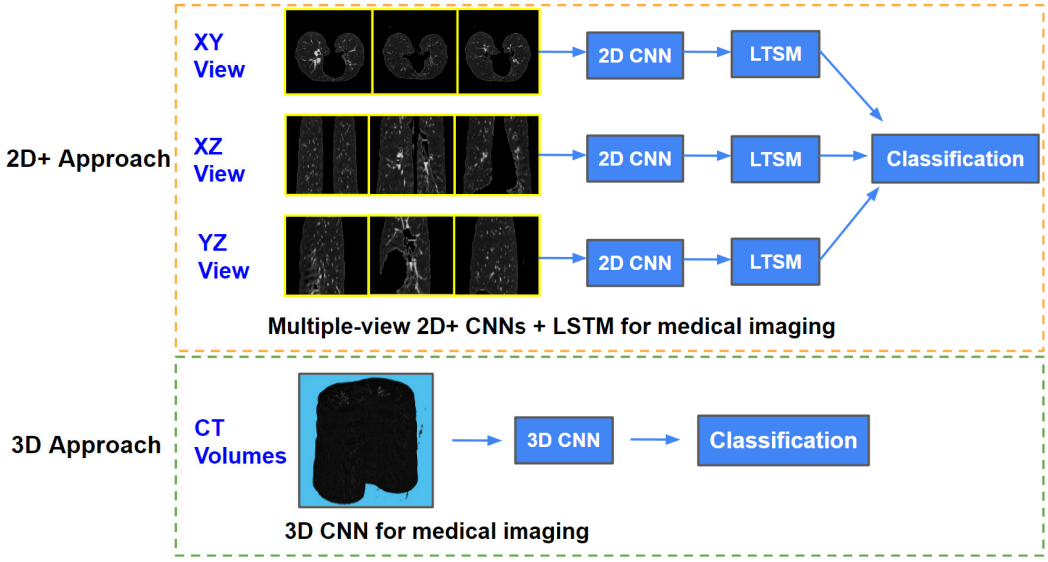

FIGURE 1. Comparison between 2D+ and 3D CNN for medical images classification.

high-quality slice-wise annotated data and well-trained third-party segmentation networks for their success. Similarly, 2D models are capable of recognising planar features and making frame-based predictions for most computer vision tasks. Nonetheless, 2D models cannot directly encode the volumetric patterns of 3D CT scan images without combining multi-view 2D convolution with another 1D network like LSTM. For 2D+ methods, slice-wise labels are required to train the network. Additionally, the input data need to be fed 3 times by different axial views in order to learn the volumetric features. On the contrary, using a 3D model is advantageous for learning the volumetric features of 3D CT images for efficient patient-level diagnosis using a simple and straightforward framework. An intuitive comparison of 2D+ and 3D models for medical imaging classification is shown in Fig. 1.

To the best of our knowledge, only two studies on 3D models for classifying CT chest scans have been reported in the literature. These include the method proposed by Yang et al. [32] and the preliminary results of our work [17]. In [32], Yang et al. proposed a 3D model-based approach consisting of a multi-step learning scheme. In this method, a pre-trained U-Net is used to segment the lung slices as a pre-processing step, followed by feeding in the segmented lung images into a 3D CNN for classification. The next step includes training 5 models to distinguish COVID-19 from CAP in the slices. In the final step, the patient-level prediction is obtained by combining predictions in the previous two prediction steps. This method has achieved good sensitivity in detecting COVID-19 and Normal patients but distinctly low sensitivity in CAP. The comparison of the performance of the related work with respect to our proposed method is shown in Section IV.

III. THE PROPOSED METHOD

Our proposed method comprises two parts: 1) Volumetric data pre-processing and 2) feature extraction and classification using RoI aware 3D ResNet transfer learning. Fig. 2 illustrates the overall algorithmic pipeline of the proposed method. For pre-processing, we propose a novel approach for 3D volumetric data, as discussed in Section III-A. In the second stage, the pre-processed CT images are fed to a modified RoI aware 3D ResNet network for transfer learning to extract features from the CT volumes followed by a tri-class classifier fully connected layer. The process of the second stage is explained in Section III-B.

A. DATA PRE-PROCESSING

As mentioned, the first stage of our proposed method is data pre-processing. The main purpose of data pre-processing is to extract the Region-of-Interest (RoI) from the volumetric data to discard the irrelevant features that might affect the model performance. Moreover, slices that are not likely to be conducive to the diagnosis are discarded thereby improving the model classification accuracy. Furthermore, all volumes are re-scaled throughout the whole process to make them compatible with the network input requirements. The whole procedure of data pre-processing includes 3 steps: 1) Pulmonary segmentation; 2) Volume resampling and 3) Salient slices extraction. The following subsections present the algorithmic details:

1) PULMONARY SEGMENTATION

Pulmonary segmentation is a critical step for most CT scan-based deep learning algorithms as it attempts to extract the RoI. Considering that pneumonia is an inflammation of the tissue in one or both lungs caused by a bacterial or viral infection, merely within the region of the thoraxes is where the abnormality should be detected. Therefore, this step is aimed at discarding irrelevant organs and tissues that might not be conducive to pneumonia diagnosis, thereby reducing the false positive rate and meanwhile compressing the volumetric image by zeroing irrelevant pixels.

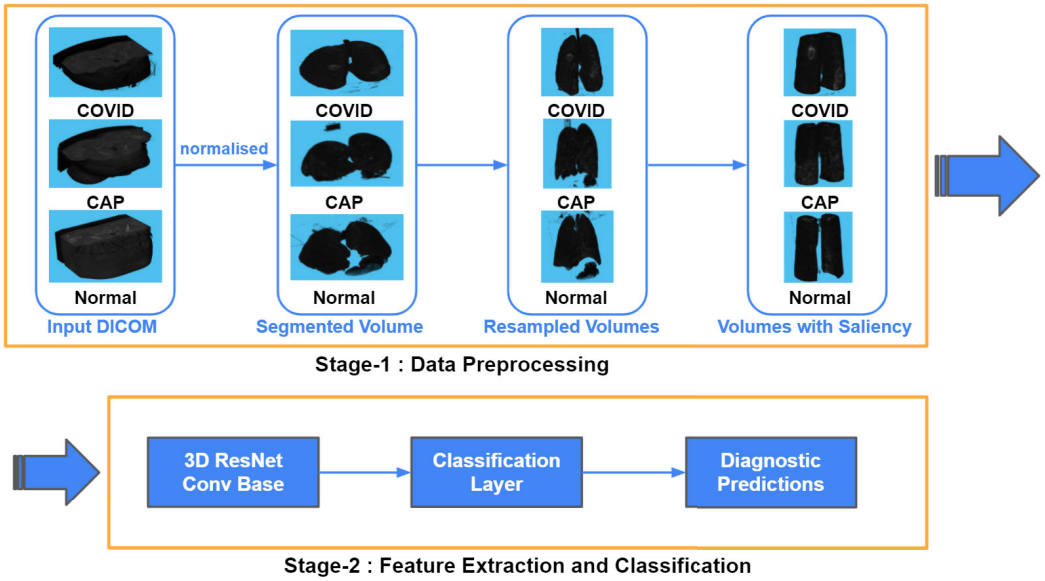

FIGURE 2. An overview of the 2-stage pipeline of our proposed method. Stage-1 includes data pre-processing while stage 2 includes feature extraction and training/classification.

The complete procedure of pulmonary segmentation is illustrated in Fig. 3. Before the segmentation, all volumes are extracted from the given Digital Imaging and Communications in Medicine (DICOM) [42], [43] format files. The pixel intensity of the original data is measured by the Hounsfield Unit (HU), ranging from 0 to 3000. Then the pixel intensity values are normalised to the range from 0 to 1. To illustrate, given a 3-dimensional volume $\boldsymbol{V_i}$, then the linear normalisation is executed as follows:

$$\boldsymbol{V_n} = (\boldsymbol{V_i} - m)\frac{M' - m'}{M - m} + m', \tag{1}$$

where m and M represent the original minimum and maximum pixel value of $\boldsymbol{V_i}$; m' and M' denote the normalised minimum and maximum pixel value of $\boldsymbol{V_n}$. Thus the normalised grey scale volume is denoted by: $\boldsymbol{V_n} \rightarrow \{0, \ldots, 1\}$.

After normalisation, the next step is pulmonary segmentation. Firstly, the contrast of the input slice image is enhanced by saturating the top and bottom 1% pixel values. This aims to improve the segmentation accuracy since natural contrast is critical for segmentation. After that, the volume is binarised using a threshold. Next, the computer vision operations, such as, inverting, border clearing and hole filling are performed sequentially to obtain the corresponding mask images. Finally, as highlighted in Fig. 3, image subtraction is performed using the mask image.

This produces the segmented RoI containing the sheer segments of the thorax. It can be seen that the segmented slices have discarded all irrelevant tissues except for the thorax. It is also observed that from the 3D perspective, the shape of the segmented lung is inclined to a real lung, as it is "squeezed". Hence, the segmented volume needs further processing before inputting to the network.

