Weakly-supervised lesion analysis with a CNN-based framework for COVID-19

Wu, Kaichao; Jelfs, Beth; Ma, Xiangyuan; Ke, Ruitian; Tan, Xuerui; Fang, Qiang

DOI:
10.1088/1361-6560/ac4316

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Wu, K, Jelfs, B, Ma, X, Ke, R, Tan, X & Fang, Q 2021, 'Weakly-supervised lesion analysis with a CNN-based framework for COVID-19', *Physics in Medicine and Biology*, vol. 66, no. 24, 245027. https://doi.org/10.1088/1361-6560/ac4316

PAPER • **OPEN ACCESS**

Weakly-supervised lesion analysis with a CNN-based framework for COVID-19

IOP Publishing Phys. Med. Biol. 66 (2021) 245027 https://doi.org/10.1088/1361-6560/ac4316

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
19 August 2021

REVISED
30 November 2021

ACCEPTED FOR PUBLICATION
14 December 2021

PUBLISHED
31 December 2021

PAPER

Weakly-supervised lesion analysis with a CNN-based framework for COVID-19

Kaichao Wu[1,2], **Beth Jelfs**[2], **Xiangyuan Ma**[1], **Ruitian Ke**[3], **Xuerui Tan**[3,*] and **Qiang Fang**[1,*]

[1] Department of Biomedical Engineering, Shantou University, Shantou, People's Republic of China
[2] School of Engineering, RMIT University, Melbourne, Australia
[3] The First Affiliated Hospital of Shantou University Medical College, Shantou, People's Republic of China
[*] Author to whom any correspondence should be addressed.

E-mail: qiangfang@stu.edu.cn **and** xrtan1@stu.edu.cn

Keywords: COVID-19, chest CT image, weakly-supervised, lesion identification, GGO

Abstract

Objective. Lesions of COVID-19 can be clearly visualized using chest CT images, and hence provide valuable evidence for clinicians when making a diagnosis. However, due to the variety of COVID-19 lesions and the complexity of the manual delineation procedure, automatic analysis of lesions with unknown and diverse types from a CT image remains a challenging task. In this paper we propose a weakly-supervised framework for this task requiring only a series of normal and abnormal CT images without the need for annotations of the specific locations and types of lesions. *Approach.* A deep learning-based diagnosis branch is employed for classification of the CT image and then a lesion identification branch is leveraged to capture multiple types of lesions. *Main Results.* Our framework is verified on publicly available datasets and CT data collected from 13 patients of the First Affiliated Hospital of Shantou University Medical College, China. The results show that the proposed framework can achieve state-of-the-art diagnosis prediction, and the extracted lesion features are capable of distinguishing between lesions showing ground glass opacity and consolidation. *Significance.* The proposed approach integrates COVID-19 positive diagnosis and lesion analysis into a unified framework without extra pixel-wise supervision. Further exploration also demonstrates that this framework has the potential to discover lesion types that have not been reported and can potentially be generalized to lesion detection of other chest-based diseases.

1. Introduction

Coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Huang *et al* 2020), and since the beginning of 2020 (Shi *et al* 2020), it has been widely spread worldwide due to person to person transmission (Chan *et al* 2020). To date, the WHO has reported that more than 260 million confirmed cases of COVID-19 globally, with more than 5 million deaths (World Health Organization 2021). Hence, there is a need for accurate diagnosis and treatment protocols.

When undergoing clinical analysis, COVID-19 patients display lesions which can clearly be seen on chest computed tomography (CT) images. Thus, CT scans play an essential role in early screening and diagnosis of COVID-19 as well as informing on treatment guidelines. Previous investigations reported several typical types of lesions shown on the chest CT of patients with COVID-19. Of the terms used to describe the clinical manifestation of such lesions in CT images, the most frequently observed are ground glass opacity (GGO), crazy paving pattern (GGO with superimposed inter- and intra-lobular septal thickening) and consolidation (see figure 1 for examples of COVID-19 lesions).

Manually annotating the lesions on CT slices requires separating lesions from the image background and setting multiple imaging parameters to identify the diverse lesions. As this process is time- and labour-consuming, automatic identification of the lesions is highly desirable in clinical studies. However, lesion analysis

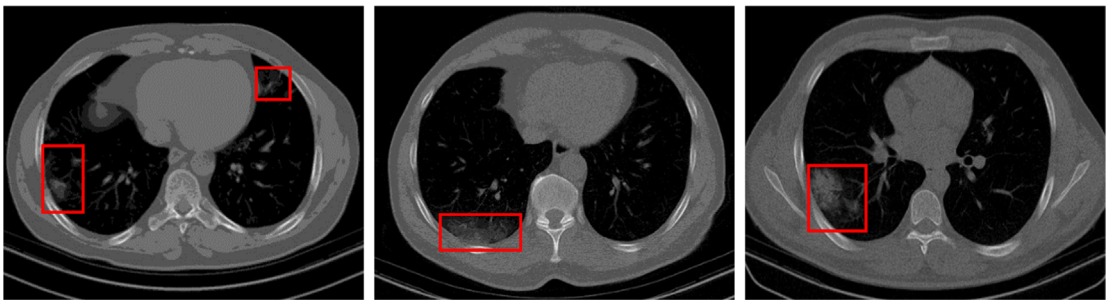

Figure 1. Example CT imaging of COVID-19 lesions (circled by red boxes). Left: GGO; Middle: crazy-paving pattern; Right: consolidation.

of COVID-19 is more challenging than traditional recognition tasks, since the imaging patterns of these lesions have a wide variety in their locations, shapes, as well as textures. In recent years, deep learning with convolutional neural networks (CNNs) has shown its value for many medical image analysis tasks, such as the disease screening (Gulshan *et al* 2016, Hirata *et al* 2020), disease grading (Yonekura *et al* 2017, Meng *et al* 2020) and lesion segmentation (Huang *et al* 2020b, Cao *et al* 2020). Hence, there is interest in harnessing the power of CNNs to detect the subtle distinctions between multiple lesions. Such distinctions can be hard for humans to detect yet can provide a reference for clinical diagnosis.

There already exist several CNN-based models for the analysis of COVID-19 based on CT scans, including AI-assisted differential diagnosis (Wang *et al* 2020a, 2020d, Li *et al* 2020, Mei *et al* 2020, Wang *et al* 2021, Ying *et al* 2021) infected lung segmentation (Jin *et al* 2020, Li *et al* 2020) and severity assessment of COVID-19 (Gozes *et al* 2020b, Huang *et al* 2020b). As COVID-19 has been proven to cause destruction of the pulmonary parenchyma, there has also been a focus on developing intelligent models dedicated to the localization, segmentation, and quantification of lung lesions in patients with this disease (Zhang *et al* 2020a, 2020b, Duran-Lopez *et al* 2020, Ghoshal and Tucker 2020, Shi *et al* 2021). However, despite being the most valuable guide to aid clinicians in making diagnoses, enacting treatment and determining a quarantine plan, the link between CT findings of COVID-19 lesions and the clinical manifestation of the disease has been paid less attention. Instead most conclusions are derived by experienced clinicians, which can be subjective.

With accurate and clear information about the lesion, valuable guidelines for clinicians to enact treatment or follow-up can be provided. Hence, in this paper, we focus on automatic identification of diverse types of lesions over a sequence of CT images. Considering the heavy delineation work, we use only weak annotations that the images are normal/abnormal with no detailed lesion information required. To achieve this we design a CNN-based framework with two branches, namely the diagnosis and lesion branches. For the diagnosis branch we develop a CNN model to automatically screen suspected COVID-19 cases. The lesion branch is connected to the diagnosis branch via a Grad++ module. This module is designed by a fact that the CNN's ability to accurately classify CT images originates from the detected lesion features in the abnormal images. By revealing the lesion features generated in the diagnostic procedure, the Grad++ module ensures the multi-lesion detector in the lesion branch can function without explicitly considering the variability in shape and texture of the lesions, significantly reducing the annotation burden.

The effectiveness of the framework is evaluated on independent datasets, and the results show the diagnosis branch can achieve robust and competitive performance with the maximum accuracy up to 99.41% and performance of the lesion identification is effective. Moreover, further exploration of this framework's potential is carried out, which demonstrates that the multi-lesion detector can detect the lesions that do not report as a clinical manifestation. The main contributions of this paper can be summarized as follows:

(i) A CNN-based framework is presented to integrate the positive CT images prediction and lesion identification.

(ii) A lesion indicator is provided by exploit feature maps under image-level supervision, which can be used to capture the lesion without explicitly considering the lesions' shape and texture.

(iii) The lesion features are extracted and then clustered into different groups with an unsupervised clustering method, the results show that the abstract representation of lesions is discriminative.

The rest of this paper is organized as follows. Section 2 summarizes the related work on artificial intelligence (AI)-assisted differential diagnosis and lesion identification for COVID-19. Section 3 introduces the proposed

Table 1. List of abbreviations.

Abbr.	Explanation	Abbr.	Explanation
COVID-19	Coronavirus disease 2019	GGO	Ground glass opacity
CNN	Convolutional neural network	CT	Computed tomography
CAM	Class activation map	LAM	Lesion activation map
FeCNN	Feature pyramid network embedded convolutional neural network	LAHm	Lesion activation heatmap
FPN	Feature pyramid network	GT	The ground-truth
COVIDx2a	A large-scale compound CT dataset	Own	CT data collected from the First Affiliated Hospital of Shantou University
CTset	CT data collected from Negin Radiology Medical Center (Iran)	ROC curve	Receiver operating characteristic curve
CNCB	CT data collected from the China National Center for Bioinformation	PR curve	Precision-recall curve
MLP	Multi-layer perceptron	AUC	The area under ROC curve
LBP	Local binary pattern descriptor	AP	The average precision
SVM	Support vector machine	Hist	Grey-scale histogram descriptor

CNN-based framework with two branches. In section 4 we describe our experimental setup and section 5 presents results. Finally, the discussion and conclusion are given in sections 6 and 7. A list of the abbreviations used in this paper is given in table 1.

2. Related work

In this section, we review the methods for AI-assisted differential diagnosis and lesion identification, which are two recent trends in the study of COVID-19 that closely relate to our work.

2.1. AI-assisted differential diagnosis

Previous methods for AI-assisted differential diagnosis can be roughly divided into two categories: binary classification and multi-class classification. Binary classification approaches aim to distinguish COVID-19 and non-COVID-19 cases. Whereas multi-class classification often focuses on three classes (Wang *et al* 2020c, Li *et al* 2020, Ying *et al* 2021): a normal or non-pneumonia class, COVID-19 cases, and other disease cases. In this work, we target binary classification, as these approaches can quickly and easily detect COVID-19 positive images with high specificity, which is valuable for the following lesion analysis.

Binary classification can be used to distinguish between COVID-19 negative versus COVID-19 positive (Zhang *et al* 2020a, Jin *et al* 2020, Narin *et al* 2021), for instance, in Jin *et al* (2020), a deep learning framework was proposed integrating lung segmentation and classification for COVID-19 detection. Alternatively binary classification can be used to distinguish COVID-19 from other diseases such as pneumonia (Wang *et al* 2020c, 2020d, Ghoshal and Tucker 2020, Wang *et al* 2021). In either case the structure of the neural network used to provide the AI-based differential diagnosis of COVID-19 can be broken down into either 2D-CNNs or 3D-CNNs.

