Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Expert Systems With Applications 213 (2023) 118939

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

EffViT-COVID: A dual-path network for COVID-19 percentage estimation

Joohi Chauhan[1], Jatin Bedi [*,1]

Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala 147004, Punjab, India

ARTICLE INFO

Keywords:
COVID-19
Percentage estimation
EfficientNet-B7
Vision transformer
Huber loss
Deep network

ABSTRACT

The first case of novel Coronavirus (COVID-19) was reported in December 2019 in Wuhan City, China and led to an international outbreak. This virus causes serious respiratory illness and affects several other organs of the body differently for different patient. Worldwide, several waves of this infection have been reported, and researchers/doctors are working hard to develop novel solutions for the COVID diagnosis. Imaging and vision-based techniques are widely explored for the prediction of COVID-19; however, COVID infection percentage estimation is under explored. In this work, we propose a novel framework for the estimation of COVID-19 infection percentage based on deep learning techniques. The proposed network utilizes the features from vision transformers and CNN (Convolutional Neural Networks), specifically EfficientNet-B7. The features of both are fused together for preparing an information-rich feature vector that contributes to a more precise estimation of infection percentage. We evaluate our model on the Per-COVID-19 dataset (Bougourzi et al., 2021b) which comprises labeled CT data of COVID-19 patients. For the evaluation of the model on this dataset, we employ the most widely-used slice-level metrics, i.e., Pearson correlation coefficient (PC), Mean absolute error (MAE), and Root mean square error (RMSE). The network outperforms the other state-of-the-art methods and achieves 0.9886 ± 0.009, 1.23 ± 0.378, and 3.12 ± 1.56, PC, MAE, and RMSE, respectively, using a 5-fold cross-validation technique. In addition, the overall average difference in the actual and predicted infection percentage is observed to be $< 2\%$. In conclusion, the detailed experimental results reveal the robustness and efficiency of the proposed network.

1. Introduction

Coronavirus, declared a pandemic in 2020 by the World Health Organization (WHO), has critically impacted the life of each human being in the world. Starting from December 2019, the world has seen continued growth in the spread of the life-threatening disease COVID (Coronavirus Disease), with around 457 million cases worldwide and 6.03 million deaths as of March 2022 (WHO, 0000). One primary reason behind the spread of this infectious disease is the lack of required testing infrastructure (medical kits) to cover the target population. A fundamental approach to stop the outspread of COVID-19 and improve the efficiency of medical treatment is the early diagnosis of the disease. Presently, the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test is used everywhere to identify/diagnose the patient with COVID-19. Even though the RT-PCR test has good accuracy, it may require 6–24 h to determine the results and is expensive. This period is very long in the present scenario, where the disease spread rate is very high. As a result, an infected person may infect other persons during testing time period.

Hence, more efficient and accurate techniques are required to reduce the time period and the related consequences such as high spread rate, mortality rate, etc. Several research techniques have been proposed to identify infected persons using alternative methods such as X-ray scans, and CT scans in the recent past (Bougourzi, Contino, Distante and Taleb-Ahmed, 2021; Lassau et al., 2021; Li, Yang, Liang, & Wu, 2021; Song et al., 2021). Among these techniques, Computed tomography (CT scans) has emerged as a powerful tool to detect/diagnose the disease with the required efficiency. Using CT scans for the COVID diagnosis is promising as the CT scans devices have wider availability, great efficiency, and can generate results in less time. Two or three days scans of a person with symptoms have been utilized to early detect the presence of COVID effectively (Chua et al., 2020). However, expert medical proficiency is required to accurately identify the presence of disease in the captured CT scans of an infected person. The problem gets resolved with the introduction of AI (Artificial Intelligence) based techniques to automatically identify the disease from the given set of CT scan images.

* Corresponding author.
E-mail addresses: joohi.chauhan@thapar.edu (J. Chauhan), jatin.bedi@thapar.edu (J. Bedi).
[1] Contributed equally to this work.

https://doi.org/10.1016/j.eswa.2022.118939
Received 14 July 2022; Received in revised form 12 September 2022; Accepted 20 September 2022
Available online 3 October 2022

In recent years, deep learning approaches have shown broad applicability in medical and healthcare domains, including cancer detection, skin lesions classification, eyes disease identification and many more (Goceri, 2021; Hsu & Tseng, 2022; Ma, Kumar, Khetan, Sen, & Bhende, 2022; Shimazaki et al., 2022). Considering this, several researchers have assessed the applicability of computer vision and machine learning techniques on the target classification task (infection detection from CT scans images). From the experimental evaluation of the studies, it is observed that the deep learning techniques, specifically Convolutional Neural Networks (CNNs) based approaches, are the right solution for the COVID-19 detection and segmentation tasks. However, the majority of existing benchmark techniques have majorly focused on utilizing CT scan images for identification (the presence or absence) or segmentation of the COVID infection in a person. Moreover, it has been identified that a few research studies have focused on analyzing and examining several critical aspects related to COVID infection, such as analyzing the evolution of infection in patients, analyzing the impact of a particular treatment on patients' health etc. In this context, the present study involves formulating a regression-based research problem to estimate the percentage of COVID infection in patients from their respective CT scans.

In contrast to the several existing benchmark COVID classification techniques, we aim to propose a fully automatic deep learning and CT scan-based approach to estimate the evolution and severity of COVID infection in patients. This kind of information will benefit medical practitioners in many ways, such as identifying patients with critical needs, treatment planning activities, progress monitoring, evolution study, and many more. The primary research contributions of the work are highlighted as follows. It presents an effective and robust method for predicting the COVID-19 infection percentage. In recent years, deep learning techniques have been widely employed and proven effective in the medical domain for the automated detection and diagnosis of diseases. In the vision task, both EfficientNet (Tan & Le, 2019) and Vision Transformer (ViT) (Dosovitskiy et al., 2020) are considered as computationally efficient and powerful. So, utilizing the benefits of both, this work embeds the features of EfficientNet-B7 and Vision Transformer to develop an automated COVID-19 diagnostic system. The proposed model has been evaluated using 5-fold cross-validation. To validate the performance of the proposed method, we compare the evaluated results with the other state-of-the-art methods and observe that our proposed method achieves state-of-the-art performance on the Per-COVID-19 dataset for all the considered evaluation metrics.

The remainder of this paper is organized as follows. Section 2 introduces the related work, Section 3 explains the proposed methodology. Section 4 presents the experimental results and discussion that includes dataset description, performance metrics, loss function and experimental results. Finally, Section 5 concludes the paper.

2. Related work

CT scan images have greatly helped radiologists worldwide in early diagnosing COVID infection. This section entails providing an in-depth explanation of the state-of-the-art research utilizing CT scan images and deep learning for COVID detection. At a broader level, these research studies can be classified into two types (a) COVID Detection (Bougourzi, Contino et al., 2021; Lassau et al., 2021; Li et al., 2021; Song et al., 2021; Ye et al., 2022) (b) COVID Segmentation (Amara et al., 2022; Bose, Chowdhury, Das, & Maulik, 2022; Elharrouss, Subramanian, & Al-Maadeed, 2022; Fan et al., 2020; Hu et al., 2022; Stefano & Comelli, 2021). The research studies falling into the aforementioned categories are explained in the subsequent paragraph.

A sufficient amount of labeled data is a prerequisite for building an efficient deep learning-based solution to a given problem. In this context, several research studies have open-sourced their datasets to help the research community to design better solutions for accurate diagnosis. Morozov et al. (2020) provided a dataset containing CT scans (anonymous) images. In He (2020), the researcher proposed to address two major problems pertaining to the COVID-19 diagnosis domain. Firstly, the authors provided an open-source CT scan dataset for the target domain. Secondly, a transfer learning-based solution is proposed to achieve high diagnosis efficiency (achieved 94% accuracy). Polsinelli, Cinque, and Placidi (2020) introduced a lighter architecture (CNN based) for COVID-19 detection from Chest CT scans. The results comparison of the approach shows that the developed approach achieved notable improvements (in terms of both accuracy and time) compared to other complex architecture-based classification approaches. Zheng et al. (2020) developed a DL model for automatic detection of COVID-19 from the 3D CT volumes. Initially, the method involved implementing Unet for the lung region segmentation. Followed by this, a three-dimensional neural architecture is employed to predict the COVID-19 infection probability. In a similar context, Wang et al. (2020) employed Unet and 3D deep neural networks for segmentation and infection identification tasks. The approach involved integrating the activation regions in the supervised network & connected components for the localization of the COVID-19 lesions.