2) VOLUME RESAMPLING

The second part of pre-processing the raw data is resampling the normalised volume $\boldsymbol{V_n}$. This resampling step aims to generate more slices for each volume, so that the volume is more solid for further preprocessing operations. Noticeably, there are two critical parameters in the resampling step: *Pixel Spacing* (s_1), *and* Slice Thickness (s_2), which specifies the axial resolution of the CT / MRI scan. In the resampling stage, an input volume $\boldsymbol{V}(s_1, s_2)$ with dimensions of $[X,\ Y,\ Z]$ is considered with the corresponding 3D meshgrid as follows:

$$\boldsymbol{V}(\boldsymbol{s_1}, \boldsymbol{s_2}) = \text{meshgrid}(\boldsymbol{x'}, \boldsymbol{y'}, \boldsymbol{z'}), \tag{2}$$

where $\boldsymbol{x'}$, $\boldsymbol{y'}$ and $\boldsymbol{z'}$ represent the 1-dimensional input arrays of coordinates to build the 3D meshgrid, as determined by:

$$\boldsymbol{x'} = s_1\boldsymbol{x}, \tag{3}$$
$$\boldsymbol{y'} = s_1\boldsymbol{y}, \tag{4}$$
$$\boldsymbol{z'} = s_2\boldsymbol{z}, \tag{5}$$

where $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$ represent vectors off all-ones with the length X, Y and Z, respectively. Next, a new 3D meshgrid $\boldsymbol{V_R}$ is defined to denote the resampled volume:

$$\boldsymbol{V_R} = \text{meshgrid}(\boldsymbol{x'_n}, \boldsymbol{y'_n}, \boldsymbol{z'_n}), \tag{6}$$

where $\boldsymbol{x'}$, $\boldsymbol{y'}$ and $\boldsymbol{z'}$ represent the 1-dimensional input arrays of coordinates the 3D meshgrid, as follows:

$$\boldsymbol{x'_n} = s_1\boldsymbol{x_n}, \tag{7}$$
$$\boldsymbol{y'_n} = s_1\boldsymbol{y_n}, \tag{8}$$
$$\boldsymbol{z'_n} = s_2\boldsymbol{z_n}, \tag{9}$$

where $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ denote vectors with the linearly spaced values of k_1, with the length X and Y, respectively. Similarly, $\boldsymbol{z}_n$

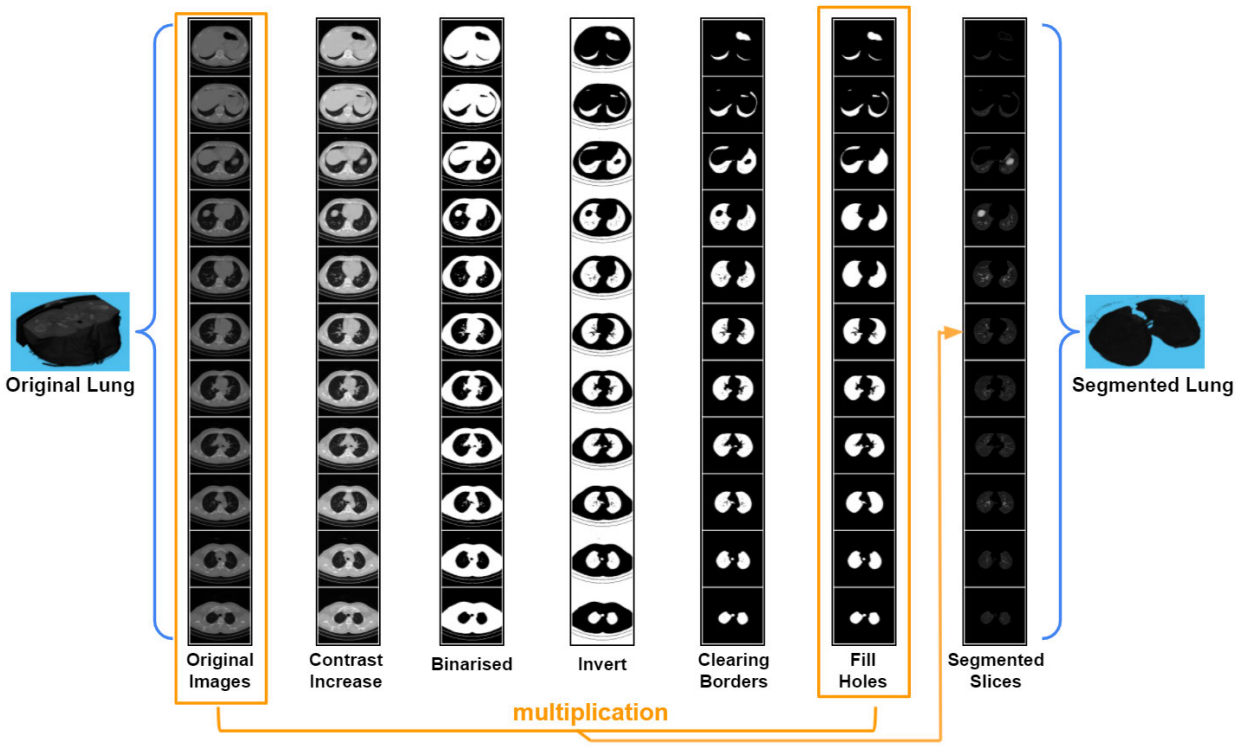

FIGURE 3. Pulmonary Segmentation Process. Note that each row denotes partial slices of the volume from the top to the bottom.

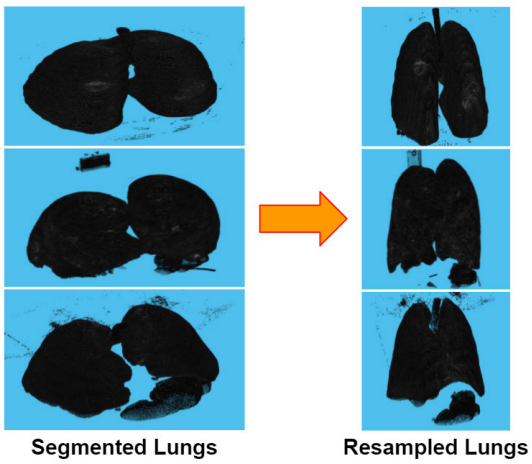

FIGURE 4. Three examples of volume resampling visualisation.

Algorithm 1 Slices With Salience Extraction

Input: $\boldsymbol{V_R}$, the input volume with dimensions of M, N, K; $\boldsymbol{V_m}(i)$, the binary segmented mask for the slice i; t, the saliency thresholding value.
Output: $\boldsymbol{V_s}$, the volume with salient features; $\boldsymbol{l'_a}$, the column vector of the slice-wise lung areas of the volume.
Require: $t \in [0\ 1]$
for $i = 1 : K$ **do**
  $\boldsymbol{l_a}(i) \leftarrow \frac{\sum(\boldsymbol{V_m}(i))}{MN}$.
end for.
Normalising: $\boldsymbol{l'_a} \leftarrow \frac{\boldsymbol{l_a}}{\mathrm{Max}(\boldsymbol{l_a})}$.
Thresholding: $\boldsymbol{l'_a} \leftarrow [\boldsymbol{l'_a} \geq t]$.
$k_t \leftarrow$ index of the first slice $s.t.\ \boldsymbol{l'_a} \geq t$.
$K_n \leftarrow$ length$(\boldsymbol{l'_a})$.
$\boldsymbol{V_s} \leftarrow \boldsymbol{V_R}\{M, N, [k_t : (k_t + K_n)]\}$.
Return $\boldsymbol{V_s}, \boldsymbol{l'_a}$.

denotes a vector with the linearly spaced values of k_2 with a length of Z. The parameters k_1 and k_2 represent the rescaling multiplier parameters corresponding to s_1 and s_2, respectively. Note that the value of k_1 and k_2 are randomly initialised and then the optimal values are experimentally acquired. Examples of a resampled volume are shown in Fig. 4.

3) SALIENT SLICES EXTRACTION

After the pulmonary segmentation, the sheer lung segment is extracted. It can be experimentally observed that not every slice is conducive to pathological diagnosis. Therefore, it is imperative to extract the slices with salient abnormalities to improve the classification accuracy without comprising much volumetric information. This step is critical in handling the scenario of not having slice-wise labels.