2.1.1. 2D-CNN modelling

Due to their fast acquisition, x-ray images are often the initial step in the study of COVID-19. Naturally, for x-ray images 2-dimensional CNN models are used (Wang *et al* 2020c, Duran-Lopez *et al* 2020, Narin *et al* 2021). However, while x-ray scans provide a fast, cost-effective examination of the chest, a CT scan provides a more detailed 3D scan. Hence, 2D-CNN models based on chest CT images also exist, for instance (Mei *et al* 2020) used CT images in their integrated model combining predictions from CT images only, non-image information only (i.e. demographic and clinical data), and the combination of image and clinical data. As a pre-processing step most 2D approaches obtain the lung-mask using segmentation or morphological operations and then make a decision using the lung region of the CT image. For instance, Hu *et al* (2020) first utilized a 2D segmentation network and then used the segmented lung image for classification of COVID-19 patients from community acquired pneumonia (CAP) and non-pneumonia scans.

2.1.2. 3D-CNN modelling

Considering the 3D structure of CT sequences, several recent methods have exploited 3D-CNNs for modelling COVID-19. Wang *et al* (2020d) employed a deep learning method for diagnosis where lung segmentation is

performed, and then the segmentation result is taken as the input of the 3D-CNN to predict the probability of COVID-19. In Gozes *et al* (2020a) a 3D model is added based on 2D analysis of each slice, where the 3D-CNN analyzes the volume for nodules and focal opacities. Though 3D CT scans can provide abundant stereoscopic information of lung involvement in COVID-19, the calculation and memory load of 3D models cannot be ignored. In our study, we have opted not to implement a 3D-CNN model and have instead attempted to mimic the way the clinicians make their decisions based on 2D CT images.

2.2. Lesion identification

Despite the significant work on lesion detection which already exists, identifying specific lesion types from COVID-19 with solely image-level supervision is challenging. Previous studies have mainly focused on separating the lesion region from the image background. With several studies employing U-Net (Ronneberger *et al* 2015) to segment the lung CT scans. For instance, to distinguish COVID-19 pneumonia from CAP (Li *et al* 2020) or to segment pulmonary opacities in the lungs to obtain quantitative measurements (Huang *et al* 2020b, Cao *et al* 2020), for longitudinal assessment of the disease. Extending beyond U-Net, Jin *et al* (2020) proposed UNet++ segmentation for locating lesions and Wang *et al* (2020b) proposed a novel COVID-19 pneumonia lesion segmentation network which is robust to noise and can deal with lesions with various scales and shapes. In addition, there is an extensively exploited lesion segmentation framework called VB-Net (Shan *et al* 2021), and with this method, Shi *et al* (2021) designed hand-crafted features based on the segmented COVID-19 infection and then screened the COVID-19 positives. However, these works all require accurate lesion annotations for the training procedure.

To address the issue of annotations, recently much effort has been directed to weakly supervised lesion detection in an attempt to achieve equivalent performance to fully supervised approaches. The use of weak supervision, e.g. image-level classification labels, relieves the rigid demand for lesion-wise annotations at a pixel-level. The class activation map (CAM) (Zhou *et al* 2016), is an effective way to localize the lesion region within CT images using diagnostic labels solely. In Hu *et al* (2020) the CAM is regarded as the class-specific saliency map and the saliency maps from different layers are joined for lesion segmentation. While Wang *et al* (2020d) combined CAM activation regions with the output of a 3D segmentation network for final lesion localization. These methods resort to CAM to indicate the pixel-wise distribution of lesions. Nonetheless, the detected lesions are class-specific and can only cover a broad range of suspected abnormal regions.

Our proposed lesion identification approach is weakly-supervised, relieving the time- and labour- intensive labelling work required by methods that need accurate pixel-wise information for lesion segmentation. Moreover, unlike previous work on lesion detection, our work aims to automatically make a differential diagnosis and use this to identify the characteristics of the COVID-19 lesion. Hence, this work requires not only high classification accuracy but also needs to identify the unique patterns of COVID-19 lesions in CT images. And in that way, the linking of clinical manifestations and provision of clinical guidelines can be allowed.

3. Methodology

In this section, we provide details of the proposed framework, which is illustrated in figure 2. In brief the framework consists of the following modules:

(i) A feature pyramid network (FPN) embedded convolutional neural network (FeCNN), to extract the multi-scale deep features of the CT image.

(ii) An ensemble feature fusion module, which integrates the multi-scale feature to make a diagnosis prediction.

(iii) A connecting module, the Grad++ module, which combines the feature maps from the last layer of network outputs and the back-propagated gradient derived from the predicted probability to generate the lesion activation heatmap (LAHm).

(iv) A multi-lesion detector, which is used to encode all the potential lesions and form a lesion distribution space using a self-supervised clustering algorithm.

(v) A clustering module, which projects the encoded lesion into the lesion space and yields the final lesion clustering module.

The FeCNN and ensemble feature fusion allow us to obtain the likelihood that the CT image is COVID-19 positive; the Grad++ then serves as a transition module connecting the diagnostic network branch to the lesion identification branch via feature maps generated by the inference procedure. In the lesion branch, the LAHm

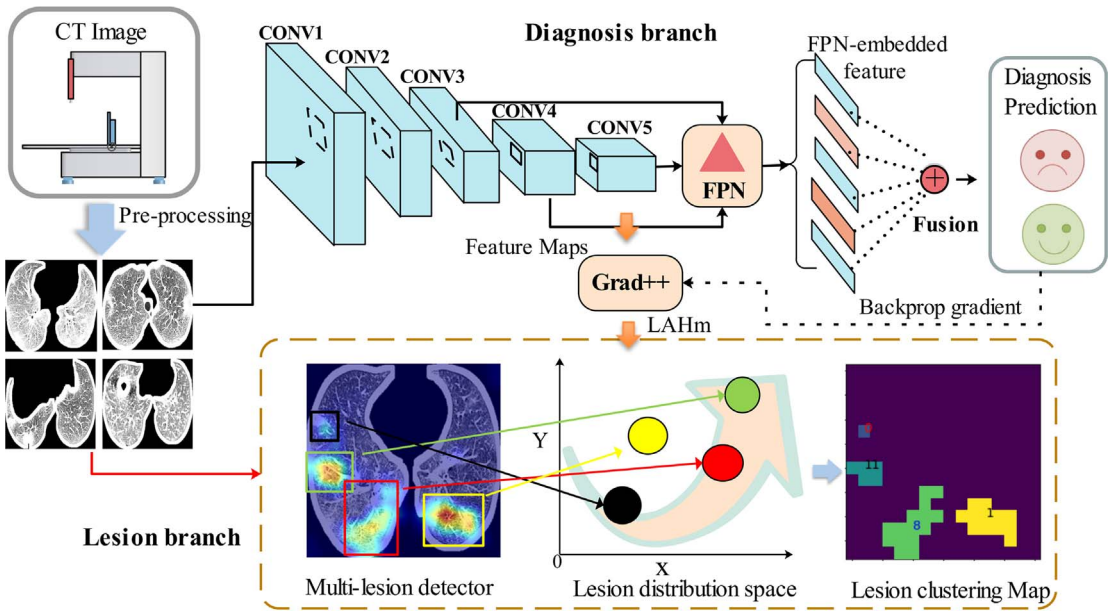

Figure 2. The pipeline of the proposed framework.

yielded by the Grad++ module is leveraged to capture the spatial information of multiple lesions which helps us locate the lesions at a pixel-level.

3.1. FPN-Embedded convolutional neural network (FeCNN) for CT slice prediction

As a supervised approach for diagnosis, FeCNN is used to learn a mapping Φ: $\mathcal{X} \rightarrow \mathcal{L}$, given the collection of training samples $\mathcal{T} = \{(x_n, l_n)\}_{n=1}^{N}$. Here, $\mathcal{X}$ denotes the data space and $\mathcal{L}$ represents the annotation space. N is the number of training samples, $l_n \in \mathcal{L}$ is the associated diagnostic label of the input CT image $x_n \in \mathcal{X}$. Specifically, the proposed FeCNN has three central parts: the backbone of the network, the FPN module, and the fusion module.

The backbone network takes advantage of the powerful encoding ability of CNNs in medical image analysis tasks. As an encoding module of the diagnosis branch, the backbone network will utilize a consecutive block structure to extract the feature maps of pre-processed CT images at each of the corresponding scales. The block-based architecture of the backbone network can facilitate the enhancement of multi-scale image features; considering this a FPN is included in the framework. FPNs have shown strong performance in computer vision tasks due to their reuse of features. The FPN leverages the pyramid-like shape of multi-scale feature maps to form a feature pyramid, aiming to further improve the semantic representation of the input feature maps at single scale. In this study, our FPN module echoes the classical plug-in proposed in Lin *et al* (2017). The features generated at each scale level of the FPN all produce a valid prediction, potentially resulting in variance in the diagnoses. Hence, inspired by the idea of ensemble fusion, a fusion module is designed to integrate the enhanced features from each scale level for the final prediction.

3.1.1. The backbone network

Ideally, the architecture of the backbone network is flexible, which means that it can be an encoding module of any classical CNN: Resnet (He *et al* 2016), VGG (Simonyan and Zisserman 2014), etc. They all share the same block-based network architecture and similar in-block elements, i.e. the convolution operation, the nonlinear activation, the batch normalization and the pooling operation. Stacking these elements with different schemes or fine-tuning their configuration provides the CNNs with distinct in-block structures.

In our work, the backbone network has five convolutional blocks denotied Conv1,…, Conv5 in figure 2. The details of the backbone network's stacking schema are illustrated in figure 3, where the numbers k, q, s in the green rectangular box represent a convolutional operation with q, $k \times k$ filters and stride, s, (q and k denote the number and the size of the convolutional kernel). Re and MP denote the residual operation and 3×3 max pooling operation, respectively. Each block yields the image feature map with the corresponding scale.

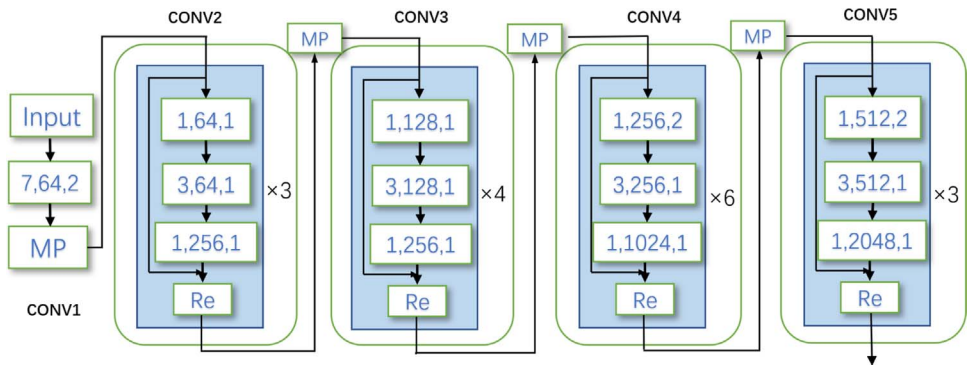

Figure 3. The stacking schema of the backbone network. The black arrow indicates the direction of data flow.