Serte and Demirel (2021) developed an artificial intelligence based classification approach to discriminate between CT scans with COVID infection and regular scans. The authors employed ResNet-50 in amalgamation with the majority voting for the classification task and achieved 96% classification accuracy. Zhao et al. (2021) performed a comparative analysis of several existing pre-trained models on the CT scans based COVID classification task. The authors aim to use the out-of-the-field knowledge of the existing pre-trained models on the target problem (CT scans based classification). From the experimental evaluation, the authors stated that the pre-trained Image net model outperforms the existing state-of-the-art solutions. From the existing literature survey, it has been observed that CNN variations have been widely adopted in the COVID-19 infection prediction domain. Considering the vast popularity, Lacerda, Barros, Albuquerque, and Conci (2021) studied the effect of network hyperparameters (backbone network, learning rate, inception modules, number of neurons) on CNN performance. The performance estimation is made by creating architectures with different values of the aforementioned hyperparameters. In 2022, Ter-Sarkisov (2022) proposed a two-stage (regional approach) framework to predict COVID-19. The first stage implements masked R-CNN to detect lesion types in CT scan images. The fused data of detections from the previous step is used for classifying the input image in the next stage. In Basu, Sheikh, Cuevas, and Sarkar (2022), the authors proposed an end-to-end deep learning pipeline (involving feature extraction, selection, and model development) for the infection identification task. The authors combined several advanced techniques such as CNN (ResNet, Inception, DenseNet), Harmony search and hill climbing for the target task. The performance evaluation on different datasets depicts that the model outperforms existing approaches.

Furthermore, to have an in-depth analysis, a comparative summary of the significant research studies in the COVID-19 detection and percentage estimation domain is presented in Table 1. From the exhaustive literature analysis, we found that majority of the research approaches have focused on segmentation and detection of COVID-19 in CT scan images. However, an important research direction on estimating the COVID infection percentage from CT scan images and analyzing its evolution is still under-explored. The current research study focused on exploring this direction and proposed a novel technique for COVID-19 percentage infection estimation using CT scan images.

3. Methodology

The architecture of the proposed EffViT-COVID network is based on the two most powerful sub-networks, i.e., EfficientNet-B7 (Tan & Le, 2019) and Vision Transformer (Dosovitskiy et al., 2020) . The overall framework of the network is presented in Fig. 1.

Table 1
Comparative summary of major research studies in the COVID-19 prediction and percentage estimation domain.

Authors	Research approach/Techniques used	Task/Dataset	Findings/Results
Polsinelli et al. (2020)	Proposed a light CNN (SqueezeNet) approach to detect COVID-19 from CT scans.	Classification task	(a) The approach results significant enhancement over complex CNN architectural designs. (b) The need for GPU acceleration is removed by using the light CNN architecture.
Serte and Demirel (2021)	(a) Developed an artificial intelligence system to detect COVID-19 from 3D CT scan volume. (b) A fusion mechanism amalgamating image level predictions to 3D CT volume is employed.	Classification task using the fine-tuning strategy.	The approach achieved 96% area under curve value for the target detection task.
Zhao, Jiang, and Qiu (2021)	Implemented transfer learning to provide a generalized solution for COVID detection task.	COVID-19 Classification task	(a) The authors signified the impact of various initialization parameters and limited dataset availability on the model results. (b) The pre-trained model on ImageNet21k achieved best classification accuracy (99.2%).
Ortiz, Rojas, Valenzuela, Herrera, and Rojas (2022)	Three phase strategy combining CNN DenseNet-161 and Clustering for the identification, and severity estimation.	Percentage Estimation	The evaluation results obtained using the approach implemented in the research are listed as follows: PC: 0.95 MAE: 5.14 RMSE: 8.47
Chaudhary, Yang, and Qiang (2022)	Amalgamation of Swin transformer (feature extraction) and multi-layer perceptron (regression).	COVID-19 percentage estimation from CT-scans	The evaluation results obtained using the approach implemented in the research are listed as follows: PC: 0.9490, MAE: 4.5042, RMSE: 8.0964
Napoli Spatafora, Ortis, and Battiato (2022)	Authors integrated Mixup Data augmentation module with Inception-v3 model for improved regression performance.	Percentage estimation from CT images	The augmentation techniques helped achieving the desired prediction performance.
Tricarico, Chaudhry, Fiandrotti, and Grangetto (2022)	Proposed feature regularization based deep regression approach for the severity prediction task.	Percentage estimation from CT scan.	The approach achieved significant improvements over baseline by generating 4.912 MAE.
Ter-Sarkisov (2022)	Two-stage workflow architecture is proposed to detect COVID. The first phase involves identifying the lesion types using R-CNN and then, the fused data is used for classification in the next phase.	Dataset consisting of 3000 images is used for the COVID detection task.	(a) The model effectiveness is validated on the basis of various evaluation measures such as sensitivity, F1-score and accuracy. (b) The regional predictions detected by first stage have contributed to the improved prediction results.

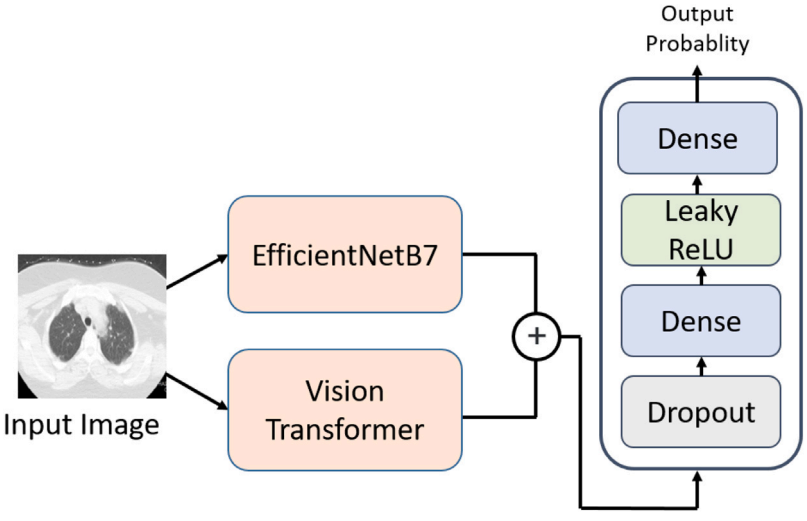

Fig. 1. Overall architecture of the proposed Network that comprises of two paths for feature extraction: one is EfficientNet-B7 and the other is Vision Transformer.

The first path of the network, i.e., EfficientNet-B7, extracts the features from the training set of the considered COVID19 dataset. EfficientNet-B7 comprises seven blocks and each block has different number of Mobile inverted Bottleneck Convolution (MBConv) layers (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018) with different filter size (3×3 or 5×5), shown in Fig. 2. A number associated with each MBConv layer in Fig. 2 denotes the ReLU non-linearity, i.e., ReLU1 and ReLU6. ReLU6 makes the function linear for negative values and also in the range of [0,6]. The initial range of ReLU is [0, inf), this may blow up the activation and explode the gradient. To deal with this problem ReLU6 is used to clip the units at 6. In the formulation, this is equivalent to imagining that each ReLU unit consists of only 6

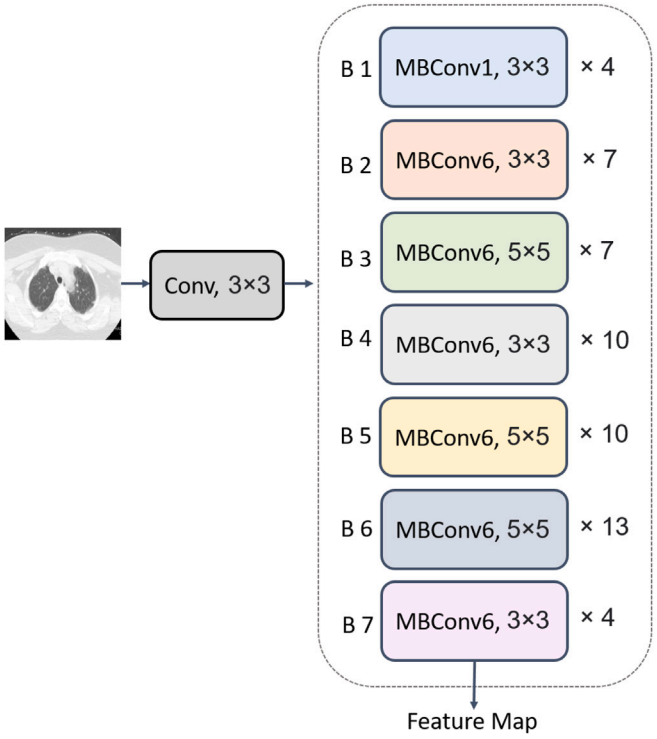

Fig. 2. Overall Architecture of EfficientNet-B7, consisting of seven blocks (B1 to B7) where MBConv (mobile inverted bottleneck convolution) is the main component of each block.