We propose a generic method that can extract slices with salience. Given a resampled volume, $\boldsymbol{V_R}$, with dimensions $M \times N \times K$, the first step is to generate a column of binary masks containing pixels of 0 and 1. The area of the lung is

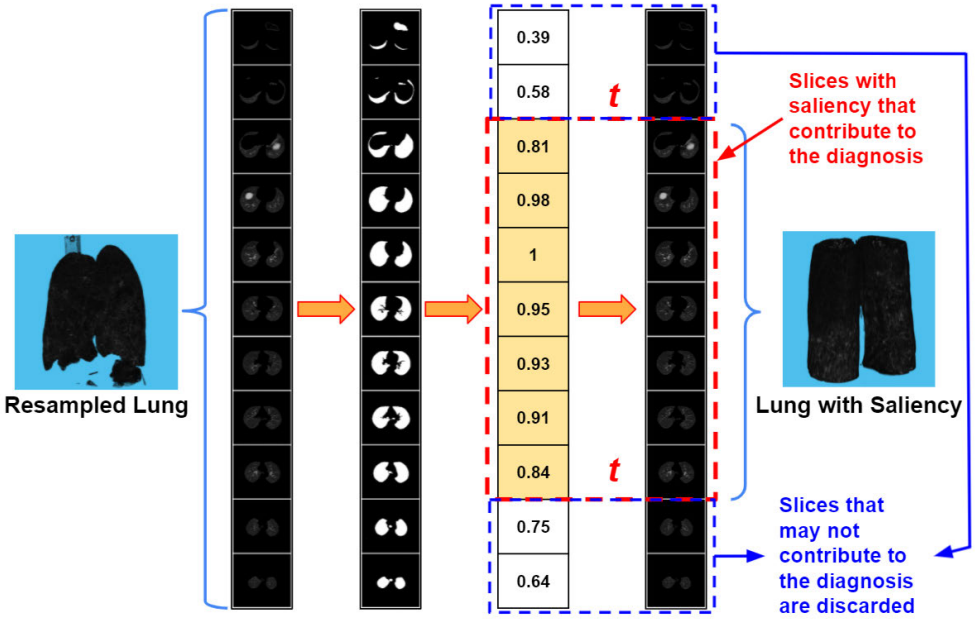

FIGURE 5. An overview of the extraction procedure of slices with salience regions. Note that the first and the second column denote partial slices of the given volume from the top to the bottom and its corresponding segmented mask, respectively. The third column denotes the normalised area vector $l'_a()$ in Algorithm 1. The parameter t refers to as the threshold value for the salience region. Whilst the red dotted line box denotes the slices with saliency and the blue dotted line boxes denote slices without saliency thus to be discarded. Hence, the slices within the red dotted line box are utilised to generate the volume with saliency, which corresponds to the output V_s as shown in Algorithm 1.

calculated by the proportion of 1s in the mask. The column of lung area $\boldsymbol{l_a}$ can be normalised by dividing the maximum area value. Next, the normalised lung area column $\boldsymbol{l'_a}$ is cut off by a customised threshold value, t. Thus, by applying the threshold t, it returns a column of indices of slices with salient features, denoted by $\boldsymbol{l'_a}()$. The final step is to reconstruct the output volume $\boldsymbol{V_s}$ by using slices with salience regions. This process is summarised in Algorithm 1.

Fig. 5 shows the mechanism of extraction of the slices with salience. Fig. 6 shows examples of volumes with salience corresponding to four different threshold values.

B. FEATURE EXTRACTION AND CLASSIFICATION

Feature extraction and classification stage is driven by the 3D ResNet model [18]. We start with the pre-trained 3D ResNet, incorporate RoI aware adaptations and perform transfer learning to train the final model using the training set and the augmented data to develop the final classifier. The following sections present the second stage of the proposed work in detail.

1) 3D ResNet

In 2D ResNet models [33], convolution layers have been designed to extract features from 2D images. Only the planar features can be computed and learned by a 2D network. However, with regard to 3D medical image-based applications, the model is required to capture volumetric representations from the data. To this end, we used 3D convolution layers for the study to extract volumetric features from the chest CT scans. 3D ResNet is realised by extending the filters of 2D ResNet to the third dimension. In this study we have considered 3D ResNet-18, 3D ResNet-50 and 3D ResNet-101 [18]. The basic information of the 3D ResNets used in this work is listed in Table 2.

In the network, a non-linear activation function follows the convolution layer to generate a feature map. The non-linearity of the activation layer grants the neural network to learn complex representations. In this study, the ReLU activation function is employed. ReLU applies an element-wise activation by setting all negative values to zeros. A pooling layer is utilised to down-sample the feature maps while retaining the most salient features. The functionality of the 3D pooling layer can be achieved by integrating a 3D window of pixels and calculating its average or maximum for average pooling or max-pooling, respectively. In this work, max-pooling was used as the first 3D pooling layer and average-pooling was used at the beginning of the classifier stage.

2) RoI AWARE 3D ResNet ARCHITECTURE

Although the input volumes have been segmented, the 3D convolution and pooling operations on the whole input data volume can result in extracting features from that are outside the RoI. Hence, it is necessary to incorporate the pulmonary segmentation based RoI on the feature maps of the intermediate layers. To achieve this, we propose a novel adaptation of the 3D ResNet architecture that supports the use of the

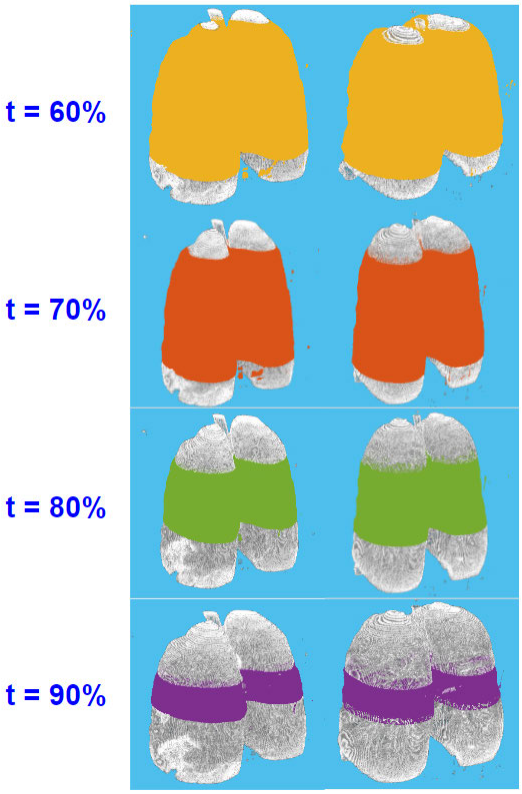

FIGURE 6. Examples of volumes with different saliency thresholding values. Coloured regions indicate the range of slices with salience information whilst the achromatic areas represent slices to be discarded. The range of slices shrinks as t is set to be larger.

TABLE 2. 3D CNNs utilised in this research.

Network	Size(MB)	# Parameters	Input Size
3D ResNet-18	46	3.4×10^7	$224 \times 224 \times 224$
3D ResNet-50	132	4.8×10^7	$224 \times 224 \times 224$
3D ResNet-101	204	8.7×10^7	$224 \times 224 \times 224$

features learned by the network are within the RoI. Fig. 7 shows the adapted ResNet architecture to achieve RoI aware 3D ResNet considering 3D ResNet-18. Note that the structure of the other 2 networks, 3D ResNet-50 and ResNet-101, is similar to that of 3D ResNet-18, but with more layers.

Firstly, the binary volume mask, $\boldsymbol{V_m}$, is generated by binarising the volume input $\boldsymbol{V_s}$ with a threshold value p, as determined by:

$$\boldsymbol{V_m}(x, y, z) = \begin{cases} 1 & \boldsymbol{V_s}(x, y, z) \geq p, \\ 0 & \boldsymbol{V_s}(x, y, z) < p. \end{cases} \tag{10}$$

Then, the activation layer outputs $\boldsymbol{V_a}(L_j)$ for a corresponding activation layer, L for the layer index j, within the neural network are element-by-element multiplied with the corresponding volume mask, $\boldsymbol{V'_m}$ as follows:

$$\boldsymbol{V'_a}(L_j) = \boldsymbol{V'_m} \odot \boldsymbol{V_a}(L_j), \tag{11}$$

where $\boldsymbol{V'_m}$ denotes the resized volume mask, $\boldsymbol{V_m}$ and $\odot$ denotes element-by-element multiplication. Since the convolution and pooling operations downsample the volume, it is not advisable to apply the RoI on the feature maps from relatively deep layers in case significant features are discarded. Note that in this case, only the first 4 activation layers were chosen for applying the RoI awareness in the network.

3) TRANSFER LEARNING AND DATA AUGMENTATION

Since the pre-trained 3D ResNets available in the MATLAB add-ons library are pre-trained using MRI brain images, it has not learnt any hierarchical features of chest CT images. Hence we set all parameters in the convolution base to be trainable allowing transfer-learning using the chest CT images. Finally, the convolution layers are followed by a custom-made ternary classifier that classifies the CT chest scans into three classes (Covid-19, normal and CAP).

Since the training set is small, in order to improve learning, several data augmentation approaches can be considered. Data augmentation used in this work includes 1) resampling volumes with different k_1 and k_2; 2) random cropping; 3) random rotation; and 4) random vertical flipping. Since the last step of data pre-processing is to determine the range of slices with salience, only 2D-based augmentation approaches were considered in this case in order to retain the range of salience unaffected. This means that the 2D slices are augmented before the construction of 3D volumes. Moreover, as the volumetric data are considerably large, the augmented data ought to be generated along with the original data before the experimental stage to reduce the computational cost in the training stage.

IV. PERFORMANCE EVALUATION

This section introduces the experimental setup and the dataset used for this study and presents the performance evaluation of the proposed RoI aware 3D ResNet transfer learning model. All algorithms were implemented using MATLAB R2020b on a PC with AMD Ryzen 3900X CPU on 32GB RAM, the GPU used for training the model is RTX 2080Ti with 11 GB VRAM.