3.1.2. FPN module

Typically, as the depth of the backbone network increases, the feature output of the deeper convolutional blocks will provide a more robust representation with strong semantics. That is also why the output features from the last layer are emphasized in most computer vision tasks. Nevertheless, compared with natural images, CT images have less object-level semantic information, and most are filled with dark pixels. At the same time, lesions in CT images show a variety of shapes, textures, and features which a single scale cannot capture. Thus, careful extraction of multi-scale semantic details is crucial and therefore, we utilize the FPN module for semantic enhancement.

Figure 4 shows the stucture of FPN; as the outputs of the first two blocks, Conv1 and Conv2 have a high computational load but only low semantic details, the feature maps of only the final three blocks Conv3, Conv4, and Conv5 (denoted as C3, C4 and C5 respectively) are fed into the FPN module. These feature maps are first convolved with a 1×1 kernel and batch normalization applied to obtain a corresponding feature map with lower dimension. The feature maps of the two relatively higher scales, i.e. C4 and C5 are upsampled to the spatial resolution of the next levels down, and thus the spatial resolution of the lower scale and the semantic information from the higher scale can be combined. Finally all three scales are convolved with a 3×3 kernel with 1-stride. Additionally to provide greater multi-scale information, two consecutive 2-stride 3×3 convolution operations are performed on C5 to obtain fine-grained feature maps C6 and C7.

To summarize the information in the multi-scale feature maps, we include a global average pooling operation (GAP). Before conducting the GAP operation, a 1×1 convolution and batch normalization is applied to the feature maps in order to keep the same dimensions, which we set to 256 in our experiments. After applying GAP operation on the multi-scale feature maps, we obtain a $1 \times 1 \times 256$ vector for each scale level.

3.1.3. Ensemble fusion

Let $\boldsymbol{P}_i \in \mathbb{R}^{1\times 256}$ indicate the feature vector of the ith scale of the FPN module. Since the variance across the scales is significant and also robust, the summarized feature vector obtained from each level can facilitate the final prediction. However, this can potentially lead to variability in the diagnoses which can be challenging to unify. Therefore, we implement an ensemble fusion method to aggregate the separate feature components.

Specifically, for each scale level $\boldsymbol{P}_i$, $i \in [0,\ 5]$, we add a ρ-way fully-connected layer with ReLU activation function to calculate a scale-level score, $\boldsymbol{s}^i$, (where ρ corresponds to the pre-set length of the scale-level score) such that

$$\boldsymbol{s}^i = \max(\boldsymbol{0},\ \boldsymbol{W}_i^T \boldsymbol{P_i} + \boldsymbol{b}_i^T), \qquad \boldsymbol{W}_i \in \mathbb{R}^{256\times C}, \quad \boldsymbol{b}_i \in \mathbb{R}^{256\times C}, \tag{1}$$

where, $\boldsymbol{W}_i$ and $\boldsymbol{b}_i$ are the weights and the bias of the fully-connected layer which are to be trained. Then, the associated weight of each feature vector, ω_i, is calculated as:

$$\omega_i = \frac{1}{2}\log\left(\frac{1-\epsilon_i}{\epsilon_i}\right), \tag{2}$$

where ϵ_i is the error rate of the scale-level prediction such that

$$\epsilon_i = \frac{1}{N}\{\mathrm{argmax}(\boldsymbol{s}^i,\ axis = -1) \neq l_n\}_1^N. \tag{3}$$

After calculating the weights, the ensemble fusion module aggregates the weighted score vectors to produce a new feature vector for the final prediction. If the final feature vector is denoted as $\boldsymbol{d} \in \mathbb{R}^{\rho\times 5}$, where

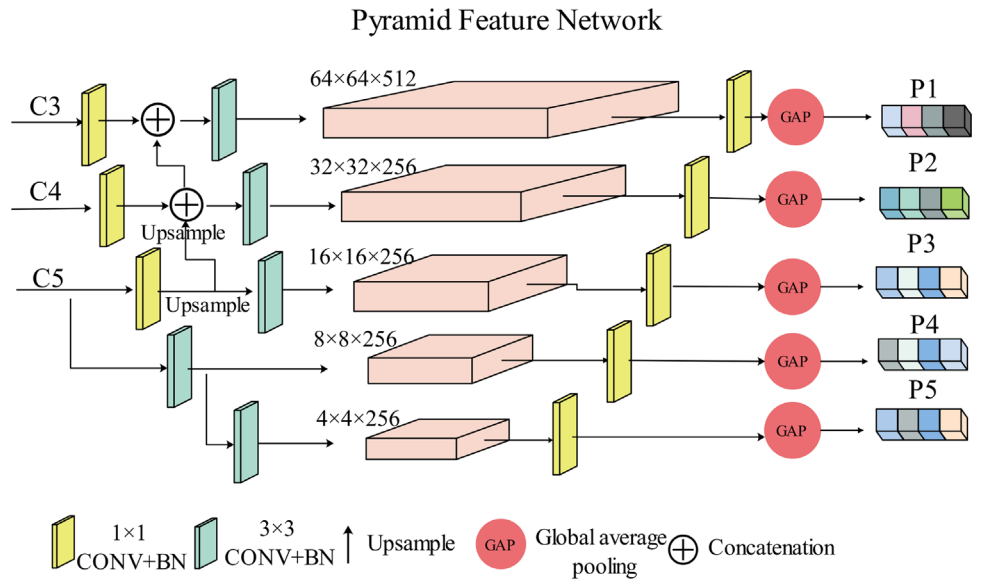

Figure 4. The feature pyramid network is utilized to reuse features of different scales extracted from the backbone network for the semantic enhancement.

$\boldsymbol{d} = [w^0\boldsymbol{s}^0,\ w^1\boldsymbol{s}^1,\ \ ..,\ w^i\boldsymbol{s}^i]^{\mathrm{T}},\ i \in [0,\ 5]$, then the final step is to add a $\mathcal{C}$-way fully-connected layer (where $\mathcal{C}$ corresponds to the number of categories). Softmax is used to predict the likelihood that the input CT slice belongs to each category and the diagnostic probability can be defined as:

$$\Phi(y_n = j|\ x_n,\ \boldsymbol{d},\ \boldsymbol{W}) = \frac{\exp\left(\mathbf{w}_j^T\boldsymbol{d}\right)}{\sum_i^{\mathcal{C}}\exp\left(\mathbf{w}_i^T\boldsymbol{d}\right)}, \qquad j \in [0,\ \mathcal{C}], \tag{4}$$

where y_n is the predicted label of the nth input CT image, and $\boldsymbol{W} = \{\mathbf{w}_j\}_{j\in\mathcal{C}}$ is the set of weighted parameters of the function Φ. Thus, the model is trained by minimizing the loss function:

$$L_{loss} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{j=0}^{\mathcal{C}} True(y_n == l_n)\log\left(\Phi(y_n = j|\ \boldsymbol{d},\ \boldsymbol{W})\right). \tag{5}$$

where, $True(\cdot)$ is a Boolean function such that $True(\cdot) = 1$ if the condition is true and 0 otherwise. As the diagnostic network is designed to predict if the CT slice is COVID-19 positive or not, $\mathcal{C}$ is set to 2 in our framework.

3.2. Lesion identification

Since the image-level label is the only human-annotated supervision that is used in our study, identifying the lesion at a pixel-level is a challenging task. However, if a trained model is able to predict whether a CT image is COVID-19 positive or not with high accuracy, it must have captured a reliable set of lesion features from the input image. Motivated by this fact, we propose taking those lesion features as clues to identify the type of lesion. Nevertheless, the difficulty with this idea is that these lesion features are theoretically invisible and unavailable. Recently, several methods have turned to CAM using the class-specific map to weakly localize the lesion area (Wang *et al* 2020d, Hu *et al* 2020) or show the suspected lesion region in order to demonstrate that the CNNs are making the correct decisions (Wang *et al* 2020d, Jin *et al* 2020). Inspired by these studies, we utilize the Grad++ module to reveal the underlying lesion using the back-propagated gradient and then leverage the multi-lesion detector to capture them.

In detail, the lesion identification problem can be formalized as follows. Let $X = \{\mathrm{x}_n\}_{n=1}^{m} \subseteq \mathcal{X}$ be a set of COVID-19 positive CT slices predicted by the diagnosis branch, where $True(\mathrm{argmax}(\Phi(x_n)) == l_n) = 1$. The target of the lesion prediction is to learn a function: $\mathcal{F}$: $X \rightarrow S$, where S is a set of possible lesion scores. For the multi-lesion detector, it is developed using the lesion activation heatmap (LAHm) generated by the Grad++ module. The intensity of the pixels in the LAHm corresponds to the presence of lesions, with which the detector can encode the lesions without explicitly considering the lesions' pixel-wise information, i.e. the shape, textures, etc. Hence, let $\mathcal{D}$ denote the detector, then the process of lesion identification can be formalized as $\mathcal{F}(\mathcal{D}(X)) \rightarrow S$.

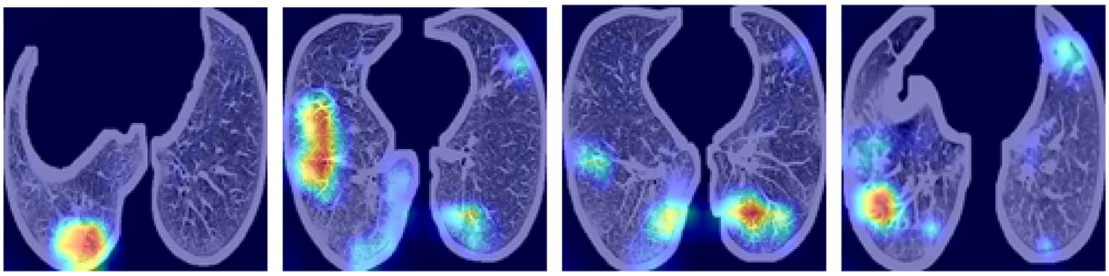

Figure 5. Example LAMs visualized in a heatmap format using a Jet color scheme, where deep color corresponds to higher activations and light to lower activations.

3.2.1. Grad++ module

Let the feature maps output by the final block of the backbone network in the FeCNN be $A_n \in \mathbb{R}^{w\times h\times \kappa}$, where w, h, κ denote the width, height, and number of feature maps respectively. Then, the lesion activation heatmap (LAHm) of the diagnosed image x_n is computed as the linear combination of $\{A_n^i\}_{i=0}^{\kappa}$, where the weight of each feature map is calculated as described in Selvaraju *et al* (2017):

$$\alpha^{\kappa} = \frac{1}{w * h}\sum_{i}^{w}\sum_{j}^{h}\frac{\partial \Phi}{\partial A_{ij}^{\kappa}}, \tag{6}$$

with α^{κ} obtained by calculating the partial derivatives of $\Phi(x_n)$ with respect to A_{ij}^{κ}. Then the LAHm is generated as follows:

$$L_{LAHm} = \mathrm{Re}\,LU\left(\sum_{\kappa}\alpha^{\kappa}A^{\kappa}\right). \tag{7}$$

The LAHm is a positive-case-specific saliency map, which activates the pixels of the feature map that contribute to the diagnostic model making a prediction. The discriminative CT image patterns which emerge are those corresponding to the COVID-19 positive cases (i.e. the lesions the CT slice shows).