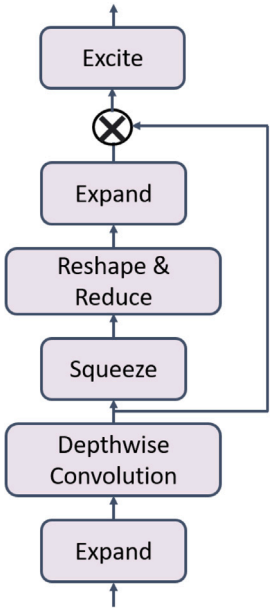

Fig. 3. MBConv block of EfficientNet architecture.

replicated bias-shifted Bernoulli units, rather than an infinite amount. The modified ReLU is represented as:

$$y = min(max(x, 0), 6) \tag{1}$$

Traditionally, the most common way for scaling convolutional neural networks is to scale up by any of one dimension; however, EfficientNet-B7 uses a simple yet effective compound scaling method (Eq. (2)) that scale all the dimension of the network in a balanced manner, i.e., depth d, width w, and resolution r.

$$d = \alpha^{\omega}$$
$$w = \beta^{\omega}$$
$$r = \gamma^{\omega}, \tag{2}$$

where $\alpha.\beta^2.\gamma^2 \approx 2$ and $\alpha, \beta, \gamma \geq 1$, ω is the compound coefficient determining the available extra resources, $\alpha, \beta,\ and\ \gamma$ are the constants that manage the distribution of these resources across the three dimensions of the network.

In EfficientNet-B7, along with the squeeze and excitation optimization, the main component of the network is a Mobile Inverted Bottleneck Convolution (MBConv) (Sandler et al., 2018). The sub-blocks of MBConv are presented in Fig. 3. Initially, the input activation maps are expanded using 1 × 1 convolution, increasing the depth of the output feature maps. Further, the 3 × 3 depth-wise convolution followed by 1 × 1 convolution performs the dimensionality reduction that reduces the number of network parameters and channels in the output feature map. The skip connections present in the network reduce the number of operations as well as the model size. In this work, ImageNet (Deng et al., 2009) weights of EfficientNet-B7 are kept frozen for all the layers that support the encoder to learn the features from the training data efficiently. f_E is the output feature vector obtained after the training process.

The second path of the network utilizes the benefits of transformers. Transformers have gained immense popularity and become state-of-the-art in many NLP tasks. The idea of its success is pre-trained on a large dataset and then fine-tuned on the smaller one, i.e., specific to the task. Transformers are computationally efficient and scalable without compromising performance. Vision transformer (ViT) is proposed as an extension of standard transformer for the image classification task; however, in computer vision, CNN networks still remain dominant (He, Zhang, Ren, & Sun, 2015). Here, we extract the feature map from the vision transformer and convolutional architecture to utilize the benefits of both transformer and CNN.

In general, transformers are based on the encoder–decoder architecture and receive a 1D sequence of tokens as input, but for the image classification task, a 2D image needs to be reshaped first, such that the input image x flattened and transformed into a sequence of 2D patches $x_p \in R^{N\times(P^2.C)}$, where $x \in R^{H\times W\times C}$ is the height and width of original image and C represents the number of channels, (P, P) is the image patch resolution, and the length of input sequence $(x_1, x_2, \ldots, x_N)$ for the transformer or the total number of patches N is calculated as HW/P^2. For performing classification, vision transformers use the encoder module that helps in the mapping of image patches sequence and the semantic label. In addition, the idea of attention mechanism is to employ attention over the different regions of the image and integrate the gathered information across the entire image. In terms of performance, it works best in case of a large size dataset, so to achieve the best out of ViT, pre-trained weights are used and the model is then trained on our COVID dataset.

The Vision Transformer comprises three parts: an embedding layer, an encoder, and a final head classifier. The classifier is used for the final prediction, but here only the final feature vector is noted and embedded into the network. The architecture of the Vision Transformer module used in this work is presented in Fig. 4. The detailed description of each part of ViT is discussed as follows:

Linear Embedding Layer: With the use of learned embedding matrix Em, the sequence of patches is linearly projected into a vector of the model dimension d. The embedding layer then concatenates all the embedded representations along with the learnable token x_{class} which helps in the classification task. Also, an additional information about the patch position is encoded and linearly added to the representations or the sequence of patches so as to keep track of the spatial position of the patches in the input image. Based on the attention mechanism, the positional embedding can be calculated as:

$$E_{mpos(pos,2i)} = \sin\left(\frac{pos}{10\,000^{\frac{2i}{d}}}\right) \tag{3}$$

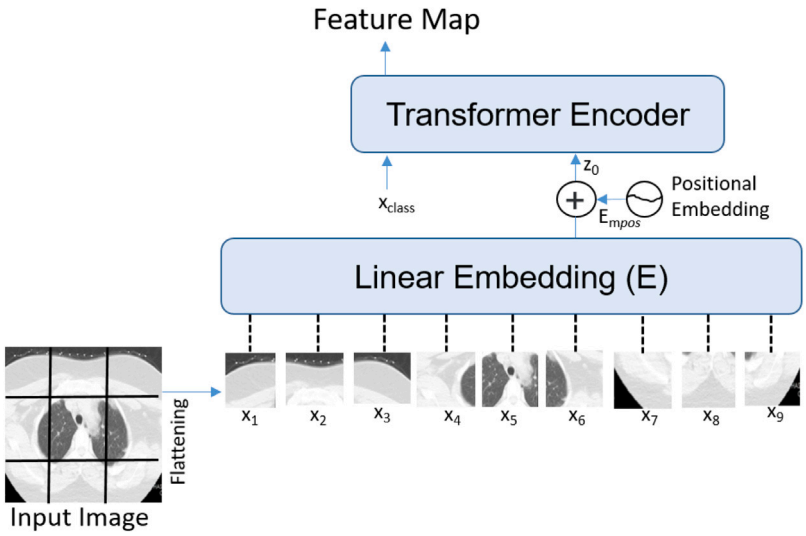

Fig. 4. Architecture of Vision Transformer module.