A. DATASET

The dataset utilised in this work, referred to as the "COVID-CT-MD" [44], was available to the participants of the ICASSP-21 COVID-19 SPGC. It comprises volumetric chest CT scans of 231 patients positive for pneumococcal infection and 76 normal people. Among the 231 patients, 171 of them tested positive for COVID-19 infection and 60 are diagnosed with community acquired pneumonia (CAP). Note that these CT scans were conducted from April 2018 to May 2020. The mean age of patients is 50 ± 16. It includes scans from 183 male and 124 female participants. It is worth mentioning that diagnosis of COVID-19 infection is derived from positive real-time reverse transcription polymerase chain reaction (rRT-PCR) tests, the test results are confirmed by an experienced thoracic radiologist. The rest of the cases were

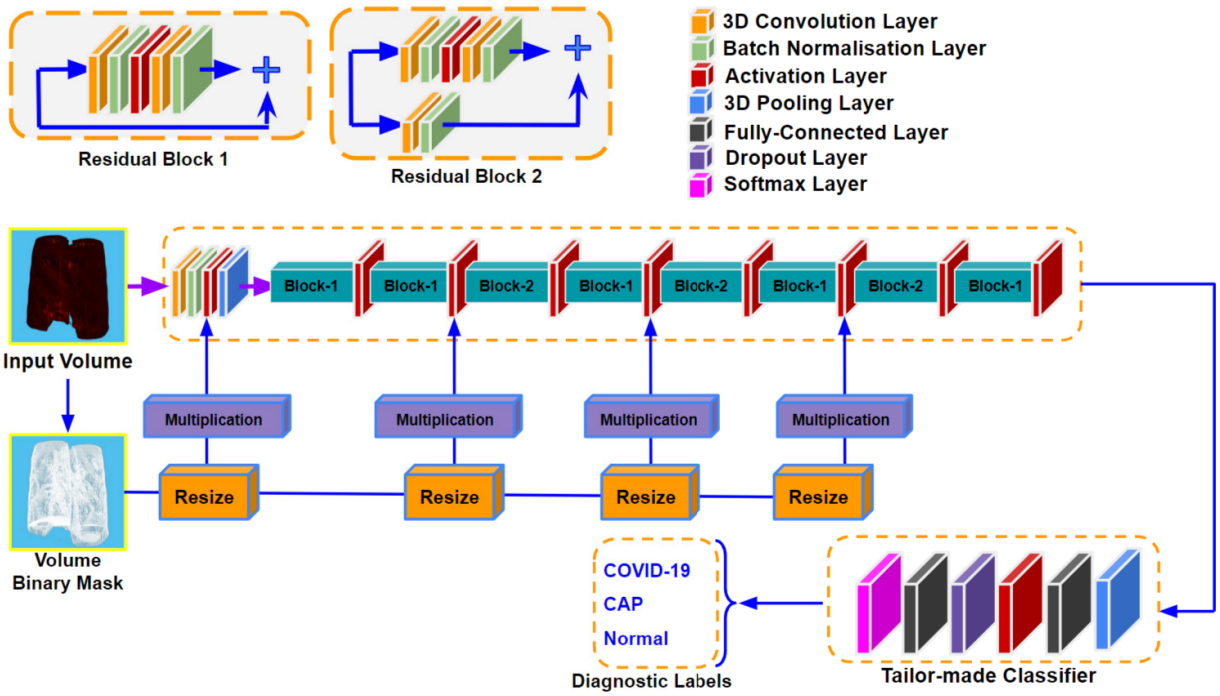

FIGURE 7. RoI aware 3D ResNet-18 architecture.

TABLE 3. Gender and age distribution in COVID-CT-MD [44].

Diagnosis	Cases	Sex Ratio	Age(year)
COVID-19	171	108M/63F	51.6 ± 14.6
CAP	60	35M/25F	57.7 ± 21.7
Normal	76	40M/36F	43.4 ± 41.1

TABLE 4. The statistical parameters(mean and standard deviation) of the Exposure values in COVID-CT-MD [44].

Diagnosis	Exposure Mean	Exposure Standard Deviation
COVID-19	157.15	32.64
CAP	138.96	43.15
Normal	155.12	35.24

confirmed based on clinical parameters and CT scans in the previous study. Additionally, a subset of 55 COVID-19 and 25 CAP cases was annotated by the radiologist with binary slice-wise labels as infected and non-infected. Table 3 shows the gender and age distribution in the dataset.

All volumetric images are saved in DICOM format [42], [43], with a reconstruction matrix with the size of 512×512. The models of CT scope scanners used for this study are SIEMENS and SOMATOM. It is worth mentioning that the radiation exposure dose varies for each volume due to different CT acquisition settings. The mean and standard deviation of the exposure values are shown in Table 4.

B. DATA PRE-PROCESSING PARAMETERS

As demonstrated in Section III-A, there are several unspecified parameters in the data pre-processing stage. Thus, before conducting the experiments it is required to determine these parameters. In the volume resampling stage, values of s_1 and s_2 can be retrieved from the original DICOM files of each volume. k_1 and k_2 are the multiplier for s_1 and s_2, respectively. Hence in this case, k_1 is set to be [1.7 2.3] and the k_2 multiplier is set to be [0.45 0.65], as acquired experimentally. The main purpose of the resampling multipliers is to let the dimensions of the resampled volume match the desired dimensions of the network input. For example, given a volume $\boldsymbol{V}$ with dimensions of [512 512 176], with k_1 and k_2 to be 2.0 and 0.6, the dimensions of the resampled volume $\boldsymbol{V_R}$ are [256 256 293].

After the resampling stage, the volumes are trimmed with a range of slices with salient features. As previously shown in Fig. 5, a critical threshold value t is used to determine the range. Therefore, in this study we selected 3 different values for t being 0.6, 0.7 and 0.8, thereby creating 3 training sets for the ablation study. In the experimental stage, datasets with different t are trained independently hence the results can reveal the optimal t that yields the best model performance. Several examples of volumes with slices of saliency by different t are shown in Fig. 6. For the ROI-aware 3D ResNet, the threshold value p is set to be 0.1 as previously mentioned in Equation.10.

C. EXPERIMENTAL SETUP

In transfer learning a small proportion of shallow layers in the convolution base of the pre-trained network tends to be frozen. These frozen layers remain untrainable and thus will not be initialised with random weights. Usually, it is not necessary to retrain the entire network since the convolution

TABLE 5. Classification layer for transfer learning.

Layer	Activations
GlobalAverage Pooling	$1 \times 1 \times 1 \times 512$
FullyConnected-1	$1 \times 1 \times 1 \times 512$
RelU	-
Dropout	0.5
Fully-Connected-2	$1 \times 1 \times 1 \times 3$
Softmax-Prob	$1 \times 1 \times 1 \times 3$
Classification	Output Label

base of the pre-trained model has already learned generic features for most computer vision tasks. Nonetheless, in this study, we trained the network from scratch as the original dataset does not include CT images. This means all layers in the convolution base of the 3D ResNet are trained and weights are updated accordingly in the training process. In addition to the convolution base, we also construct the classification phase thereby classifying the extracted features, the layout is explicitly illustrated in Table 5.

The model hyper-parameters are defined as follows. In this work, all models are trained with an initial learning rate of 10^{-4} for 60 epochs in total. Each epoch contains 314 iterations so the total iterations are more than $18K$. Note that the learning rate degrades by 0.2 for every 20 epochs to improve the overfitting problem. Adam optimiser is used to update the model weights during training. It is also worth mentioning that due to the limit of GPU VRAM, the mini-batch size is set to be 2.

During the ICASSP-21 SPGC competition, only the training dataset was available to the participants. Hence to evaluate the performance of the proposed method, we used 5-fold cross validation. For that, the training dataset is split into 5 folds and each experiment uses 4 of them for training and 1 for testing, with this repeated for all 5 combinations to get the overall performance metrics. As the test dataset became available after the competition period, we also report the performance for the test set with the model trained using the overall training set in this paper.

D. EVALUATION METRICS

The primary function of this developed model is to predict the given CT scan with a diagnostic class, the most intuitive criterion is the overall accuracy (AC), which can be determined by the ratio of correctly predicted assessments to all testing assessments as follows:

$$\mathrm{AC}(\%) = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100, \tag{12}$$

where TP, TN, FP and FN represent the number of true-positive, true-negative, false-positive and false-negative predictions, respectively. In addition to overall accuracy, two other measures, Sensitivity (SEN) and Specificity (SPE), are used. They are defined as follows:

$$\mathrm{SEN}(\%) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100, \tag{13}$$

$$\mathrm{SPE}(\%) = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}} \times 100. \tag{14}$$

E. ABLATION STUDY RESULTS

For the ablation study in this research, we have considered the 3 variables: 1) ResNet variants; 2) RoI-based modification to the backbone network; 3) the threshold value t for extracting salience slices. The main objective of the ablation study is to investigate how different settings affect the model performance and which combination yields the best classification accuracy. We employ both 5-fold cross validation and of the competition test set, as detailed in Section IV-C to evaluate the performance of the model in these ablation studies.

Table 6 shows the explicit ablation study results using the 5-fold cross validation for 18 experiments corresponding to ablation study variable combinations. We have compared three threshold values, $t = 60\%, 70\%$ and 80% and three backbone networks, ResNet-18, ResNet-50 and ResNet-101 and their corresponding RoI aware modified versions. It can be seen that the RoI aware modified 3D ResNet-101 with $t = 80\%$ yielded the highest overall accuracy of 89.6%. It also shows the best sensitivity for COVID-19 and Normal classes and the highest specificity for CAP and COVID-19 classes. It achieved 100% sensitivity for the class of Normal patients, which is an encouraging finding because it means all patients with and without a disease are correctly diagnosed.