To further analyze the lesions, the LAHm is first normalized to [0,1] and then thresholded using Otsu's method (Otsu 1979) to obtain a preliminary binary mask which segments the LAHm into lesion region and background. With the binary mask obtained, the components are searched to detect potential lesions by identifying connected components and eliminating fuzzy boundaries between them. To provide a better separation of lesions the search mode is set to 4-connectivity, meaning that those pixels within 4 orthogonal hops are considered as neighbors. Since 1 pixel on the LAHm equates to a lesion area of 1024 (i.e. 32×32) when projected to the resolution of the original image. To preserve as many potential lesion regions as possible, connected components with more than 1 pixel are accepted as candidate lesion regions. These two steps ensure that, as much as possible, all potential lesions are kept and the boundaries between the lesions are also as clear as possible. Inspired by Zhu *et al* (2017), the LAHm and the binary mask are coupled via Hadamard product and a lesion activation map (LAM) is created by upsampling the coupled map to the original image for potential lesion localization, example LAMs are shown in figure 5.

3.2.2. Multi-lesion detector

If we denote the LAHm of the input lung image x_n as $L \in \mathbb{R}^{w\times h}$, where w and h correspond to the width and height of the feature map A. The multi-lesion binary mask divides the LAHm into multiple lesion regions. If we recall that the higher the value of the LAM, the higher the corresponding local area's contribution to predicting the correct diagnostic label. Thus, we adopt the method used in Lin *et al* (2020) for multi-lesion feature extraction. That is, for each potential lesion region, we locate the spatial maxima:

$$(x^*, y^*)_i = \mathrm{argmax}\,(L^i), \qquad i \in [0, m], \tag{8}$$

where L_i is ith sub-region of L, the candidate lesion locations and m is the number of candidate lesions. These extrema correspond to points deemed maximally salient for the differential diagnosis task by the proposed network.

Considering the LAHm is the weighted linear combination of the feature maps, $A \in \mathbb{R}^{w\times h\times \kappa}$, used by the classification network, we utilize these spatial maxima $(x^*, y^*)_i$ to extract a local feature vector describing the ith lesion region based on the corresponding component of A for each channel. Thus, the lesion detector $\mathcal{D}$ can be formalized as follows:

$$\mathcal{D}(x_n^i) = A((x^*, y^*)_i), \qquad i \in [0, m], \tag{9}$$

where x_n^i denotes the ith lesion of the input image x_n.

Typically, the feature maps A have low resolution but high channel dimensionality, in our case $\mathcal{D}(x_n^i)$ is $16 \times 16 \times 2048$. Therefore, the length of the encoded lesion feature is 2048. If we run the feature detector for each candidate lesion of input CT image, then, a set of the feature vectors for multiple lesions is yielded, on which we build a feature space for all input CT images.

3.2.3. Lesion clustering

As there are no extra supervised annotations except the image-level label, the identification of the lesions is accomplished by clustering lesions based on the encoded lesion representation using an unsupervised machine learning approach. Simply, we apply a k-means clustering algorithm on the extracted lesion feature space to group the detected lesions. Where each of the K lesion clusters represents a potential lesion type. Hence, for a predicted positive CT slice, x_n, the lesion score ls_n^i of the ith detected lesion is defined as:

$$ls_i = \min_k \frac{e^{-d(i,k)}}{\sum_k e^{-d(i,k)}}, \tag{10}$$

where $\{ls_n^i\}_{i=0}^m \subseteq S$, $d(\cdot)$ is the euclidean distance between the ith lesion and the kth optimal cluster centre. This formulation produces a smooth probability distribution over the K clusters, in which the lesion score decreases the likelihood of belonging to this cluster increases.

4. Experimental setup

The following section describes our experimental setup including the datasets used to evaluate our proposed framework and the metrics used to evaluate the performance.

4.1. Dataset

With Institutional Review Board (IRB) approval, 39 CT scans collected from 13 patients at the First Affiliated Hospital of Shantou University (denoted as Own) are included in this study and all patients included in this dataset have provided written informed consent. Besides our own dataset, we also evaluate the performance of the proposed frame-work on two publicly available data sources, the small scale dataset: Radio-2 (Knipe and Iqbal 2020), and the compound COVID-19 dataset (COVIDx2a) (Gunraj *et al* 2020). The COVIDx2a dataset itself has been collected from several different data spaces including the China National Center for Bioinformation (CNCB) (Zhang *et al* 2020b), the COVID-19 diagnosis dataset (CTset) from Negin Radiology Medical Center (Iran) (Rahimzadeh *et al* 2021), and the CT dataset provided by the multi-national, national institutes for health (NIH) cnsortium for CT AI in COVID-19 via the Cancer Imaging Archive (TCIA) public website (Clark *et al* 2013, An *et al* 2020, Harmon *et al* 2020). The dataset from the TCIA public website is also the official data used by the MICCIA grand challenge on COVID-19 lesion segmentation 2020 (An *et al* 2020). Not all of the slices from these sources were used in COVIDx2a and the data for other types of pneumonia have been excluded in our experiment. In total 155 541 CT slices including 94 548 from positive patients were used in this investigation. The details of these datasets, including the number of positive cases, the number of positive slices and the annotations, are listed in table 2.

Note that only a portion of the CT slices from the Corona (Ma *et al* 2020) and CNCB (Zhang *et al* 2020b) datasets are released with annotations of the infection area. Of the 750 CT slices from 150 COVID-19 patients of the CNCB dataset, 549 of these slices are also marked with lesion type (e.g. 2 in the lesion mask denotes GGO, 3 denotes consolidation).

4.2. Network implementation details

4.2.1. Basic setup

In the diagnosis branch, the backbone network for CT image feature extraction is initialized with the no-top-weights trained from ImageNet. The ρ of the fusion module in the diagnosis branch is set to 2. For each database, 80%, 15%, and 5% of data split randomly for training, testing, and validation. The network was trained for 50 epochs using Adam optimizer with a constant learning rate of 1e-5 and a dropout rate of 0.5, the batch size is set to 10.

4.2.2. Preprocessing

All CT images of each dataset were preprocessed in a unified manner before training and testing. A normalization window was first set to normalize each image to 8-bit pixel intensity values, i.e. 0-255. And the lung was segmented out by morphological operation. Since the lung of several CT images at the beginning and

Table 2. Details of the datasets used in this study. Pos. refers to the number of positive COVID-19 cases and Slices are the corresponding positive slices. Img., Les. A. and Les. T. refer to the annotation of image category, lesion area and lesion type respectively. The checked box denotes that part of the dataset has this annotation.

Dataset		Details		Annotation		
		Pos.	Slices	Img.	Les. A.	Les. T
COVIDx2a (Gunraj *et al* 2020)	Radio-1 (Bell and Hacking 2020)	—	3175	✓		
	CTset (Rahimzadeh *et al* 2021)	95	2282	✓		
	LIDC (Armato *et al* 2015)	—	3999	✓		
	HUST (Ning *et al* 2020)	1521	37 306	✓		
	TCIA (An *et al* 2020)	632	11 818	✓		
	Corona (Ma *et al* 2020)	20	1213	✓	☑	
	CNCB (Zhang *et al* 2020b)	409	31 070	✓	☑	☑
Radio-2 (Knipe and Iqbal 2020)		9	829	✓	✓	
Own		13	2856	✓	☑	☑

end of a CT sequence is usually closed, the average pixel intensity per CT image in the sequence was calculated and they were discarded if their average pixel intensity was below 0.08. After that, all the cropped lung images are resampled to the same spatial resolution, 512×512. Inputting the lung region instead of the whole CT image manually helps our model focus on pulmonary differentiation, ignoring the effects of air or fat.

4.2.3. Data augmentation

As one method to tackle the overfitting problem, a data augmentation scheme was applied in the training stage. The data augmentation included a random affine transformation and color adjustment. The affine transformation was composed of rotation (0°–360°), horizontal and vertical flip, and resolution shifting (0.05). The color adjustment includes brightness ($0\% \pm 50\%$) and contrast ($0\% \pm 30\%$). For each training sample, the parameters were randomly generated, and the augmentation was identically applied.

The training procedure of our FeCNN was carried out on a NVIDIA RTX 2080ti GPU with 11 GB of GPU memory. During the testing procedure, the data augmentation strategy was not applied. The trained model gives the diagnostic probability as the likelihood of being COVID-19 positive. Using the predicted probabilities and corresponding ground-truth labels, statistical analysis of the model performance is conducted.

4.3. Evaluation metrics

We independently evaluated the performance of each of the three parts of our model: the diagnosis prediction, lesion detection, and lesion identification, for the datasets described in section 4.1.

4.3.1. Diagnosis prediction

To evaluate the performance of the diagnosis network, the testing dataset was used with the trained model. For each testing CT image, the COVID-19 positive and negative probability are predicted. The performance is evaluated against the ground truth labels through the diagnostic accuracy, the precision-recall (PR) curve, and the receiver operating characteristic (ROC) curve. If the true positives (TP) are the number of correctly detected COVID positive cases, the false positives (FP) the number of detected positive cases that are actually negative, and the false negatives (FN) are the number of rejected positive cases that are truly positive. Then the precision $= \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ and the recall $= \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$. The ROC curve is created by plotting the TP rate (TPR) against the FP rate (FPR) at various thresholds. Finally, the average precision (AP) and the area under ROC curve (AUC), which summarize the PR curve and ROC curve, are also calculated.

4.3.2. Lesion detection

To quantitatively analyse the performance of our weakly-supervised lesion detection module, and in line with the results presented in Wang *et al* (2020d) we calculate the lesion hit rate as the evaluation metric. First bounding boxes for the highlighted regions of the LAM are calculated, by employing the connected component operation. This is then repeated for the ground-truth (GT) lesion boxes, marked by the GT lesion masks. Next the ratio of the area of the LAM box which overlaps the GT lesion box is calculated. To determine if the lesion is successfully detected we check if the ratio is over a specified threshold and the spatial maximal of the LAM box is inside the overlapping region. The hit rate is then calculated as the quotient of the number of successful hits and the number of the GT lesions.