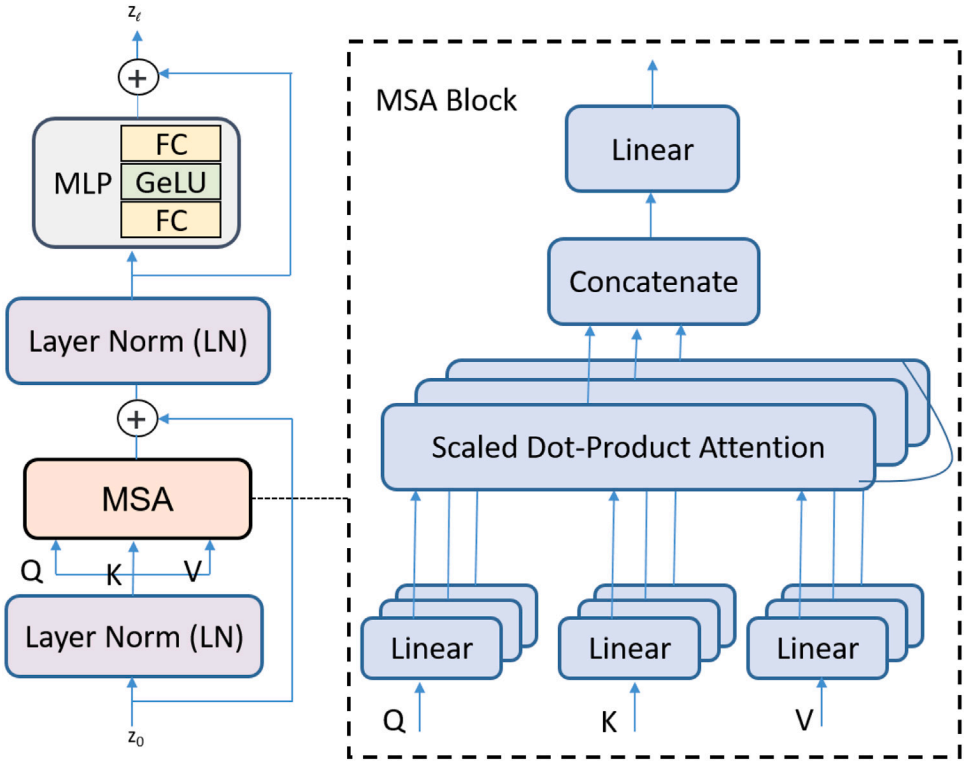

Fig. 5. Architecture of Transform Encoder and its internal blocks.

$$E_{mpos(pos,2i+1)} = \cos\left(\frac{pos}{10\,000^{\frac{2i}{d}}}\right) \tag{4}$$

The joint embedding sequence of patches with z_0 token can be represented as:

$$z_0 = \left[x_{class}; {x_p}^1 E_m; {x_p}^2 E_m; \ldots; {x_p}^N E_m\right] + E_{mpos} \tag{5}$$

where the patch embedding $E_m \in \mathbb{R}^{(P^2.C)\times d}$, the positional embedding $E_{mpos} \in \mathbb{R}^{(N+1)\times d}$, and x_{class} is class label and ${x_p}^N$ are the image patches.

Encoder: The resulting sequence from the embedding layer is passed to the transformer encoder. There are L identical layers in the encoder, and each layer has two major components, one is a multi-head self-attention block (MSA), and the other is a fully connected feed-forward dense block (MLP), as shown in Fig. 5. MSA is the central block of the transformer, and its role is to determine the relationship and importance of single patch embedding with other embeddings within the sequence. It splits the input into several heads and, based on the previous operation, computes the scaled dot product attention for all the heads individually; that is, each head is learning a different level of self-attention. Later, the output of all the attention heads is concatenated and fed into the MLP layer consisting of two fully connected layers and a GELU (Hendrycks & Gimpel, 2016) non-linearity. Both MSA (Eq. (6)) and MLP (Eq. (7)) layers are followed by LayerNorm (LN) and after each block residual connections are applied, as presented in

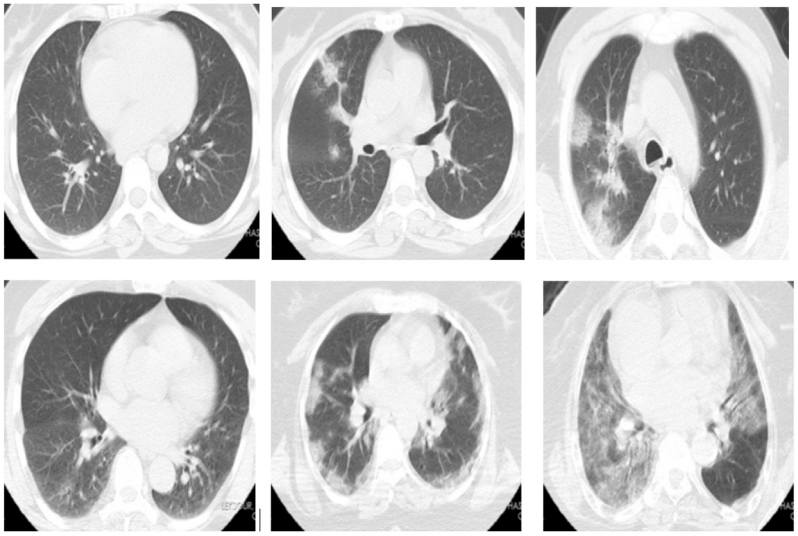

Fig. 6. Some Sample CT Images of COVID-19 patients with infection percentage: $0\%, 8\%, 16\%, 48\%, 65\%,$ *and* 85%, respectively (starting from first row, left to right) Bougourzi et al. (2021).

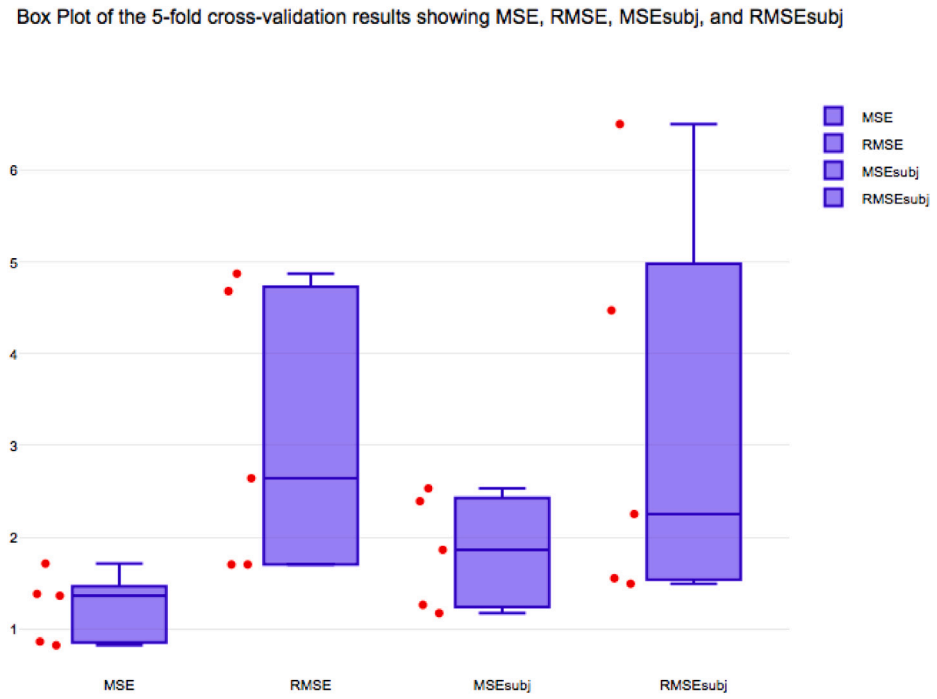

Fig. 7. Boxplot of the 5-fold cross validation results showing MSE, RMSE, subject-wise MSE, subject-wise RMSE.

Fig. 5. The final feature map obtained from the transformer encoder be denoted as f_V.

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1}, l = 1, 2, \ldots, L \tag{6}$$

$$z_l = MLP(LN(z'_l)) + z'_l, l = 1, 2, \ldots, L \tag{7}$$

The feature vectors from both the module or path of the network are added and a single feature vector F_{EV} is thus attained (Eq. (8)). Further, a dropout of 0.3 is applied and the output is fed into a dense layer followed by Leaky ReLU and finally a linear layer to achieve the output probability.

$$F_{EV} = f_E + f_V \tag{8}$$

First, the dataset images are resized to 244×244 at the time of training and the AdamW optimizer (Loshchilov & Hutter, 2017) is used with an initial learning rate of $\exp(-5)$ and weight decay of 0.01. As part of data augmentation, we adopted random cropping of resolution 224×224 and random rotation with an angle in the range of $[-10, 10]$. Initially, the proposed network is trained for 30 epochs with initial learning rate $\exp(-5)$, a learning rate decay of 0.1 after every 10 epochs until the loss converges.

We implemented the proposed architecture using the PyTorch libraries (Paszke et al., 2019) and all the experiments are conducted on a machine with NVIDIA TitanX 12 GB GPU, Intel core 8th generation i7 and 16 GB RAM.