The corresponding confusion matrices for the 18 experiments are shown in Fig. 8. Confusion matrices show the model performance in terms of showing a heatmap corresponding to the percentages of predicted labels with respect to the actual labels. They show a high percentage (darker shade) of diagonal cells and a low percentage (lighter shades) off-diagonal cells, confirming the high performance of the models. The superior performance of the RoI aware modified 3D ResNet-101 with $t = 80\%$ is also evident in the corresponding confusion matrix which shows the highest percentages in the diagonal elements and the lowest percentages in the off-diagonal elements.

Furthermore, the models trained with the complete training set are tested with a second dataset (the competition test set) and the corresponding results for the ablation study combinations are shown in Table 7. Here also, it can be seen that the RoI aware 3D modified ResNet-101 with $t = 80\%$ has resulted in the highest overall accuracy of 90%. It resulted in the highest sensitivity for all three classes, with 96.1% sensitivity achieved for the Normal class. It also shows the highest specificity for COVID-19 and CAP classes. The corresponding confusion matrices are illustrated in Fig. 9. The superior performance of the proposed ROI-Aware ResNet-101 with saliency threshold value $t = 80\%$ is evident from the corresponding confusion matrix, as it shows the highest percentages in the diagonal cells and the lowest

TABLE 6. Ablation Study Results for the 5-fold cross validation using the training set (**Bold** font values correspond to the highest value of the column).

No	Network	RoI aware ResNet	t	AC	CAP SEN	COVID-19 SEN	Normal SEN	CAP SPE	COVID-19 SPE	Normal SPE
1	ResNet-18	N	60%	74.6%	63.3%	72.5%	88.1%	55.9%	85.6%	71.3%
2	ResNet-18	N	70%	78.8%	68.3%	76.6%	92.1%	61.2%	87.9%	76.1%
3	ResNet-18	N	80%	86.3%	78.3%	84.2%	97.4%	70.2%	92.3%	**88.1%**
4	ResNet-18	Y	60%	75.2%	65.0%	72.5%	89.5%	57.4%	85.6%	72.3%
5	ResNet-18	Y	70%	81.1%	71.7%	79.5%	92.1%	64.2%	89.5%	79.6%
6	ResNet-18	Y	80%	87.3%	80.0%	85.4%	97.4%	71.6%	93.4%	**88.1%**
7	ResNet-50	N	60%	76.9%	66.7%	74.9%	89.4%	64.6%	88.3%	71.2%
8	ResNet-50	N	70%	82.4%	71.7%	80.7%	96.1%	69.4%	92.6%	84.4%
9	ResNet-50	N	80%	88.9%	**81.6%**	86.5%	**100%**	81.7%	95.5%	84.4%
10	ResNet-50	Y	60%	78.1%	70.0%	76.6%	90.8%	63.7%	88.9%	72.6%
11	ResNet-50	Y	70%	82.7%	71.7%	81.9%	96.1%	69.4%	93.3%	76.8%
12	ResNet-50	Y	80%	89.3%	**81.6%**	87.1%	**100%**	81.7%	95.5%	85.4%
13	ResNet-101	N	60%	78.5%	68.3%	76.0%	92.1%	62.1%	90.3%	72.6%
14	ResNet-101	N	70%	83.1%	71.6%	80.7%	97.4%	71.7%	92.0%	72.2%
15	ResNet-101	N	80%	87.6%	78.3%	85.6%	**100%**	82.5%	93.6%	80.9%
16	ResNet-101	Y	60%	80.1%	70.0%	77.8%	93.4%	67.7%	91.1%	71.7%
17	ResNet-101	Y	70%	84.7%	71.6%	84.2%	96.1%	76.8%	90.0%	78.5%
18	ResNet-101	Y	80%	**89.6%**	80.0%	**88.0%**	**100%**	**84.2%**	**96.2%**	81.7%

percentages in the off-diagonal cells compared to other confusion matrices.

F. 3D GRAD-CAM

In this work, we also show the 3D gradient class activation maps (Grad-CAMs) in order to add the explainability for the classifier predictions. The algorithm to generate the 3D Grad-CAMs is illustrated in Fig. 10. The final activation layer is used for generating score maps. Then the scores map is resized and normalised to generate the pixel labels that indicate the infection area. Finally, the volume is visualised with the highlighted pixel labels. Thus, the highlighted area indicates the salience regions of the chest volumes corresponding to the class label. This visualisation approach would be a promising technique to assist radiologists with their labelling work. Nonetheless, to develop a robust and applicable model for real-world applications, a substantial amount of patient data is required for further research.

G. DISCUSSION

As can be seen from Table 6 and Table 7, 3D ResNet-101 based models have resulted in the best model performance in overall classification accuracy compared to its shallower variants 3D ResNet-18 and 3D ResNet-50. Moreover, it appears that volumes with $t = 80\%$ yielded superior accuracy and sensitivity compared to those volumes with $t = 60\%$ and $t = 70\%$. For 5-fold cross validation, the highest overall accuracy is 89.6% and achieved by the RoI aware modified 3D ResNet-101 on volumes with $t = 80\%$, this model also achieved the highest sensitivity for COVID-19 and Normal for 88.0% and 100%. Whilst the highest sensitivity of CAP is 81.6% and achieved by both 3D ResNet-50 and the RoI aware modified 3D ResNet-50 on volumes with $t = 80\%$. For specificity, the maximum percentages of CAP and COVID-19 are 84.2% and 96.2% respectively and are achieved by modified ResNet-101 on volumes with $t = 80\%$. The highest specificity for Normal is achieved by both ResNet-18 and modified ResNet-18 on volumes with $t = 80\%$.

For the competition test set in Table 7, the highest overall accuracy is 90.0%, achieved by the RoI aware modified 3D ResNet-101 on volumes with $t = 80\%$. Note that this model also attained the highest values in most other metrics. The best sensitivity values achieved in detecting CAP, COVID-19 and Normal are 96.4%, 88.2% and 96.1%, respectively. The best specificity values achieved are 80.0%, 91.7% and 97.1%, respectively. The models in this work exhibited relatively low sensitivity in detecting CAP and COVID-19 in comparison with Normal. This problem could be because the symptoms of pneumonia in CAP and COVID-19 are similar in many aspects. Hence, the given data are insufficient to develop a model that can consistently distinguish COVID-19 from CAP. In summary, our both evaluation methods demonstrated that the RoI aware modified 3D ResNet-101 with $t = 80\%$ yielded the best model performance.

Table 8 compares the performance of the best model in this work to those of the related works. Our proposed method in this work has achieved the best overall accuracy and

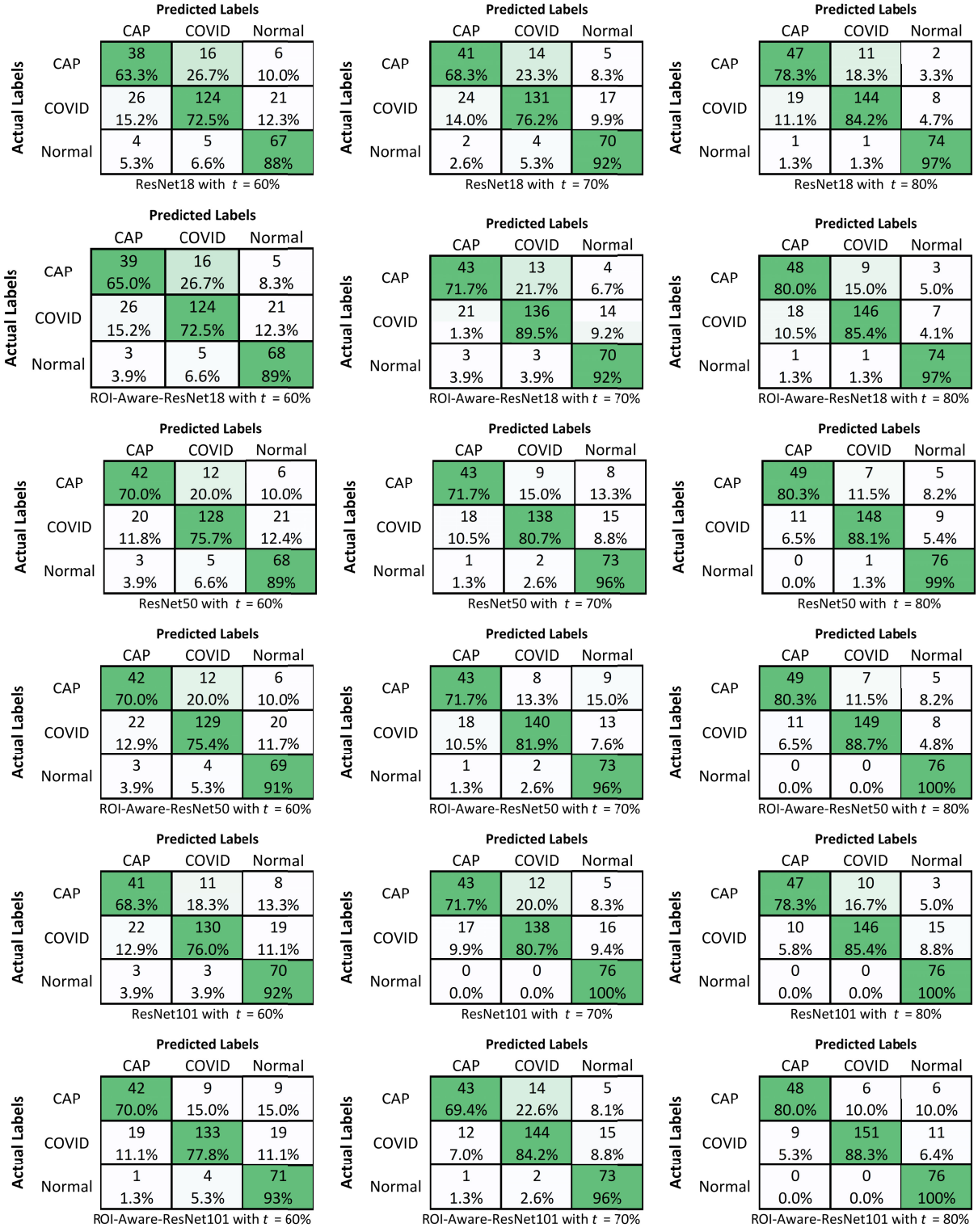

FIGURE 8. Confusion Matrices showing model performance using 5-fold cross validation for various saliency threshold values ($t = 60\%$, $t = 70\%$ and $t = 80\%$ in columns 1-3, respectively), for different 3D networks (3D ResNet-18, 3D ResNet-50 and 3D ResNet-101, in rows 1,3 and5 respectively) and their RoI-Aware modified versions (the proposed method) in rows 2, 4 and 6, respectively.

comprehensively better performance compared to our previous work in [17].