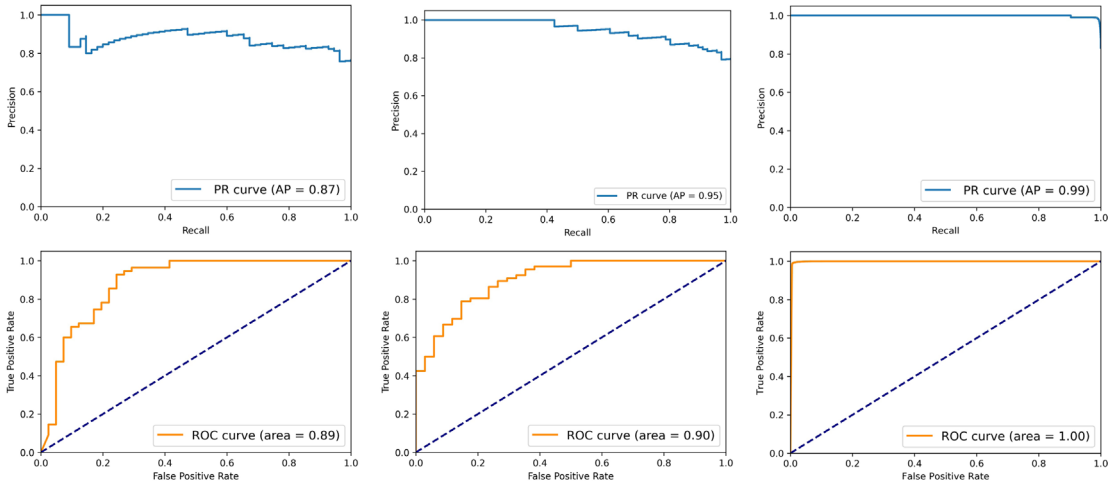

Figure 6. The PR curves (top) and the ROC curves (bottom) of data radio-2 (left), Own (middle) and COVIDx2a (right).

4.3.3. Lesion identification

The evaluation of the lesion detection is on the lesion-level. For the lesion clusters of a dataset, experienced radiologists label the detected lesions. According to the guidance of the specialist, we then calculate the sensitivity (SEN) and specificity (SPE) of the clusters as the evaluation metrics, where SEN = TPR, SPE = 1—FPR. TPR is the ratio of the number of correctly identified lesions over the total number in a cluster and FPR is the ratio of the number of incorrectly identified lesions over the total number in a negative cluster.

5. Results

In this section, we present our results and evaluate our method against the state-of-the-art in order to validate the effectiveness of our framework for COVID-19 classification and lesion identification.

As COVIDx2a is a compound dataset, we first evaluated the classification performance of our diagnostic network on this data. To validate the performance on the available independent datasets, the model is tested on our Own dataset and the Radio-2 (Knipe and Iqbal 2020). The testing threshold is set to 0.5, i.e. if the predicted likelihood of COVID-19 is over 0.5, the CT image is classified as COVID-19 positive, and vice versa. Overall, our FeCNN achieves accuracies of 0.99 for the compound dataset COVIDx2a, and of 0.95, 0.85 for the other two independent datasets Own and Radio-2, respectively. The accuracies for positive and negative classification are above 0.99 and 0.98 on the dataset COVIDx2a, and on the other two datasets, the positive classification metric is 0.94 and 0.83, with the corresponding negative classification results being approximately 0.91 and 0.96 respectively. Using these results, the evaluation metrics introduced in the previous section are calculated.

5.1. COVID-19 diagnostic prediction

Figure 6 shows the PR curves and the ROC curves of the three datasets, respectively. From the PR curves we can see that our model exhibited relatively high discrimination of COVID-19 positive cases, especially on the datasets Own and COVIDx2a, with both datasets showing high AP values of 0.95 and 0.99, respectively. For the dataset Radio-2 as the recall increases the precision fluctuates in a narrow range and when the recall is nearly 1, the precision is almost 0.8. The results obtained from the COVIDx2a dataset are particularly impressive (see figure 6), even when the recall reaches 0.90, the precision is still over approximately 0.95. The likely explanation for this result is due to the fact that our diagnosis network tends to predict with high certainty (e.g. with a value of either 0.99 or 0.01). If we analyze the distribution of predicted probabilities, we learn that the percentage of slices with probability ranging from 0 to 0.1 and from 0.90 to 1 is up to 83% on average (this percentage is over 90% in COVIDx2a). If we look at the corresponding ROCs, then for the Radio-2 the model obtains an AUC of 0.89 and similarly for the our Own dataset the AUC is 0.90. In both cases when the FPR is less than 0.3 the FeCNN cannot achieve robust performance. In contrast if we compare the COVIDx2a, the results are much higher with an AUC of approximately 1.00. This somewhat surprising result is similar to the result reported in the study (Gunraj *et al* 2020), which can probably be attributed to the COVIDx2a having been filtered to the common abnormal CT slices by an experienced radiologist; in other words, only CT images showing the significant variance are kept.

To further validate the effectiveness of the diagnosis prediction, we evaluated it against existing baseline methods, including those based on the combination of hand-crafted features and a classifier and those based on

Table 3. The diagnostic accuracy of the proposed method and the baseline methods on datasets: Own, Radio-2 (Knipe and Iqbal 2020), and COVIDx2a (Gunraj *et al* 2020).

		Dataset		
	Method	Own	Radio-2	COVIDx2a
Feature-based methods	LBP + SVM	0.717	0.789	0.898
	LBP + MLP	0.691	0.733	0.887
	64 hist+SVM	*0.718*	*0.807*	*0.939*
	64 hist+MLP	0.706	0.744	0.896
Deep learning methods	Resnet50	0.908	0.851	0.976
	VGG16	0.854	0.825	0.955
	Xception (Chollet 2017)	0.913	0.854	0.979
	WsNet (Panwar *et al* 2020)	0.928	0.854	0.986
	DeCoVNet (Wang *et al* 2020d)	0.916	0.847	0.987
	COVIDNet (Gunraj *et al* 2020)	0.944	0.849	0.991
	FeCNN	**0.957**	**0.859**	**0.994**

Table 4. The effectiveness of different modules in the diagnosis branch. The diagnostic accuracy for each of the datasets is given for four different combinations of the modules.

	Modules			Accuracy		
#	Backbone	FPN	Fusion	Own	Radio-2	COVIDx2a
1	✓			0.908	0.851	0.976
2	✓	✓		0.922	0.843	0.983
3	✓		✓	0.941	0.853	0.987
4	✓	✓	✓	**0.957**	**0.859**	**0.994**

deep-learning techniques. In particular, we applied two image feature descriptors the local binary pattern (LBP) operator (Zhang *et al* 2004), and 64 bins grey-scale histogram (Hist). We also show the results of these features with two different classifiers, the three-layer multi-layer perceptron (MLP) and the support vector machine (SVM). Here, the three layers of the MLP have 25, 10, and 2 nodes respectively and are composed of a batch normalization operation, a fully connected layer, and a tanh activation function, and the L2 penalty (regularization term) parameter is 0.01. The baseline classification CNNs tested are the Resnet50 (He *et al* 2016), VGG16 (Simonyan and Zisserman 2014), Xception (Chollet 2017). The architecture of the backbone networks remain unchanged, but two dense layers with ReLU activation function and one dense layer with softmax are added at the top to map the output of the CNNs to COVID-19 likelihood. In addition, three recent methods which have been demonstrated to be effective are included for comparison, these are: the weakly supervised multi-scale network used in Hu *et al* (2020) (denoted as WsNet), the 2D DeCoVNet proposed in Wang *et al* (2020d) and the COVIDNet-CT reported in Gunraj *et al* (2020). All the baseline methods are trained in the same environment as the proposed framework.

Table 3 reports the performances of all of the methods in terms of accuracy on the testing datasets. From this table, we can observe that within the hand-crafted feature-based approaches, the diagnostic performance of the feature descriptors with SVM are better than with MLP, and for all three datasets, the combination of 64 Hist + SVM, which reaches accuracies of 71.78%, 80.72% and 93.92% respectively, is more persuasive than the other feature-based methods. Compared with the feature-based methods, the CNNs show competitive results, which are largely capable of beating the feature-based methods. Among the baseline CNNs, Xception could generally achieve the state-of-art performance. The WsNet, DeCoVNet, COVIDNet and FecNN all achieved similar performance for the Radio-2 dataset. This can be attributed to the fact that all four networks share similar backbone architecture (e.g. all are embedded with the residual connection), and the volume of data limits any divergance in performance. For the other two datasets the FeCNN achieved better performance than the other CNNs, surpassing the maximum by over 1.3% for dataset Own and 0.3% for dataset COVIDx2a. The advantage of our model is not as great for the COVIDx2a dataset, but as highlighted earlier the COVIDx2a dataset has already been filtered to include only CT images with significant variance therefore making it easier to discriminate between the COVID-19 positive and negative cases which is indicated by the very high diagnostic accuracy of all of the CNN methods.

To verify the effectiveness of the different modules in the diagnosis branch we carried out an ablative study. The baseline is the single backbone network, without the FPN and the ensemble fusion, on top of which the original fully-connected layer is added for the diagnosis. From the final results (see table 4), we can conclude as

Table 5. Hit rate results for weakly-supervised lesion detection with different overlap thresholds.

	Dataset		
Threshold	Ratio-2	COVIDx2a	Own
0.1	72.4%	75.3%	74.1%
0.2	70.5%	74.5%	72.3%
0.3	68.5%	73.5%	66.5%
0.4	64.8%	69.8%	65.3%
0.5	63.2%	68.3%	65.3%
0.6	60.7%	59.2%	52.7%
0.7	50.6%	50.3%	40.2%
0.8	45.2%	33.1%	20.9%
0.9	30.2%	27.9%	15.8%

Table 6. The results of weakly-supervised lesion detection.

	Dataset		
Method	Ratio-2	COVIDx2a	Own
LAM	**63.2**%	**68.3**%	**65.3**%
CAM	35.8%	45.8%	38.5%
Norm-grad	34.2%	43.1%	41.4%

follows that: first, adding either the FPN or the fusion module can improve the performance on dataset Own but this is not significant for the other two datasets; second, overall the addition of the fusion module to the backbone network can improve the performance more than the backbone network with just the FPN; third, the combination of all three modules, as in our configuration, achieved the best performance surpassing the baseline by more than 3% accuracy on average across all three datasets. These results illustrate the effectiveness of these two components and, thus, suggest that CNNs with the FPN module and the fusion module have the potential to lead to significant progress in deep-learning methods for COVID-19 diagnosis.

5.2. Lesion detection

Table 5 reports the hit rates achieved when the overlap threshold used to determine whether a lesion is successfully detected or not is varied. The results in table 5 are for the datasets Radio-2, a subset of the COVIDx2a dataset—as the examples are from the CNCB dataset we denote it more specifically as CNCB, and our Own dataset. As can be seen, when the threshold is 0.1, the hit rate reaches 72.4%, 75.3%, 74.1% on the three datasets respectively. Since this threshold is relatively small, these results are more indicative of the percentage of the spatial maxima of each identified lesion which are correctly located in the overlapping region of a true lesion. We can also see that the hit rates for the different datasets follow a general trend: as the threshold increases, the hit rate gradually decreases. This is to be expected, as a successful hit needs to ensure the location of the spatial maxima is inside the overlapping region, and then the area of the overlap exceeds the threshold.