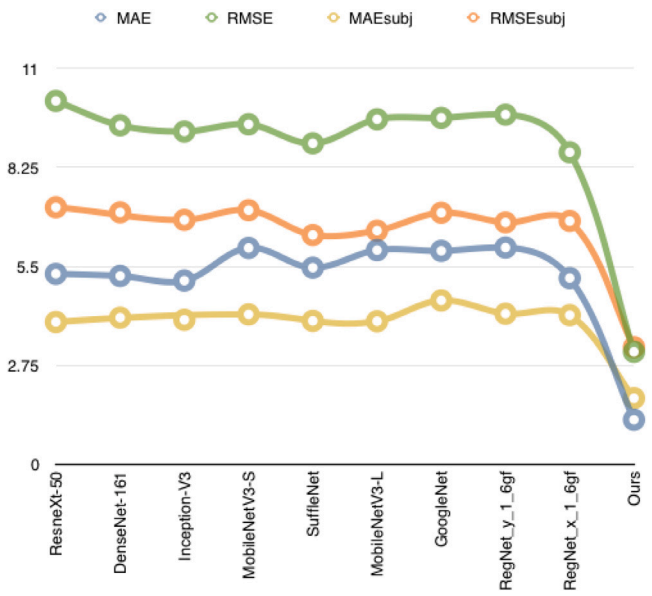

Fig. 8. Quantitative performance analysis by the different methods in terms of MAE, RMSE, Subject-wise MAE and Subject-wise RMSE.

4. Experimental results and discussion

4.1. Dataset description

The dataset comprises a total of 183 CT scans of Coronavirus infected patients. The data have CT scans of both male and female COVID-19 patients of different age groups ranging from 27–70 years. As reported in Bougourzi, Distante et al. (2021) the COVID-19 CT scans were collected from two hospitals of Algeria from June to December 2020. Each CT-scan comprises around 40–70 slices, and there are 150 CT-scans taken with Hitachi ECLOS CT-Scanner having 5 mm slice thickness and 33 CT-scans of 3 mm slice thickness with Toshiba Alexion CT-Scanner collected from Hakim Saidane Biskra and Ziouch Mohamed Tolga hospital, respectively. Based on the ratio of infected lung area and the overall lung size, two radiologists estimated the percentage of COVID-19 infection. There are 3986 labeled slices, which are converted into PNG followed by manual cropping of the lung region. Furthermore, the dataset is divided into five patient-independent folds, i.e., no inter-fold similarity in patient slices. Along with that, the fold slices have an almost similar distribution. Some of the sample CT slices from the dataset are shown in Fig. 6.

4.2. Performance metrics

In order to quantitatively evaluate the capability and effectiveness of the proposed approach and demonstrating the comparative analysis with other state-of-the-art methods, we employ the most widely-used slice-level metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Pearson Correlation coefficient (PC), can be defined as:

$$MAE = \frac{1}{m}\sum_{j=1}^{m}|x_j - \hat{x_j}| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(x_j - \hat{x_j})^2} \tag{10}$$

$$PC = \frac{\sum_{j=1}^{m}(x_j - \overline{x_j})(\hat{x_j} - \overline{\hat{x_j}})}{\sqrt{\sum_{j=1}^{m}(x_j - \overline{x_j})^2}\sqrt{\sum_{j=1}^{m}(\hat{x_j} - \overline{\hat{x_j}})^2}} \tag{11}$$

where, $X = (x_1, x_2, \ldots, x_m)$ are the test data ground truth of COVID-19 percentages for m number of slices and $\hat{X} = (\hat{x_1}, \hat{x_2}, \ldots, \hat{x_m})$ are their corresponding estimations.

Moreover, we also compute the subject-wise metrics, i.e., MAE_{subj}, $RMSE_{subj}$, and PC_{subj}, and can be defined as:

$$MAE_{subj} = \frac{1}{s}\sum_{j=1}^{s}|x_{s_j} - \hat{x_{s_j}}| \tag{12}$$

$$RMSE_{subj} = \sqrt{\frac{1}{s}\sum_{j=1}^{s}(x_{s_j} - \hat{x_{s_j}})^2} \tag{13}$$

$$PC_{subj} = \frac{\sum_{j=1}^{s}(x_{s_j} - \overline{x_{s_j}})(\hat{x_{s_j}} - \overline{\hat{x_{s_j}}})}{\sqrt{\sum_{j=1}^{s}(x_{s_j} - \overline{x_{s_j}})^2}\sqrt{\sum_{j=1}^{s}(\hat{x_{s_j}} - \overline{\hat{x_{s_j}}})^2}} \tag{14}$$

where, $X_s = (x_{s_1}, x_{s_2}, \ldots, x_{s_m})$ are the test data ground truth of COVID-19 infection percentages of the slices of each patient's CT scan. $\hat{X_s} = (\hat{x_{s_1}}, \hat{x_{s_2}}, \ldots, \hat{x_{s_m}})$ are their corresponding patient-level estimations.

4.3. Loss function

Mean Square Error (MSE) is the simplest and most widely used loss function. It is the average of the squared difference between predictions and the ground truth across the whole dataset. As the difference is squared for every data point, MSE will never be negative. However, squaring of difference magnifies the error, and this is one of the major drawbacks of MSE. MSE for N batch size is defined as:

$$L_{MSE} = \frac{1}{N}\sum_{j=1}^{N}(x_j - \hat{x_j})^2 \tag{15}$$

Mean Absolute Error (MAE) covers the drawback of MSE by taking the absolute value of the difference between predictions and ground truth. Thus, all the errors will be weighted on the same linear scale. But it fails in the case of outliers prediction because the more significant errors will be weighted the same as smaller ones.

For learning outliers, MSE performs well, whereas for ignoring, MAE is preferred. To utilize the benefits of both, we used a custom loss function, 'Huber Loss' (Huber, 1992), that balances the MSE and MAE together. MAE and dynamic Huber Loss can be defined as:

$$L_{MAE} = \frac{1}{N}\sum_{j=1}^{N}|x_j - \hat{x_j}| \tag{16}$$

$$L_{Huber} = \frac{1}{N}\sum_{j=1}^{N}z_j \tag{17}$$

$$z_j = \begin{cases} \frac{1}{2}(x_j - \hat{x_j})^2 & for|x_j - \hat{x_j}| \le \delta \\ \delta|x_j - \hat{x_j}| - \frac{1}{2}\delta^2 & Otherwise \end{cases} \tag{18}$$

where N is the batch size, δ is the controlling hyper-parameter that decreasing from 15 to 1 in the training phase, $X = (x_1, x_2, \ldots, x_N)$ is the ground truth COVID-19 percentages, and $\hat{X} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N)$ is the estimated percentages.

If error is less than δ, use MSE, and MAE, otherwise. Let say, $\delta = 1$, so, for the data points having loss value greater than 1, Huber loss magnifies the loss values until they are greater than 1, and when loss values of these data points drop down below 1, it maintains a quadratic function near the center.

4.4. Experiment results

This section presents the performance of the proposed method and the state-of-the-art methods (Bougourzi, Distante et al., 2021; Huang, Liu, Van Der Maaten, & Weinberger, 2017; Xie, Girshick, Dollár, Tu, & He, 2017; Zhang, Zhou, Lin, & Sun, 2018) on the COVID-19 test set. The

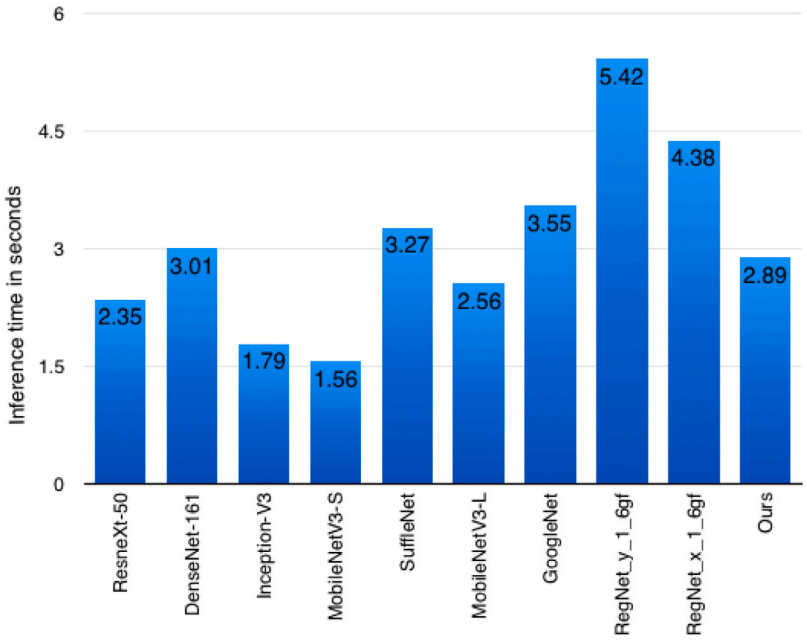

Fig. 9. Average inference time in seconds for the proposed approach and the state-of-the-art methods.