Furthermore, it is evident that the threshold value, t of slices with salience is the most impactful component of the model performance. Models with $t = 70\%$ increased the overall accuracy by approximately 5% compared to that of models with $t = 60\%$ and the models with $t = 80\%$ increased nearly by 7% compared to that of

TABLE 7. Model performance evaluation using the competition test set (Bold font values correspond to the highest value of the column).

No	Network	RoI aware ResNet	t	AC	CAP SEN	COVID-19 SEN	Normal SEN	CAP SPE	COVID-19 SPE	Normal SPE
1	ResNet-18	N	60%	74.6%	63.3%	72.5%	88.1%	55.9%	85.6%	71.3%
2	ResNet-18	N	70%	77.7%	82.1%	80.4%	72.5%	63.9%	78.8%	88.1%
3	ResNet-18	N	80%	80.0%	82.1%	82.4%	76.5%	63.9%	82.4%	90.7%
4	ResNet-18	Y	60%	79.2%	89.3%	82.4%	70.6%	64.1%	82.4%	90.0%
5	ResNet-18	Y	70%	79.2%	92.9%	80.4%	70.6%	59.1%	87.2%	92.3%
6	ResNet-18	Y	80%	79.2%	**96.4%**	84.3%	64.7%	56.3%	89.6%	**97.1%**
7	ResNet-50	N	60%	86.2%	85.7%	82.4%	90.2%	70.6%	87.5%	95.8%
8	ResNet-50	N	70%	85.4%	85.7%	80.4%	90.2%	68.6%	87.2%	95.8%
9	ResNet-50	N	80%	86.2%	86.2%	84.0%	88.2%	71.4%	89.4%	93.8%
10	ResNet-50	Y	60%	87.7%	89.3%	84.3%	90.2%	73.5%	89.6%	95.8%
11	ResNet-50	Y	70%	86.9%	89.3%	82.4%	90.2%	71.4%	89.4%	95.8%
12	ResNet-50	Y	80%	86.9%	89.3%	84.3%	88.2%	71.4%	91.5%	93.8%
13	ResNet-101	N	60%	86.9%	85.7%	86.3%	92.2%	77.4%	**91.7%**	92.2%
14	ResNet-101	N	70%	85.4%	82.1%	82.4%	90.2%	69.7%	89.4%	92.0%
15	ResNet-101	N	80%	89.2%	82.1%	**88.2%**	94.1%	79.3%	90.0%	94.2%
16	ResNet-101	Y	60%	88.5%	85.7%	86.3%	92.3%	77.4%	**91.7%**	92.2%
17	ResNet-101	Y	70%	85.4%	82.1%	82.4%	90.2%	69.7%	89.4%	92.0%
18	ResNet-101	Y	80%	**90.0%**	85.7%	86.3%	**96.1%**	**80.0%**	**91.7%**	94.2%

models with $t = 70\%$. Hence, the proposed algorithm for the extraction of slices with salience is proven to be effective in discarding irrelevant slices, thereby, increasing the classification accuracy. However, it must be mentioned that the concept of salience slices is merely experimentally acquired in this study and not acknowledged by any medical practitioners.

Moreover, it must be emphasised that although the RoI aware modified 3D ResNet architecture merely improved the model performance by a limited share, the modified network is effective in increasing the localisation accuracy of the 3D Grad-CAMs. In Fig. 11, several examples of the comparison between the 3D Grad-CAMs by 3D ResNet and RoI aware modified 3D ResNet are shown. It can be seen that for the same input volume, the activation area of the RoI aware modified 3D ResNet is within the thorax, whereas the unmodified 3D ResNet exhibits a leakage of the activated area out of the lung regions. As aforementioned, the excitation areas of 3D Grad-CAMs can be utilised as an auxiliary visualisation technique for radiologists to localise the possible infected areas of the lung. Since the RoI aware modified network is advantageous in rendering and retaining the excitation within the thorax, we envisage that the modified network is preferable to yield localisation accurately.

Finally, it can be seen that compared to our early work [17], the classification accuracy and sensitivity have been improved after optimising the data pre-processing approach and using a larger extent of data augmentation. It must be noted that the COVID-CT MD dataset is the only publicly available data set available at the time of writing of this paper. As stated in [19], the common drawback of using deep learning for CT is that few studies are able to demonstrate good reproducibility. Our proposed method is implementable and reusable for any subsequent datasets. We have demonstrated this by using the model trained on the first dataset for classifying another dataset (Table 7 and Fig. 9), without further retraining. Similarly, transfer learning can be used in extending the model into newly emerging datasets. In addition, the utilisation of 3D Grad-CAMs based on the RoI-aware ResNet architecture provides the explainability of the classification and enhances the wider applicability of the study in real-world scenarios.

H. LIMITATIONS OF THE STUDY

The main limitation of the proposed method is that even the optimum model cannot fully discriminate COVID cases from CAP. We speculate that could be because SARS-CoV-2 shares similar pathologies with other viral pneumonia. In addition, the severity of infection is unknown as it could cause a failure for deep learning classifiers to differentiate COVID from CAP according to [45].

Another limitation of the study is that for the modified ResNet architecture, we can merely add up to 4 RoI-Aware layers due to the memory limit of the workstation. We envisage that there could be a further improvement to the overall accuracy if more RoI-Aware layers are applied.