If we set the overlap threshold to 0.5, i.e. in a successful hit, the spatial maxima of the segmented lesion needs to be inside the overlapping region and the area of the overlap needs to be over half of the union pixels, then our weakly-supervised framework achieves hit rates of 63%, 68% and 65%, respectively. While these results are not especially high, considering there is no pixel-level lesion annotation they are acceptable. If we compare our results with the CAM method and the recently proposed weakly supervised method, Norm-grad (Rebuffi *et al* 2019), both achieve 39% on average making our results a significant improvement (see table 6). In principle, the generation of the LAM shares the rationale with these two methods. Thus, we can attribute the improvement in the hit rate for the lesion detection to the LAHm with morphological operation, which allows us to keep most of the potential lesions and separates the lesions by eliminating the fuzzy boundaries between lesions, thus improving sensitivity to the distribution of the actual lesions, and improving the hit rate.

Example results of the lesion detection on datasets Radio-2, CNCB, and Own are given in figure 7, showing the original positive CT image, its cropped and resized lung image, the corresponding LAMs and the detected lesion map. As the other parts of COVIDx2a lack lesion annotations, examples of LAMs for these datasets are given in figure 8.

From the examples in figures 7(a) and (c), we can see that the detected solid boxes generally cover the labelled lesion regions. However, from the detected lesions, we can also observe that the LAM has two main limitations.

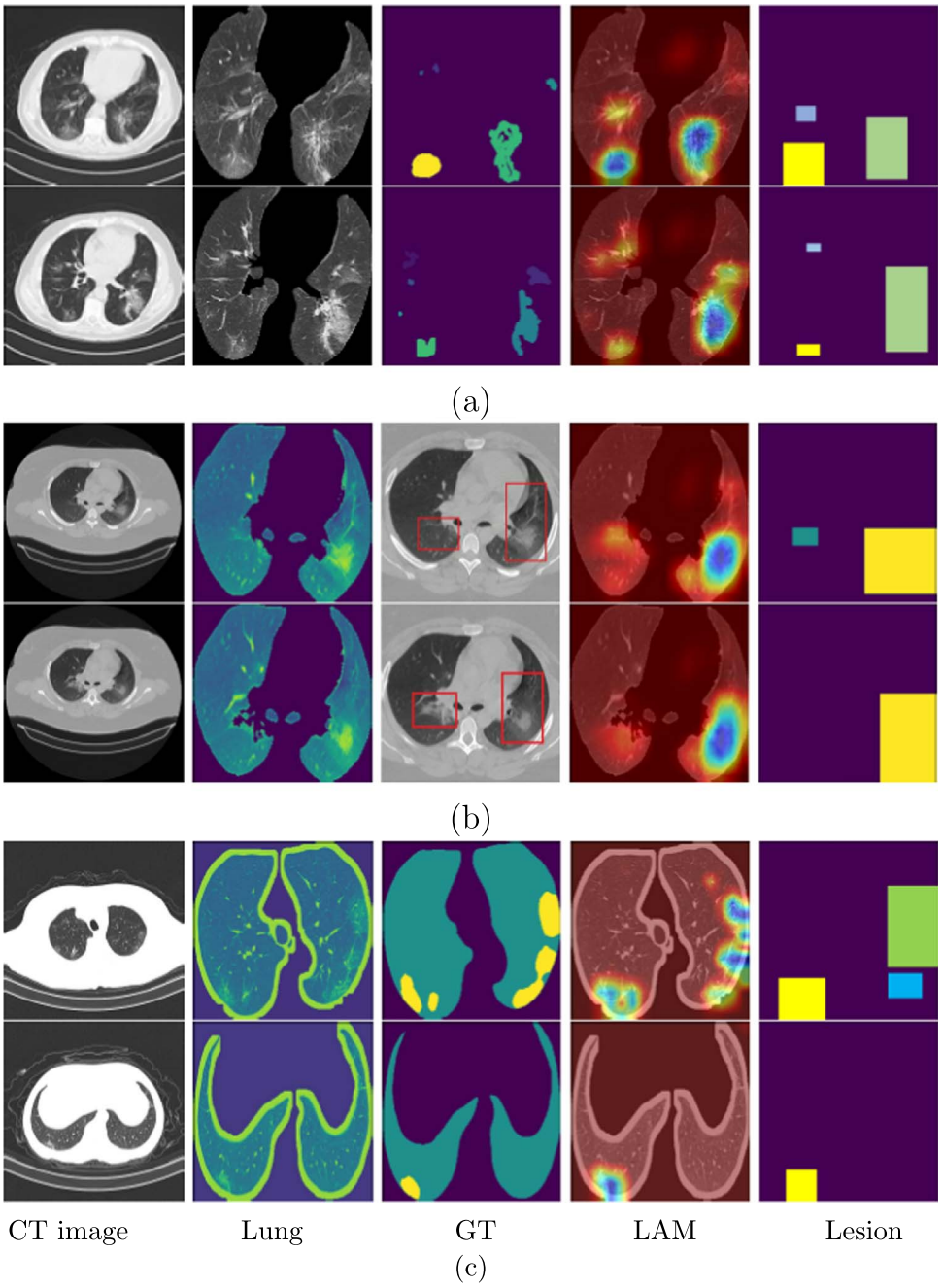

Figure 7. Example results of lesion detection for dataset (a) Radio-2, (b) Own and (c) CNCB.

First, the LAM does not always distinguish two lesions that are close together. As shown in the second case of figure 7(a) and the first case of figure 7(c), the two patches annotated as lesions are identified but presented as an integrated one. Though we set up strategies to keep the potential lesions separated as well as possible, the low activation around two patches will link them together if they are very close together. Thus, the effectiveness of the lesion detector will be affected. Second, the LAM is not very sensitive to small lesions. If we look at the ground-truth lesion maps, there are relatively small patches (compared to the image resolution) that are marked as lesions, as shown in the first (the green patches) and last case (the blue patches) of figure 7(a). When the candidate lesions are identified they fail to hit these small patches. Empirically, we have found the minimum threshold of the patch size to be 225. Hence, when the number of these patches compared to that of the candidates differ largely, it will result in a high number of false positives and the relatively low hit rate. Significantly, there also exist artefacts in the detected lesions that are probably caused by the heartbeat, breathing, or the diaphragm moving during scanning resulting in a certain proportion of pseudo lesions that also decrease the hit rate.

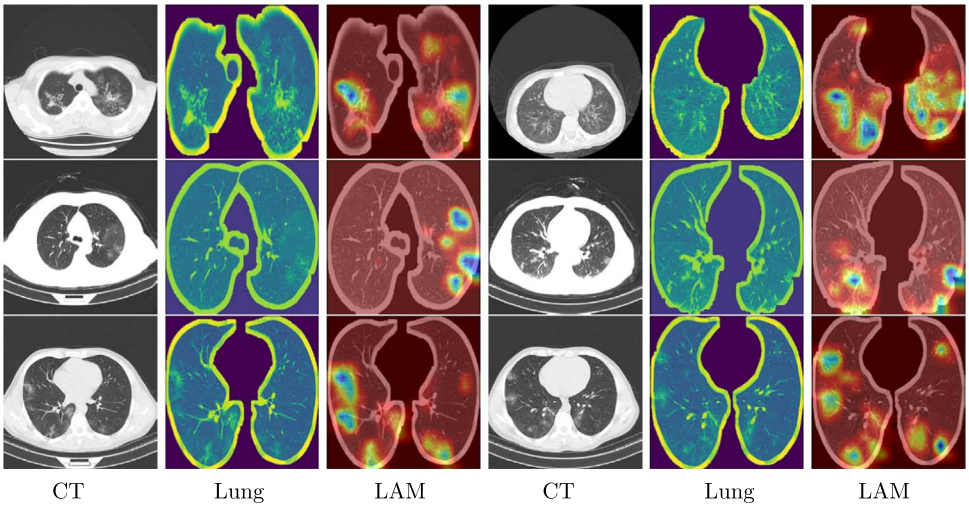

Figure 8. Lesion distribution maps for the COVIDx2a dataset, which corresponds to dataset TCIA, HUST and CTset from the top to the bottom, respectively.

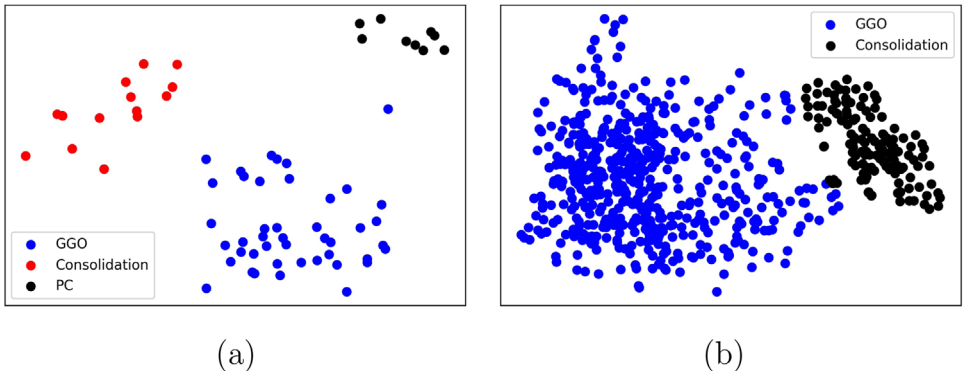

Figure 9. The encoded lesion feature space of dataset (a) Own and (b) CNCB. Each sample in the feature space represents a lesion.

5.3. Lesion identification

The hit rate indicates the ratio of correctly detected lesions, having detected the lesions we now verify the performance of the lesion identification using the extracted lesion features. Considering the our Own dataset, based on the radiologist guidelines we set $K = 3$, which results in three types of radiological manifestation for this data. The lesions are GGO, partial consolidation, and consolidation. The lesion is categorized as partial consolidation if the consolidation area of the lesion is over 10% but less than 80%. To visualize the cluster performance, we reduce the dimensions of the encoded lesion feature to 50 using truncated singular value decomposition and then use t-SNE (van der Maaten and Hinton 2008) to visualize these lesions in 2D space (see figure 9(a)).