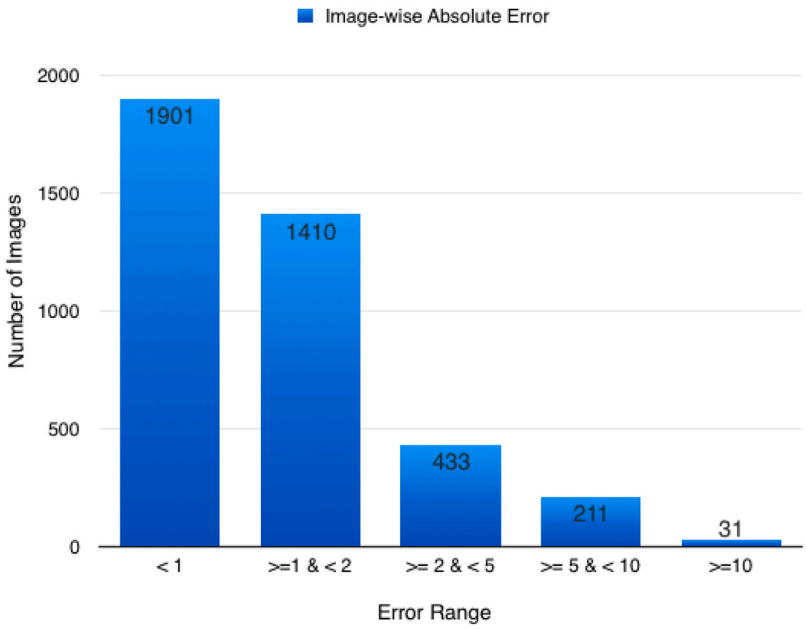

Fig. 10. Image-wise absolute error analysis.

Table 2
5-fold cross-validation results and the overall average performance by the proposed EffViT model.

Fold	$PC \uparrow$	$MAE \downarrow$	$RMSE \downarrow$	$PC_{subj} \uparrow$	$MAE_{subj} \downarrow$	$RMSE_{subj} \downarrow$
Fold 1	0.9924	1.38	2.64	0.9951	1.86	2.25
Fold 2	0.9781	1.71	4.87	0.9861	2.53	4.47
Fold 3	0.9795	1.36	4.68	0.9712	2.39	6.50
Fold 4	0.9957	0.82	1.70	0.9976	1.17	1.55
Fold 5	0.9974	0.86	1.70	0.9987	1.26	1.49
Mean	0.9886	1.23	3.12	0.9897	1.83	3.25
STD	0.0092	0.378	1.56	0.0115	0.608	2.18

results are presented in Tables 2 and 3. Here, we have adopted the five-fold cross-validation techniques, and the results on all five folds by the proposed method are listed in Table 2. It is to be noted that folds 2 and 3 are the most challenging ones as compared to folds 1, 4, and 5. This is probably due to the presence of more challenging patients data in these folds. Fig. 7 present the boxplot of the 5-fold cross validation, and Table 2 presents the evaluation results for each validation fold in more detail. We can observe that overall PC varies from 0.9924 to 0.9974, the MSE varies from 0.82 to 1.71, and RMSE varies from 1.70 to 4.87.

Compared with the benchmark percentage estimation approaches (Bougourzi, Distante et al., 2021), our proposed approach performs better for all considered performance metrics. Our method outperforms the best performing method reported by Bougourzi, Distante et al. (2021), i.e., Inception-V3 by 5.21% in terms of PC, and a difference of 3.87 and 6.13 is observed, in MAE and RMSE, respectively. Moreover, for the subject-wise results, our methods show comparable better

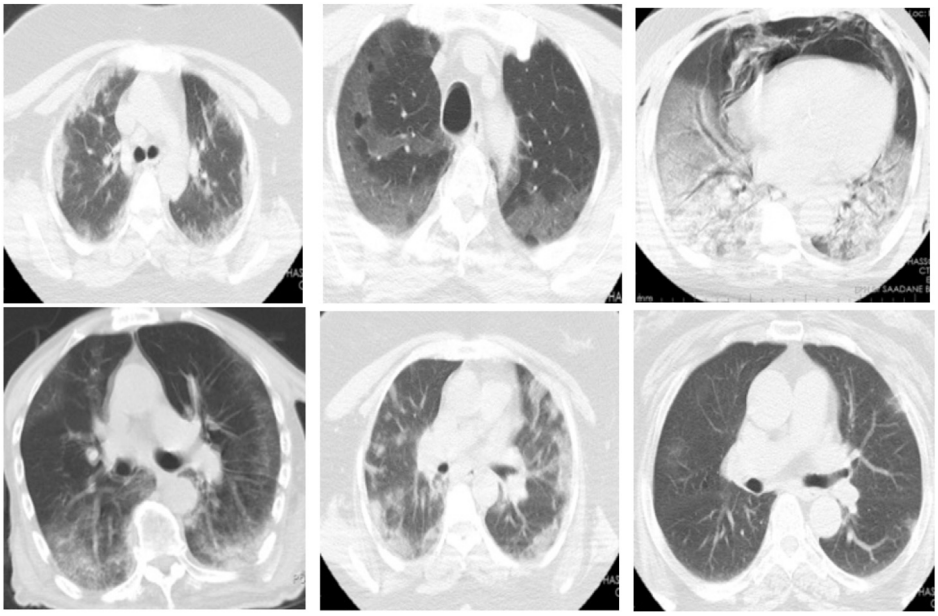

Fig. 11. Qualitative results by the proposed EffViT method on some of the CT Images from test set of COVID-19 dataset (Bougourzi, Distante et al., 2021). Starting from first row, left to right; actual: 60% pred: 60%, actual: 38% pred: 39%, actual: 85% pred: 87%, actual: 26% pred: 26%, actual: 90% pred: 91%, actual: 5% pred: 5%, respectively.

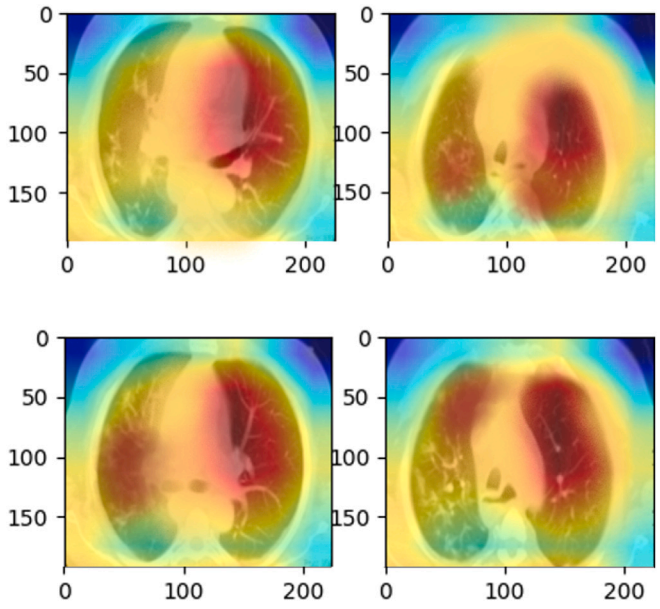

Fig. 12. GradCAM maps of some slices of a patient's CT scan.

Table 3
Overall average performance by the proposed EffViT method and the state-of-the-art methods.