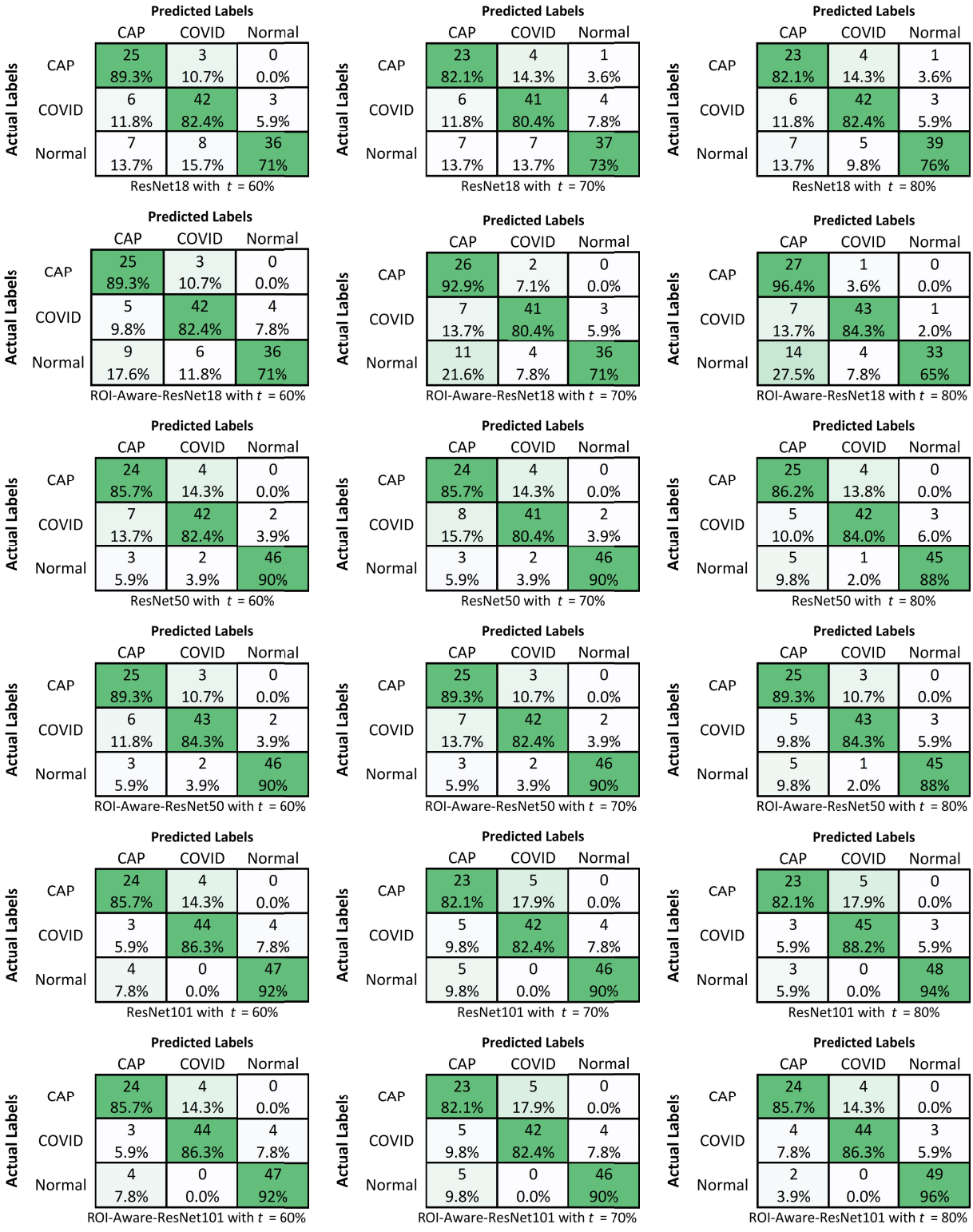

FIGURE 9. Confusion Matrices showing model performance using the competition test dataset for various saliency threshold values ($t = 60\%$, $t = 70\%$ and $t = 80\%$ in columns 1-3, respectively), for different 3D networks (3D ResNet-18, 3D ResNet-50 and 3D ResNet-101, in rows 1,3 and5 respectively) and their RoI-Aware modified versions (the proposed method) in rows 2, 4 and 6, respectively.

Furthermore, we have employed 3D Grad-CAM to visualise the area that activated the diagnostic prediction, which could be a promising auxiliary approach for radiologists to localise the infection. Nonetheless, it needs to be mentioned that since the excitation area is based on the pre-processed volumes so it is not applicable to retrieve and visualise the

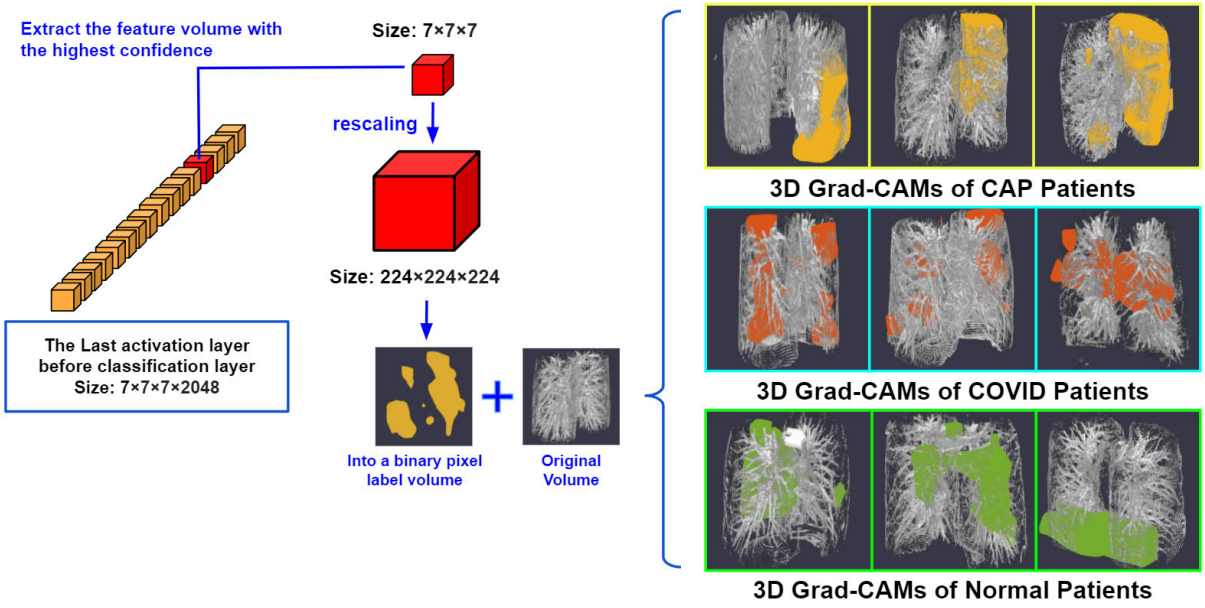

FIGURE 10. 3D Grad-CAMs to illustrate the salient regions corresponding to different class labels. The series of cubes represent the output feature tensors from the last activation layer of the network. The red cube corresponds to the tensor of features that excited the final decision the most. The excitation area is created by binarising the rescaled tensor to 0 and 1. The final 3D Grad-CAMs are generated by combining the excitation with the volume. Note that the volumes here refer to the pre-processed volume used for training.

TABLE 8. The overall performance comparison of the proposed method with related works.

Method	Overall Accuracy	COVID-19 SEN	CAP SEN	Normal SEN
Our proposed method	**90.0%**	88.2%	**96.4%**	**97.1%**
Chaudhary *et al.* [28]	**90.0%**	85.7%	90.0%	94.3%
Garg *et al.* [29]	88.9%	88.6%	90.0%	88.6%
Li *et al.* [30]	86.7%	85.7%	85.0%	88.6%
Our Previous Work [17]	85.6%	82.8%	80.0%	91.4%
Bougourzi *et al.* [31]	81.1%	**91.4%**	45.0%	91.4%
Yang *et al.* [32]	80.0%	88.6%	35.0%	**97.1%**

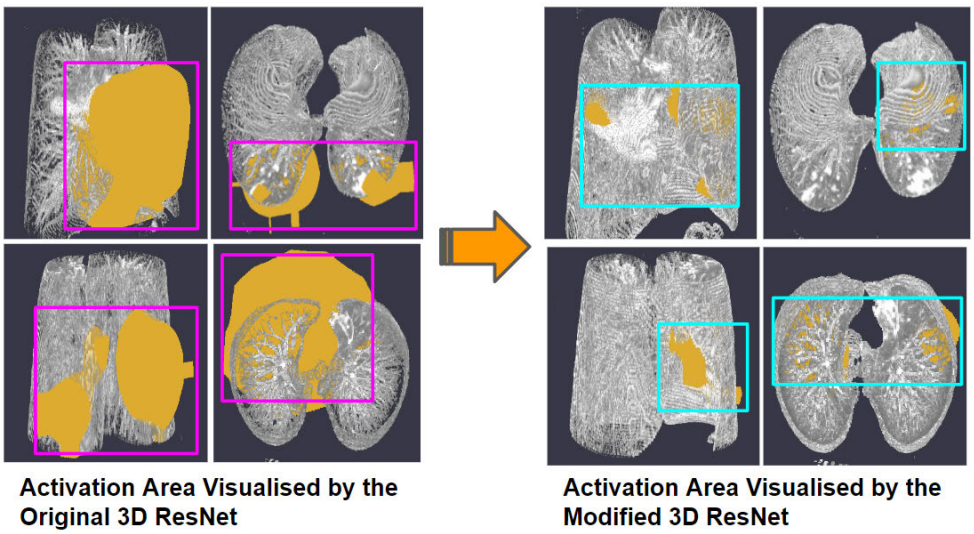

FIGURE 11. Comparison of the 3D CAM of ResNets and ROI-Aware ResNets. The left and right columns illustrate the 3D Grad-CAM of ResNet without and with the modification of RoI-Aware layers, respectively. The highlighted areas indicate the activation of the corresponding networks. Note that the first and second rows show the same volume by different views.

excitation area within the original volume, meaning less plausibility to validate the predicted excitation area. Thus, to justify the RoI-aware 3D Grad-CAM predicted areas, larger-scale data and the expertise of qualified radiologists (for verification) are required to make the localisation more accurate and robust.

I. CODES AVAILABILITY

The software codes developed for the work presented in this paper are available in the GitHub repository, RoI-Aware-ResNet.[2]

V. CONCLUSION

In this paper, we have proposed a novel deep learning strategy for COVID-19 volumetric CT scan classification that does not require slice-wise annotation. We achieved this by introducing an RoI aware modified 3D ResNet architecture. Our proposed scheme includes CT volumetric data pre-processing and RoI aware modified 3D ResNet for feature learning followed by diagnostic classification. Our approach mainly solved the problem of the need for high-calibre slice-wise annotations for the development of deep learning models for the classification of volumetric CT scans.

As can be seen from the results, the RoI-aware modified 3D ResNet-101 with the saliency threshold $t = 80\%$ yielded the best overall accuracy of 90.0%. The highest sensitivity for detecting COVID-19, CAP and Normal are 88.2%, 96.4% and 96.1%, respectively. The highest specificity for detecting COVID-19, CAP and Normal are 91.7%, 80.0% and 97.1%, respectively. Our proposed method shows excellent diagnostic accuracy and outperforms the existing methods. It outperforms the existing 3D model by 10%, confirming the potential of the 3D deep learning networks. It is encouraging to see the excellent performance of the 3D approach that does not require slice-wise annotations for model training.

REFERENCES

[1] M. Lotfi, M. R. Hamblin, and N. Rezaei, "COVID-19: Transmission, prevention, and potential therapeutic opportunities," *Clinica Chim. Acta*, vol. 508, pp. 254–266, Sep. 2020.

[2] R. Zhang, Y. Li, A. L. Zhang, Y. Wang, and M. J. Molina, "Identifying airborne transmission as the dominant route for the spread of COVID-19," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 26, pp. 14857–14863, 2020.