The results of the lesion identification are listed in table 7. As can be seen, of the 77 detected lesions, 51 out of 54 GGO, 12 out of 14 consolidation, and 4 out of 9 partial consolidation are detected. The SENs of the three clusters are 94.44%, 44% and 85.71% respectively and the SPEs are 96.37%, 92.85% and 80.00%. Of the 9 lesions which are labelled as partial consolidation, 3 lesions are wrongly classified as GGO and 2 as consolidation. The identification rate of partial consolidation is not significant, which is to be expected as even an experienced radiologist cannot be 100% sure if the lesion belongs to partial consolidation or consolidation. The lesion identification is also evaluated on the CNCB dataset which has pixel-level lesion annotation. Since CNCB only has two types of annotation, $K = 2$ for this dataset. There we detect 862 lesions in total. Of the 612 recognized as GGO, 541 are annotated as GGO proactively, while of the 250 labelled consolidation, 199 are identified. Thus, the SENs of the two clusters are 88.39% and 76.53% respectively, and the SPEs are 72.70% and 92.39%. The encoded lesion feature space can be seen in figure 9(b). Figure 10 shows example identification results, including the input CT images, LAMs and the lesion clustering maps. The patches in a single slice are of the same type. The numbers on these patches denote the clustering index, and the positions of the numbers indicate the location of

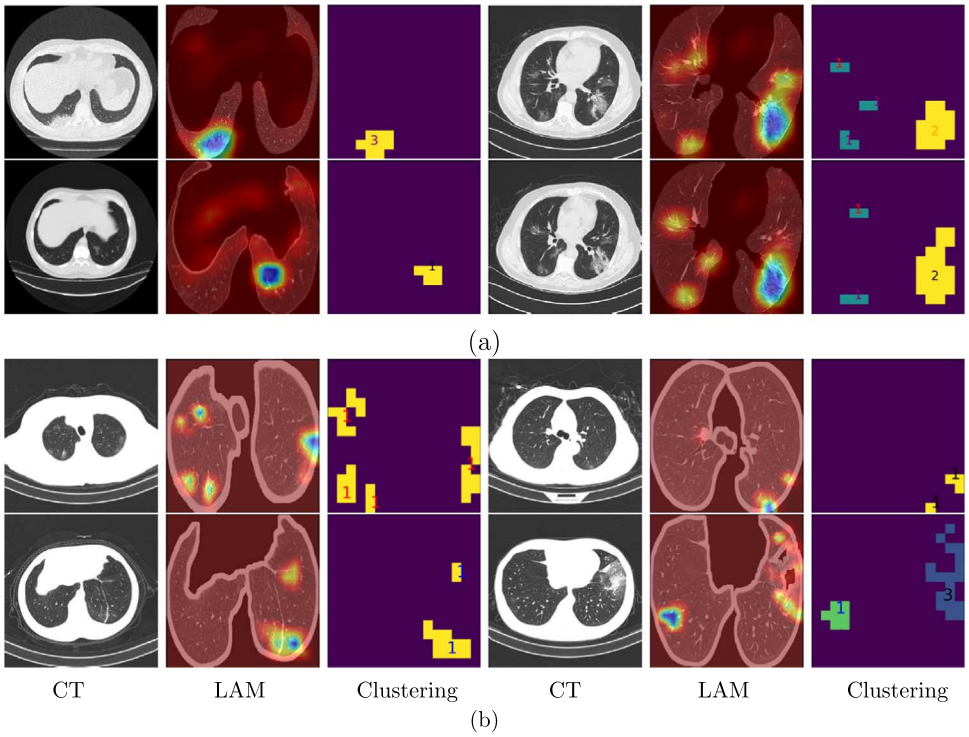

Figure 10. Lesion clustering maps for dataset (a) Own and (b) CNCB. 1 represents GGO, 2 partial consolidation and 3 consolidation.

Table 7. Lesion identification results for our Own dataset.

	Detected		
Ground truth	GGO	Partial consolidation	Consolidation
GGO	**51**	3	1
Partial consolidation	2	**4**	1
Consolidation	1	2	**12**
Sensitivity	**94.44%**	**44%**	**85.71%**
Specificity	**86.37%**	**92.85%**	**80.00%**

the local maxima. From these it is easy to observe that the *k*-means clustering algorithm can successfully use the encoded lesion features to distinguish the different lesions.

6. Discussion

This study has developed a CNN-based integrated framework for COVID-19 diagnosis and lesion analysis using CT scan data and weak annotations. The framework consists of two branches, the first, the diagnosis branch, learns the CT image representation for prediction of abnormal CT images. Simultaneously, this branch provides lesion information from abnormal CT images. Next the lesion identification branch learns the COVID-19 lesion representation capturing the cues with a multi-lesion detector for analysis. Our proposed framework shares several similarities with the work in Wang *et al* (2020d) for example the use of weak labels and the generation of binary masks of the lesion without pixel-level lesion supervision. However, our method has the following advantages:

(i) A feature enhanced network, FeCNN, which can achieve state-of-the-art diagnosis prediction with average precisions for the test datasets of 87%, 95% and 99% respectively.

(ii) A robust framework which has been demonstrated on datasets with a variety of scales, achieving areas under the ROC curves all over 89%.

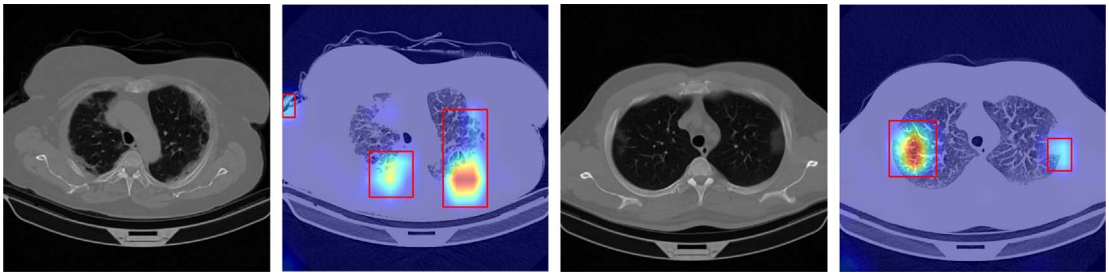

Figure 11. LAMs for dataset CTset of COVIDx2a without lung segmentation.

(iii) COVID-19 lesion classification without the need to explicitly consider the circumstances and unique attributes of the lesions. The lesion detector also provides a solution for multiple types of lesion occurring in the same CT image.

(iv) We have built a sizeable COVID-19 lesion feature space, which offers a new approach for the study of COVID-19 lesions.

This study was inspired by the fact that if the network can make an accurate prediction of COVID-19 then it must be capturing the differences caused by the lesions during the inference procedure. Hence, the capability of the lesion detector comes from extracting the lesion features from the FeCNN and we can use these hidden cues for lesion identification. However, this method has the following limitations:

(i) The detector is based on using a 2D slice, which means that it does not take into consideration the 3D distribution of lesions. Thus, there exists a semantic gap between mapping the lesion feature based on 2D slice to its clinical manifestation.

(ii) The detector can only encode the lesions with clear boundaries, in other words, if two different lesions are very close to each other the detector may consider these as one lesion.

In this paper although we did not focus on lesion segmentation specifically, from the LAM we can see that our FeCNN learns the imaging patterns of the COVID-19 lesions, and by combining the LAHm with a morphological operation the lesion detection rate we obtained was still relatively high. In the following we discuss the role of the lung segmentation, the diagnosis performance in a three-way classification task with other pneumonia CT images, the potential to identify different lesion types and future applications of our work.

6.1. Effectiveness of lung segmentation

In our work several pre-processing steps were carried out to allow us to evaluate the model on different data at the same scale and use only slices where the lungs are clearly shown. As part of the pre-processing lung segmentation is carried out. The segmentation procedure can potentially impact on the final performance of the lesion identification. There are existing works which can achieve state of the art performance in lung segmentation (Ronneberger *et al* 2015, Shan *et al* 2021), while in this study, we segment the lung from CT slice with the morphological operation. This is mainly because our method is weakly supervised, which means that the only available supervision is image-level (normal or abnormal). The morphological operation can segment the lung out without any prior label related to the lung area. To obtain quantitative results for the morphological operation, we tested this method on the dataset CNCB (Zhang *et al* 2020b), the result shows that we achieve Dice coefficient of 0.82 and therefore still have room for further improvement. This is to be expected because the morphological operation cannot segment the lung area from the severe fibrosis and effusions well, thus lowering the lung segmentation performance.

To further evaluate how segmentation affects the final performance, we test different configurations on the dataset CTset (Rahimzadeh *et al* 2021). The results show that the prediction performances do not differ much achieving 98.8% with lung segmentation and 98.1% without. However, without lung segmentation we cannot use the features from the network for lesion identification. From figure 11 we can see that the locations relating to the evidence the network uses to make its decision contain areas of air or fat. Therefore, even though the lung segmentation may not bring a significant improvement in the network prediction, its significance is that it makes the lesion recognition more interpretable.

Table 8. The results of weakly-supervised lesion detection by varying overlap thresholds.

	Type		
Threshold	COVID-19	Normal	Pneu
0.1	0.934	0.958	0.957
0.2	0.934	0.958	0.957
0.3	0.934	0.958	0.957
0.4	0.934	0.958	0.957
0.5	0.933	0.957	0.956
0.6	0.925	0.949	0.947
0.7	0.918	0.939	0.936
0.8	0.910	0.922	0.923
0.9	0.895	0.891	0.898

6.2. Three-way classification

The results in section 5 have shown the power of the proposed framework in the binary task. To further verify the effectiveness of this framework, we test the diagnosis branch in a three-way classification task. As the name suggests, the three-way classification task aims to distinguish COVID-19, normal and other pneumonia CT slices.

The basic experiment setup for the three-way classification is the same as described in section 4, while accordingly, the ρ in the fusion module and the $\mathcal{C}$ in softmax layer are set to 3. The output of the diagnosis branch is a 3-d tensor, each dimension of which indicates the probability that the CT image belongs to either COVID-19, normal or other pneumonia. We trained the adjusted method on the dataset COVIDx2a, this is a large-scale dataset that contains the three types CT slices (in total 194 922 CT images, 94 548 are COVID-19, 60 065 are normal, 58 321 belong to other pneumonia). The dataset was split following the public splitting file with 42 286 COVID-19, 25 496 normal, 35 996 pneumonia CT images for training.

Overall, our FeCNN achieves 0.95 accuracy in the three-way classification task. Although this is 0.04 less than the binary classification, it is still an acceptable result, especially considering the increase in complexity and uncertainty compared to the binary task. To verify the stability of the diagnosis prediction for the three types of CT images, we obtained a series of classification accuracies for each individual category by varying the probability threshold as shown in table 8. We can see that the classification accuracies for an individual type are higher than 0.9 when the threshold ranges from 0.2 to 0.8. And if we use the winner-take-all strategy, i.e. the threshold is 0.5, the COVID-19, normal and other pneumonia CT images can be recognized by accuracy of 93.3%, 95.7%, 95.6%, respectively. Even when we select the probability threshold to 0.9, the diagnosis accuracy is high with 89.5%, 89.1%, 89.8% for the three type, which shows the powerful and stable diagnosis capacity.

6.3. Potential of lesion identification

In the evaluation of the lesion identification, we set different K values in advance. This is due to the diversity of the data, e.g. guided by an experienced radiologist K is empirically set to a value of 3. However, different values of K may result in different lesion clusters. To explore how the value of K affects the lesion clusters, we use the dataset CTset (Rahimzadeh *et al* 2021), which has 436 detected lesions and test the k-means algorithm with different values of K on this lesion space.

For each value of K the inertias i.e. the sum of the distances of the samples to their closest cluster centre is calculated. Then we identify $K = 12$ as optimum using the elbow method. From the final result, we can find that even though there exist clusters with slight overlap, the result with 12 clusters appears reasonable. Each group has a clear boundary with the neighboring groups, suggesting that purely from the perspective of the lesions, there may be more than three types of lesion in the data space. And recently, several longitudinal researches (Ng *et al* 2020, Pan *et al* 2020), concluded that the lesions of COVID-19 would not stay in a dormant state. Hence, the lesion patterns could be diverse in intermediate states during the dynamic evolution of COVID-19. Therefore, it is reasonable to include the abstracted lesion feature into the clustering process, as this could help to discriminate the lesion patterns with tiny differences. More significantly, it implies that COVID-19 may exhibit other lesion types that have not been reported up to now. Equivalently, it could also be interpreted that the encoded lesion features are sensitive and the subtle discrimination of lesions is hard for humans to identify.