Methods	$PC \uparrow$	$MAE \downarrow$	$RMSE \downarrow$	$PC_{subj} \uparrow$	$MAE_{subj} \downarrow$	$RMSE_{subj} \downarrow$
ResneXt-50	0.9207	5.29	10.10	0.9532	3.95	7.14
DenseNet-161	0.9341	5.23	9.42	0.9582	4.07	7.00
Inception-V3	0.9365	5.10	9.25	0.9603	4.01	6.79
MobileNetV3-S	0.9374	6.01	9.45	0.9540	4.16	7.06
SuffleNet	0.9409	5.46	8.92	0.9613	3.98	6.37
MobileNetV3-L	0.9427	5.95	9.59	0.9598	3.97	6.49
GoogleNet	0.9438	5.93	9.63	0.9577	4.55	6.99
RegNet_y_1_6gf	0.9442	6.02	9.72	0.9598	4.18	6.72
RegNet_x_1_6gf	0.9443	5.17	8.67	0.9590	4.14	6.76
Ours	**0.9886**	**1.23**	**3.12**	**0.9897**	**1.83**	**3.25**

performance, an improvement of 2.94% and deduction of 2.18/3.54 is observed in terms of PC_{subj} and $MAE_{subj}/RMSE_{subj}$, respectively. The performance difference can be visualized from Fig. 8.

We report the average inference of the proposed method and other approaches in Fig. 9. It can be observed that, the average inference time for the percentage prediction task using the proposed method is less than that of second best performing method $RegNet\_x\_1\_6gf$ and overall, the average inference time for these different methods is observed to be within 1.56 s–5.42 s.

Further, we analyze the image-wise results and calculated the absolute error of actual vs predicted percentage value for each image of the dataset, shown in Fig. 10. We observed that almost 83% of the images are under error value of 2 and only 6% images have error value greater than 5.

To more comprehensively interpret the effectiveness and significance of the proposed method, we analyzed the qualitative results and observed that there is not much variation in the actual and predicted infection percentage for the majority of the CT images of patients in the COVID-19 dataset. The overall average difference in the actual and predicted percentage is 1.77. The Qualitative results of the proposed method on some sample CT images from the test set are presented in Fig. 11. From Fig. 11, it can be noted that the model effectively predicts the percentage of COVID-19 infection even when the infection percentage of COVID-19 is at higher side, i.e., 90%, or it is low, i.e., 5%. Based on the proposed network, we also generated the GradCAM maps for some slices of a patient's CT scan. The Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) provides visual explanations that showcase how the CNN is internally learning. The visualization presented in Fig. 12 shows the active areas or regions where the model is focusing. From the maps, it can be observed that the model is mainly focusing on the right side internal region and the left upper and lower lobe. Also, based on the ground truth and visual inspection of the selected slices, the COVID-19 symptoms are mainly present in the same region as displayed in GradCAM maps.

Further, we extended the quantitative and qualitative analysis and noted that out of 3986 slices or CT images, only 31 images are there with a predicted percentage difference greater than 10. However, based on the visual examination of a patient's CT images, we observed a difference in the appearance of the lungs in different slices with the same infection percentage. In Fig. 13, for all three images, the

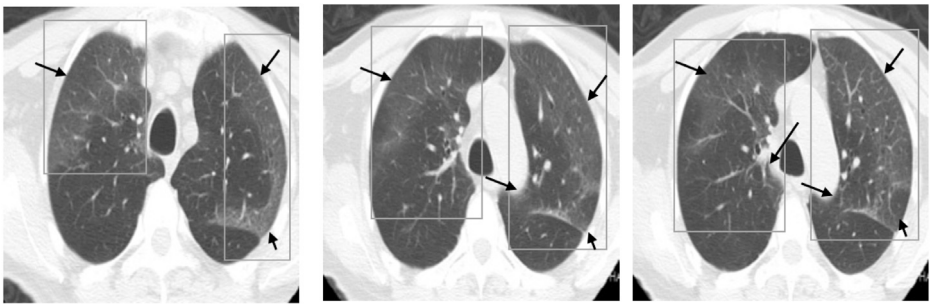

Fig. 13. Results on some sample CT Images of a COVID-19 patients with actual infection percentage 25% for all the three images and predicted as $27\%, 61\%, 64\%$, respectively.

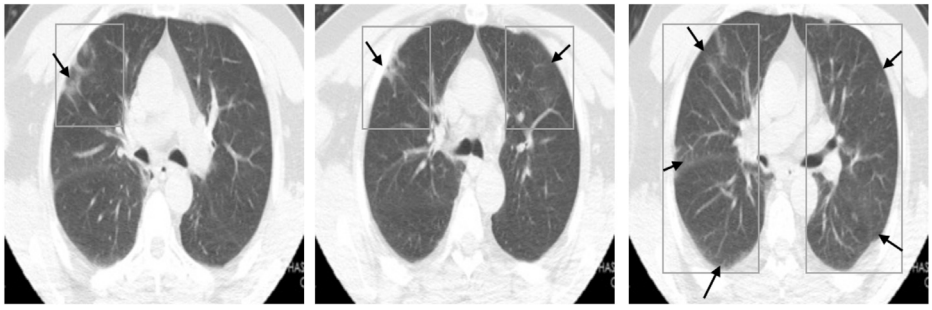

Fig. 14. Results on some sample CT Images of a COVID-19 patients with actual infection percentage 2% for all the three images and predicted as $3\%, 4\%, 27\%$, respectively. Bounding box and arrows represents the region with ground glass opacity (GGO).

Table 4
Overall average performance by the proposed EffViT method and the state-of-the-art methods for the noisy data.

Methods	$PC \uparrow$	$MAE \downarrow$	$RMSE \downarrow$
ResneXt-50	0.8410	10.31	19.15
DenseNet-161	0.8435	10.02	18.93
Inception-V3	0.8542	9.87	17.02
MobileNetV3-S	0.8496	9.66	16.78
SuffelNet	0.8679	9.59	16.55
MobileNetV3-L	0.8722	9.34	15.93
GoogleNet	0.8774	9.01	15.75
RegNet_y_1_6gf	0.8845	8.98	14.98
RegNet_x_1_6gf	0.8936	8.85	13.54
Ours	**0.9568**	**4.01**	**6.35**

actual infection percentage is 25%, and the predicted percentage is $27\%, 61\%, 64\%$, respectively. Similarly, in Fig. 14, for all three images, the actual infection percentage is 2%, and the predicted percentage is $3\%, 4\%, 27\%$, respectively. When examined closely, different levels of ground glass opacity (GGO) (Cozzi et al., 2021) can be noted in different slices of CT scan with the same infection percentage. Ground glass opacity is defined on CT scans as a hazy gray area, with increased density inside the lungs and preserves the margins of pulmonary vessels. Research (Cozzi et al., 2021; Shi et al., 2020; Wang et al., 2020) shows that in the case of COVID-19 pneumonia, GGO is the most common abnormality present in the CT scans of patients. In general, the CT scan of a healthy chest appears black, whereas the haze gray region on the lung indicates that the air spaces inside the lungs are partially filled with some fluid. In Fig. 13, as compared to 1st image, 2nd and 3rd image is having high ground glass opacity. The bounding box and arrows represent the region with GGO. Similarly, in Fig. 14, in the last image high hazy gray region is present as compared to 1st and 2nd. The presence of variations in glass opacity might have influenced and confused the model, leading to erroneous results in some cases.

Moreover, for verifying the robustness of the proposed method, we added the Gaussian noise to the data and evaluated the performance of the proposed model and the other approaches. From Table 4, we observed a comparatively high error for noisy data as compared to the original data. However, it has been noted that the difference in performance of noisy and original data by our method is quite less as compared to the other approaches. Overall, our method performs significantly better when compared to other methods for both original and noisy data.

5. Conclusions

The present research study proposed a framework integrating a convolutional neural network with a transformer architecture for predicting the COVID-19 infection percentage using the CT data. Moreover, we used the Huber loss as our loss function that overcame the drawback of MSE and MAE losses and employed the 5-fold cross-validation approach for the performance evaluation of the model. The prediction results reveal that the proposed EffViT-COVID network outperforms the other state-of-the-art methods and, as compared to other methods, a significant difference is observed in subject-wise as well as overall PC, MAE, and RMSE evaluation metrics. Further, the average difference between actual and predicted COVID-19 infection percentage by the proposed network is $< 2\%$. The results demonstrated the efficacy and robustness of the network. In future, we plan to explore and study the other COVID-19 datasets, such that more informative data can be used for better COVID-19 diagnosis. In addition, several data augmentation and post-processing techniques can be tried out to improvising the performance.