[3] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, and D. Beltekian, M. Roser. (Mar. 2020). *Coronavirus (COVID-19) Vaccinations Statistics and Research*. Accessed: Feb. 1, 2022. [Online]. Available: https://ourworldindata.org/covid-vaccinations?country=OWID_WRL

[4] E. J. Haas, F. J. Angulo, J. M. McLaughlin, E. Anis, S. R. Singer, F. Khan, N. Brooks, M. Smaja, G. Mircus, K. Pan, J. Southern, D. L. Swerdlow, L. Jodar, Y. Levy, and S. Alroy-Preis, "Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in israel: An observational study using national surveillance data," *Lancet*, vol. 397, no. 10287, pp. 1819–1829, May 2021.

[5] C. Aschwanden, "Five reasons why COVID herd immunity is probably impossible," *Nature*, vol. 591, no. 7851, pp. 520–522, Mar. 2021.

[6] A. O. C. Costa, H. de Carvalho Aragão Neto, A. P. L. Nunes, R. D. de Castro, and R. N. de Almeida, "COVID-19: Is reinfection possible?" *EXCLI J.*, vol. 20, pp. 522–536, Mar. 2021.

[2]https://github.com/lestrance/RoI-Aware-ResNet

[7] J. West, S. Everden, and N. Nikitas, "A case of COVID-19 reinfection in the UK," *Clin. Med.*, vol. 21, no. 1, pp. e52–e53, Jan. 2021.

[8] Y. Aiyar, V. Chandru, M. Chatterjee, S. Desai, and A. Fernandez, "India's resurgence of COVID-19: Urgent actions needed," *Lancet*, vol. 397, no. 10291, pp. 2232–2234, Jun. 2021.

[9] E. C. Sabino, L. F. Buss, M. P. S. Carvalho, and C. A. Prete, "Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence," *Lancet*, vol. 397, no. 10273, pp. 452–455, Feb. 2021.

[10] C. J. Smith and A. M. Osborn, "Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology," *FEMS Microbiology Ecology*, vol. 67, no. 1, pp. 6–20, Jan. 2009.

[11] N. Ravi, D. L. Cortade, E. Ng, and S. X. Wang, "Diagnostics for SARS-CoV-2 detection: A comprehensive review of the FDA-EUA COVID-19 testing landscape," *Biosensors Bioelectron.*, vol. 165, Oct. 2020, Art. no. 112454.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[13] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, "Deep learning in medical imaging: General overview," *Korean J. Radiol.*, vol. 18, no. 4, pp. 570–584, 2017.

[14] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[15] H. J. Koo, S. Lim, J. Choe, S.-H. Choi, H. Sung, and K.-H. Do, "Radiographic and CT features of viral pneumonia," *RadioGraphics*, vol. 38, no. 3, pp. 719–739, May 2018.

[16] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[17] S. Xue and C. Abhayaratne, "COVID-19 diagnostic using 3D deep transfer learning for classification of volumetric computerised tomography chest scans," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 8573–8577.

[18] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3154–3160.

[19] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. Aviles-Rivero, C. Etmann, and C. McCague, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Mach. Intell.*, vol. 3, pp. 199–217, Mar. 2021.

[20] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, K. Cao, D. Liu, G. Wang, Q. Xu, X. Fang, S. Zhang, J. Xia, and J. Xia, "Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy," *Radiology*, vol. 296, no. 2, pp. E65–E71, Aug. 2020.

[21] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[22] X. Wu, H. Hui, M. Niu, L. Li, L. Wang, B. He, X. Yang, L. Li, H. Li, J. Tian, and Y. Zha, "Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: A multicentre study," *Eur. J. Radiol.*, vol. 128, Jul. 2020, Art. no. 109041.

[23] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, and Q. Ni, "A deep learning system to screen novel coronavirus disease 2019 pneumonia," *Engineering*, vol. 6, no. 10, pp. 1122–1129, Oct. 2020.

[24] M. Yousefzadeh, P. Esfahanian, S. M. S. Movahed, S. Gorgin, D. Rahmati, A. Abedini, S. A. Nadji, S. Haseli, M. B. Karam, A. Kiani, M. Hoseinyazdi, J. Roshandel, and R. Lashgari, "*AI − Corona*: Radiologist-assistant deep learning framework for COVID-19 diagnosis in chest CT scans," *PLoS ONE*, vol. 16, pp. 1–20, May 2021.

[25] K. Purohit, A. Kesarwani, D. Ranjan Kisku, and M. Dalui, "COVID-19 detection on chest X-ray and CT scan images using multi-image augmented deep learning model," in *Proc. ICMC*, D. Giri, K.-K. Raymond Choo, S. Ponnusamy, W. Meng, S. Akleylek, and S. P. Maity, Eds. Singapore: Springer, 2022, pp. 395–413.

[26] S. Hu, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E. F. Fang, W. Menpes-Smith, J. Xia, H. Ye, and G. Yang, "Weakly supervised deep learning for COVID-19 infection detection and classification from CT images," *IEEE Access*, vol. 8, pp. 118869–118883, 2020.

[27] S. Heidarian, P. Afshar, N. Enshaei, F. Naderkhani, M. J. Rafiee, F. Babaki Fard, K. Samimi, S. F. Atashzar, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "COVID-FACT: A fully-automated capsule network-based framework for identification of COVID-19 cases from chest CT scans," *Frontiers Artif. Intell.*, vol. 4, May 2021, Art. no. 598932.
[28] S. Chaudhary, S. Sadbhawna, V. Jakhetiya, B. N. Subudhi, U. Baid, and S. C. Guntuku, "Detecting COVID-19 and community acquired pneumonia using chest CT scan images with deep learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 8583–8587.
[29] P. Garg, R. Ranjan, K. Upadhyay, M. Agrawal, and D. Deepak, "Multi-scale residual network for COVID-19 diagnosis using CT-scans," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 8558–8562.
[30] B. Li, Q. Zhang, Y. Song, Z. Zhao, Z. Meng, and F. Su, "Diagnosing COVID-19 from CT images based on an ensemble learning framework," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 8563–8567.
[31] F. Bougourzi, R. Contino, C. Distante, and A. Taleb-Ahmed, "CNR-IEMN: A deep learning based approach to recognise COVID-19 from CT-scan," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 8568–8572.
[32] Z. Yang, Y. Hou, Z. Chen, L. Zhang, and J. Chen, "A multi-stage progressive learning strategy for COVID-19 diagnosis using chest computed tomography with imbalanced data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 8578–8582.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[34] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013.
[35] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *Eur. Radiol. Experim.*, vol. 4, no. 1, p. 50, Aug. 2020.
[36] J. Choi, H. Seo, S. Im, and M. Kang, "Attention routing between capsules," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1981–1989.
[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[38] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Mach. Learn. Res.*, vol. 97, Jun. 2019, pp. 6105–6114.
[39] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.
[40] R. E. Schapire, *Explaining AdaBoost*. Berlin, Germany: Springer, 2013, pp. 37–52.
[41] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. SIGKDD*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794.
[42] P. Mildenberger, M. Eichelberg, and E. Martin, "Introduction to the DICOM standard," *Eur. Radiol.*, vol. 12, no. 4, pp. 920–927, Apr. 2002.
[43] M. Mustra, K. Delac, and M. Grgic, "Overview of the DICOM standard," in *Proc. 50th Int. Symp. (ELMAR)*, vol. 1, 2008, pp. 39–44.
[44] P. Afshar, S. Heidarian, N. Enshaei, F. Naderkhani, M. J. Rafiee, A. Oikonomou, F. B. Fard, K. Samimi, K. N. Plataniotis, and A. Mohammadi, "COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning," *Sci. Data*, vol. 8, no. 1, p. 121, Apr. 2021.
[45] F. Shi, L. Xia, F. Shan, B. Song, D. Wu, Y. Wei, H. Yuan, H. Jiang, Y. He, Y. Gao, H. Sui, and D. Shen, "Large-scale screening to distinguish between COVID-19 and community-acquired pneumonia using infection size-aware classification," *Phys. Med. Biol.*, vol. 66, no. 6, Mar. 2021, Art. no. 065031.

SHUOHAN XUE (Student Member, IEEE) received the B.E. degree in mechatronics engineering from the North University of China, Shanxi, China, in 2015, and the M.Sc. degree in electronic and electrical engineering from The University of Sheffield, U.K., in 2019, where he is currently pursuing the Ph.D. degree with the Department of Electronic and Electrical Engineering.

CHARITH ABHAYARATNE (Member, IEEE) received the B.E. degree in electrical and electronic engineering from The University of Adelaide, Australia, in 1998, and the Ph.D. degree in electronic and electrical engineering from the University of Bath, U.K., in 2002. He is currently a Lecturer with the Department of Electronic and Electrical Engineering, The University of Sheffield, U.K. He has published over 90 peer-reviewed papers in leading journals, conferences, and book editions. His research interests include visual content analysis, visual content security, machine learning, and multidimensional signal processing. He was a recipient of the European Research Consortium for Informatics and Mathematics (ERCIM) Postdoctoral Fellowship to carry out research from the Centre of Mathematics and Computer Science (CWI), The Netherlands, from 2002 to 2004, and the National Research Institute for Computer Science and Control (INRIA), Sophia Antipolis, France. He currently serves as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE ACCESS, and *Journal of Information Security and Applications* (JISA) (Elsevier).