6.4. Future applications

In the clinical study of COVID-19 a significant amount of work has been put into analysis of CT scans to investigate the course and severity of the disease. By observing changes in the CT findings, a set of systematic rules can be developed to assess the severity of the COVID-19 patient. For instance, when enlarged regions of

GGO with superimposed inter- and intra-lobular septal thickening (crazy-paving pattern) are observed, the patient may be in a serious situation. Hence, with automatic recognition of the lesions from CT images, we could further investigate how the detected lesions can be mapped to the severity of COVID-19. Similarly, the lesion information can be provided to clinicians to assist them in making a diagnosis, enabling doctors to act early to changes in patients' condition early and enact treatment strategies.

Even though this paper has aimed to provide a solution for lesion analysis for COVID-19, the proposed approach is not COVID-19 specific. It is easy to see that our framework can potentially be generalized to lesion detection of other chest-based diseases. For other diseases where lesions can be observed with a CT scan, the proposed framework in this paper can naturally serve as a preliminary and cost-effective step to explore the clinical manifestation of this disease's lesions.

7. Conclusion

Identifying diverse types of lesions from CT images without pixel-wise labels is a challenging task. This paper presents a novel and effective integrated framework for this purpose. In particular, we leverage the power of neural networks to extract a deep representation of the CT images and bridge this representation to COVID-19 lesions through a multi-lesion detector. The obtained results prove that the diagnostic network can detect COVID-19 positive CT images and the lesion identification branch can successfully distinguish the lesion types. Furthermore, the proposed system has the capability to detect unreported lesions and hence can assist clinicians to assess the severity of the disease and to enact the treatment plan efficiently.

ORCID iDs

Kaichao Wu https://orcid.org/0000-0001-6988-624X
Beth Jelfs https://orcid.org/0000-0002-6844-7154
Xiangyuan Ma https://orcid.org/0000-0002-8107-6861
Qiang Fang https://orcid.org/0000-0003-3209-6417

References

An P *et al* 2020 CT images in COVID-19 [data set], the cancer imaging archive (https://doi.org/10.7937/TCIA.2020.GQRY-NC81)
Armato S G III *et al* 2015 Data from LIDC-IDRI, the cancer imaging archive (https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX)
Bell D and Hacking C 2020 Reference article Radiopaedia.org (https://doi.org/10.53347/rID-73913)
Cao Y, Xu Z, Feng J, Jin C, Han X, Wu H and Shi H 2020 Longitudinal assessment of COVID-19 using a deep learning-based quantitative CT pipeline: Illustration of two cases *Radiol.: Cardiothoracic Imaging* **2** e200082
Chan J F-W *et al* 2020 A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster *The Lancet* **395** 514–23
Chollet F 2017 Xception: deep learning with depthwise separable convolutions *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE: Honolulu, HI, USA) pp 1251–8
Clark K *et al* 2013 The cancer imaging archive (TCIA): maintaining and operating a public information repository *J. Digit. Imaging* **26** 1045–57
Duran-Lopez L, Dominguez-Morales J P, Corral-Jaime J, Vicente-Diaz S and Linares-Barranco A 2020 COVID-XNet: a custom deep learning system to diagnose and locate COVID-19 in chest x-ray images *Appl. Sci.* **10** 1-12
Ghoshal B and Tucker A 2020 Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection, arXiv:2003.10769
Gozes O, Frid-Adar M, Greenspan H, Browning P D, Zhang H, Ji W, Bernheim A and Siegel E 2020a Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis, arXiv:2003.05037
Gozes O, Frid-Adar M, Sagie N, Kabakovitch A, Amran D, Amer R and Greenspan H 2020b A weakly supervised deep learning framework for COVID-19 ct detection and analysis *Thoracic Image Analysis* (*Lecture Notes in Computer Science*) ed J Petersen *et al* vol 12502 (Cham: Springer) pp 84–93
Gulshan V *et al* 2016 Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs *JAMA* **316** 2402–10
Gunraj H, Wang L and Wong A 2020 COVIDNet-CT: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest CT images *Front. Med.* **7**
Harmon S A *et al* 2020 Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets *Nat. Commun.* **11** 1–7
He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE: Las Vegas, NV, USA) pp 770–8
Hirata Y *et al* 2020 Deep learning for screening of pulmonary hypertension using standard chest x-ray *Eur. Heart J.* **41** ehaa946.2246
Hu S *et al* 2020 Weakly supervised deep learning for COVID-19 infection detection and classification from CT images *IEEE Access* **8** 118869–83
Huang C *et al* 2020 Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China *The Lancet* **395** 497–506
Huang L, Han R, Ai T, Yu P, Kang H, Tao Q and Xia L 2020b Serial quantitative chest CT assessment of COVID-19: Deep-learning approach *Radiol.: Cardiothoracic Imaging* **2** e200075

Jin C *et al* 2020 Development and evaluation of an artificial intelligence system for covid-19 diagnosis *Nat. Commun.* **11** 1–14
Knipe H and Iqbal S 2020 *COVID-19 (summary)* Reference article, Radiopaedia.org (https://doi.org/10.53347/rID-75235)
Li L *et al* 2020 Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT *Radiol.* **296** E65–71
Lin T Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S, Belongie S and Belongie S 2017 Feature pyramid networks for object detection *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (Honolulu, HI, USA: IEEE) pp 2117–25
Lin Z, Sun J, Davis A and Snavely N 2020 Visual chirality *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (Seattle, WA, USA: IEEE) pp 12295–303
Ma J, Ge C, Wang Y, An X, Gao J and Yu Z 2020 COVID-19 CT lung and infectionsegmentation dataset (version 1.0) [data set], Zenodo (https://doi.org/10.5281/zenodo.3757476)
Mei X *et al* 2020 Artificial intelligence-enabled rapid diagnosis of patients with COVID-19 *Nature Med.* **26** 1–5
Meng L, Dong D, Li L, Niu M, Bai Y, Wang M, Qiu X, Zha Y and Tian J 2020 A deep learning prognosis model help alert for COVID-19 patients at high-risk of death: a multi-center study *IEEE J. Biomed. Health Inf.* **24** 3576–84
Narin A, Kaya C and Pamuk Z 2021 Automatic detection of coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks *Pattern Anal. Appl.* **24** 1207–20
Ng M *et al* 2020 Imaging profile of the COVID-19 infection: radiologic findings and literature review *Radiol. Cardiothoracic Imaging* **2** e200034
Ning W *et al* 2020 Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning *Nat. Biomed. Eng.* **4** 1197–207
Otsu N 1979 A threshold selection method from gray-level histograms *IEEE Trans. Syst., Man, Cybern.* **9** 62–6
Pan F *et al* 2020 Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia *Radiol* **295** 715–21
Panwar H, Gupta P, Siddiqui M K, Morales-Menendez R, Bhardwaj P and Singh V 2020 A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest x-ray and CT-Scan images *Chaos, Solitons Fractals* **140** 110190
Rahimzadeh M, Attar A and Sakhaei S M 2021 A fully automated deep learning-based network for detecting COVID-19 from a new and large lung CT scan dataset *Biomed. Signal Process. Control* **68** 102588
Rebuffi S-A, Fong R, Ji X, Bilen H and Vedaldi A 2019 NormGrad: finding the pixels that matter for training arXiv:1910.08823
Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation *MedicalImage Computing and Computer-Assisted Intervention* (*Lecture Notes in ComputerScience*) ed N Navab *et al* vol 9351 (Cham: Springer) pp 234–41
Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D 2017 Grad-cam:Visual explanations from deep networks via gradient-based localization *Proc. IEEE Int. Conf. on Computer Vision* (Venice, Italy: IEEE) pp 618–26
Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, Xue Z, Shen D and Shi Y 2021 Abnormal lung quantification in chest ct images of covid-19 patients with deep learning and its application to severity prediction *Med. Phys.* **48** 1633–45
Shi F, Wang J, Shi J, Wu Z, Wang Q, Tang Z, He K, Shi Y and Shen D 2020 Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19 *IEEE Rev. Biomed. Eng.* **14** 4–15
Shi F, Xia L, Shan F, Wu D, Wei Y, Yuan H, Jiang H, Gao Y, Sui H and Shen D 2021 Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification *Phys. Med. Biol.* **66** 065031
Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition arXiv:1409.1556
van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605 Available: http://jmlr.org/papers/v9/vandermaaten08a.html
Wang B *et al* 2020a AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system *Appl. Softw. Comput.* **98** 106897
Wang G, Liu X, Li C, Xu Z, Ruan J, Zhu H, Meng T, Li K, Huang N and Zhang S 2020b A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images *IEEE Trans. Med. Image* **39** 2653–63
Wang L, Lin Z Q and Wong A 2020c COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images *Sci. Rep.* **10** 1–12
Wang S *et al* 2021 A deep learning algorithm using CT images to screen for corona virus disease (COVID-19) *Eur. Radiol.* **31** 6096–104
Wang X, Deng X, Fu Q, Zhou Q, Feng J, Ma H, Liu W and Zheng C 2020d A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT *IEEE Trans. Med. Image* **39** 2615–25
World Health Organization 2021 Coronavirus disease (COVID-19) (https://who.int/emergencies/diseases/novel-coronavirus-2019/) (Accessed: 20 October 2021)
Ying S *et al* 2021 Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images *IEEE/ACM Trans.Comput. Biol. Bioinform.* **18** 2775-2780 Early Access
Yonekura A, Kawanaka H, Prasath V S, Aronow B J and Takase H 2017 Improving the generalization of disease stage classification with deep CNN for glioma histopathological images *Proc. IEEE Int. Conf. on Bioinformatics and Biomedicine* (Kansas City, MO, USA: IEEE) pp 1222–6
Zhang G, Huang X, Li S Z, Wang Y and Wu X 2004 Boosting local binary pattern (LBP)-based face recognition *Advances in Biometric Person Authentication* (*Lecture Notes in Computer Science*) vol 3338 (Berlin: Springer) pp 179–86
Zhang H *et al* 2020a Automated detection and quantification of COVID-19 pneumonia: CT imaging analysis by a deep learning-based software *Eur. J. Nucl. Med. Mol. Imaging* **47** 2525–32
Zhang K *et al* 2020b Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography *Cell* **181** 1423–33
Zhou B, Khosla A, Lapedriza A, Oliva A and Torralba A 2016 Learning deep features for discriminative localization *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (Las Vegas, NV, USA: IEEE) pp 2921–9
Zhu Y, Zhou Y, Ye Q, Qiu Q and Jiao J 2017 Soft proposal networks for weakly supervised object localization *Proc. IEEE Int. Conf. on Computer Vision* (Venice, Italy,: IEEE) pp 1841–50