CRediT authorship contribution statement

Joohi Chauhan: Conceived the study, Performed the numerical experiments, Analyzed the data and wrote the manuscript. **Jatin Bedi:** Conceived the study, Performed the numerical experiments, Analyzed the data and wrote the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

All authors approved the final version of the manuscript.

References

Amara, K., Aouf, A., Kennouche, H., Djekoune, A. O., Zenati, N., Kerdjidj, O., et al. (2022). COVIR: A virtual rendering of a novel NN architecture O-Net for COVID-19 CT-scan automatic lung lesions segmentation. *Computers & Graphics*.

Basu, A., Sheikh, K. H., Cuevas, E., & Sarkar, R. (2022). COVID-19 detection from CT scans using a two-stage framework. *Expert Systems with Applications*, Article 116377.

Bose, S., Chowdhury, R. S., Das, R., & Maulik, U. (2022). Dense dilated deep multiscale supervised U-Network for biomedical image segmentation. *Computers in Biology and Medicine*, *143*, Article 105274.

Bougourzi, F., Contino, R., Distante, C., & Taleb-Ahmed, A. (2021). Recognition of COVID-19 from CT scans using two-stage deep-learning-based approach: CNR-IEMN. *Sensors*, *21*(17), 5878.

Bougourzi, F., Distante, C., Ouafi, A., Dornaika, F., Hadid, A., & Taleb-Ahmed, A. (2021). Per-COVID-19: A benchmark dataset for COVID-19 percentage estimation from CT-Scans. *Journal of Imaging*, *7*(9).

Chaudhary, S., Yang, W., & Qiang, Y. (2022). Swin transformer for COVID-19 infection percentage estimation from CT-Scans. In *International conference on image analysis and processing* (pp. 520–528). Springer.

Chua, F., Armstrong-James, D., Desai, S. R., Barnett, J., Kouranos, V., Kon, O. M., et al. (2020). The role of CT in case ascertainment and management of COVID-19 pneumonia in the UK: insights from high-incidence regions. *The Lancet Respiratory Medicine*, *8*(5), 438–440.

Cozzi, D., Cavigli, E., Moroni, C., Smorchkova, O., Zantonelli, G., Pradella, S., et al. (2021). Ground-glass opacity (GGO): A review of the differential diagnosis in the era of COVID-19. *Japanese Journal of Radiology*, *39*(8), 721–732.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. CoRR, abs/2010.11929.

Elharrouss, O., Subramanian, N., & Al-Maadeed, S. (2022). An encoder–decoder-based method for segmentation of COVID-19 lung infection in CT images. *SN Computer Science*, *3*(1), 1–12.

Fan, D.-P., Zhou, T., Ji, G.-P., Zhou, Y., Chen, G., Fu, H., et al. (2020). Inf-net: Automatic COVID-19 lung infection segmentation from CT images. *IEEE Transactions on Medical Imaging*, *39*(8), 2626–2637.

Goceri, E. (2021). Diagnosis of skin diseases in the era of deep learning and mobile technology. *Computers in Biology and Medicine*, *134*, Article 104458.

He, X. (2020). Sample-efficient deep learning for COVID-19 diagnosis based on CT scans. *IEEE Transactions on Medical Imaging*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. CoRR, abs/1512.03385.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415.

Hsu, B. W.-Y., & Tseng, V. S. (2022). Hierarchy-aware contrastive learning with late fusion for skin lesion classification. *Computer Methods and Programs in Biomedicine*, Article 106666.

Hu, H., Shen, L., Guan, Q., Li, X., Zhou, Q., & Ruan, S. (2022). Deep co-supervision and attention fusion strategy for automatic COVID-19 lung infection segmentation on CT images. *Pattern Recognition*, *124*, Article 108452.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics* (pp. 492–518). Springer.

Lacerda, P., Barros, B., Albuquerque, C., & Conci, A. (2021). Hyperparameter optimization for COVID-19 pneumonia piagnosis based on chest CT. *Sensors*, *21*(6), 2174.

Lassau, N., Ammari, S., Chouzenoux, E., Gortais, H., Herent, P., Devilder, M., et al. (2021). Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. *Nature communications*, *12*(1), 1–11.

Li, C., Yang, Y., Liang, H., & Wu, B. (2021). Transfer learning for establishment of recognition of COVID-19 on CT imaging using small-sized training datasets. *Knowledge-Based Systems*, *218*, Article 106849.

Loshchilov, I., & Hutter, F. (2017). Fixing weight decay regularization in Adam. CoRR, abs/1711.05101.

Ma, D., Kumar, M., Khetan, V., Sen, P., & Bhende (2022). Clinical explainable differential diagnosis of polypoidal choroidal vasculopathy and age-related macular degeneration using deep learning. *Computers in Biology and Medicine*, *143*, Article 105319.

Morozov, S., Andreychenko, A., Pavlov, N., Vladzymyrskyy, A., Ledikhova, N., Gombolevskiy, V., et al. (2020). MosMedData: Chest CT scans with COVID-19 related findings dataset. arXiv preprint arXiv:2005.06465.

Napoli Spatafora, M. A., Ortis, A., & Battiato, S. (2022). Data augmentation for COVID-19 infection percentage estimation. In *International Conference on Image Analysis and Processing* (pp. 508–519). Springer.

Ortiz, S., Rojas, F., Valenzuela, O., Herrera, L. J., & Rojas, I. (2022). Determination of the severity and percentage of COVID-19 infection through a hierarchical deep learning system. *Journal of Personalized Medicine*, *12*(4), 535.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Polsinelli, M., Cinque, L., & Placidi, G. (2020). A light CNN for detecting COVID-19 from CT scans of the chest. *Pattern Recognition Letters*, *140*, 95–100.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510–4520).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618–626).

Serte, S., & Demirel, H. (2021). Deep learning for diagnosis of COVID-19 using 3D CT scans. *Computers in Biology and Medicine*, *132*, Article 104306.

Shi, H., Han, X., Jiang, N., Cao, Y., Alwalid, O., Gu, J., et al. (2020). Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *The Lancet Infectious Diseases*, *20*(4), 425–434.

Shimazaki, A., Ueda, D., Choppin, A., Yamamoto, A., Honjo, T., Shimahara, Y., et al. (2022). Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. *Scientific Reports*, *12*(1), 1–10.

Song, Y., Zheng, S., Li, L., Zhang, X., Zhang, X., Huang, Z., et al. (2021). Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(6), 2775–2780.

Stefano, A., & Comelli, A. (2021). Customized efficient neural network for COVID-19 infected region identification in CT images. *Journal of Imaging*, *7*(8), 131.

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri, & R. Salakhutdinov (Eds.), *Proceedings of machine learning research: Vol. 97*, *Proceedings of the 36th international conference on machine learning* (pp. 6105–6114). PMLR.

Ter-Sarkisov, A. (2022). COVID-CT-mask-net: prediction of COVID-19 from CT scans using regional features. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, 1–12.

Tricarico, D., Chaudhry, H. A. H., Fiandrotti, A., & Grangetto, M. (2022). Deep regression by feature regularization for COVID-19 severity prediction. In *International conference on image analysis and processing* (pp. 496–507). Springer.

Wang, X., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., et al. (2020). A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. *IEEE Transactions on Medical Imaging*, *39*(8), 2615–2625.

Wang, Y., Dong, C., Hu, Y., Li, C., Ren, Q., Zhang, X., et al. (2020). Temporal changes of CT findings in 90 patients with COVID-19 pneumonia: A longitudinal study. *Radiology*, *296*(2), E55–E64.

WHO, (0000). coronavirus (COVID-19) dashboard, URL: https://covid19.who.int/.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492–1500).

Ye, Q., Gao, Y., Ding, W., Niu, Z., Wang, C., Jiang, Y., et al. (2022). Robust weakly supervised learning for COVID-19 recognition using multi-center CT images. *Applied Soft Computing*, *116*, Article 108291.

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6848–6856).

Zhao, W., Jiang, W., & Qiu, X. (2021). Deep learning for COVID-19 detection based on CT images. *Scientific Reports*, *11*(1), 1–12.

Zheng, C., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., et al. (2020). Deep learning-based detection for COVID-19 from chest CT using weak label. *MedRxiv*.