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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
MIDCAN: A multiple input deep convolutional attention network for Covid-19 diagnosis based on chest CT and chest X-ray

Yu-Dong Zhang\textsuperscript{a,1}, Zheng Zhang\textsuperscript{b,c,1}, Xin Zhang\textsuperscript{d,e}, Shui-Hua Wang\textsuperscript{c,e}\textsuperscript{*}

\textsuperscript{a}School of Informatics, University of Leicester, Leicester, LE1 7RH, UK
\textsuperscript{b}Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen 518055, China
\textsuperscript{c}Department of Computer and Information Science, University of Macau, Macau 999078, China
\textsuperscript{d}Department of Medical Imaging, The Fourth People’s Hospital of Huai’an, Huai’an, Jiangsu Province, 223002, China
\textsuperscript{e}School of Mathematics and Actuarial Science, University of Leicester, LE1 7RH, UK

\section*{A R T I C L E   I N F O}

Article history:
Received 3 March 2021
Revised 14 May 2021
Accepted 23 June 2021
Available online 14 July 2021
Edited by: Maria De Marsico

Keywords:
Deep learning
Data harmonization
Multiple input
Convolutional neural network
Automatic differentiation
COVID-19
Chest CT
Chest X-ray
Multimodality

\section*{A B S T R A C T}

\textbf{Background:} COVID-19 has caused 3.34m deaths till 13/May/2021. It is now still causing confirmed cases and ongoing deaths every day.

\textbf{Method:} This study investigated whether fusing chest CT with chest X-ray can help improve the AI’s diagnosis performance. Data harmonization is employed to make a homogeneous dataset. We create an end-to-end multiple-input deep convolutional attention network (MIDCAN) by using the convolutional block attention module (CBAM). One input of our model receives 3D chest CT image, and other input receives 2D X-ray image. Besides, multiple-way data augmentation is used to generate fake data on training set. Grad-CAM is used to give explainable heatmap.

\textbf{Results:} The proposed MIDCAN achieves a sensitivity of 98.10±1.88%, a specificity of 97.95±2.26%, and an accuracy of 98.02±1.35%.

\textbf{Conclusion:} Our MIDCAN method provides better results than 8 state-of-the-art approaches. We demonstrate using multiple modalities can achieve better results than individual modality. Also, we demonstrate that CBAM can help improve the diagnosis performance.

© 2021 Elsevier B.V. All rights reserved.

\section{1. Introduction}

COVID-19 (also known as coronavirus) pandemic is an on-going infectious disease caused by severe acute respiratory syndrome (SARS) coronavirus 2 [1]. As of 13/May/2021, there are over 161,144 confirmed cases and over 3.34m deaths attributed to COVID-19. The cumulative deaths of the top 10 countries are shown in Fig. 1.

The main symptoms of COVID-19 are a low fever, a new and ongoing cough, a loss or change in taste and smell. In UK, the vaccines approved were developed by Pfizer/BioNTech, Oxford/AstraZeneca, and Moderna. The joint committee on vaccination and immunization (JCVI) [2] determines the order in which people will be offered the vaccine. At April/2021, people aged 50 and over, people of clinically (extremely) vulnerable, people living or working in the care homes, and health care providers, people with a learning disability are being offered.

Two COVID-19 diagnosis methods are available. The first method is viral testing to test the existence of viral RNA fragments [3]. The shortcomings of swab test [4] are two folds: (i) the swab samples may be contaminated and (ii) it needs to wait from several hours to several days to get the results. The other method is chest imaging. There are two main different chest imaging available: chest computed tomography (CCT) and chest X-ray (CXR).

CCT is one of the best chest imaging techniques so far, because it provides the finest resolution and it is capable of recognizing extremely small nodules [5]. It provides high-quality volumetric 3D chest data. On the other hand, CXR performs poor on soft tissue contrast, and it only provides 2D image [6].

In this paper, we aim to fuse CCT and CXR images, and expects the fusion can improve the performance compared to using CCT or CXR individually. Besides, we create a novel multiple input deep convolutional attention network (MIDCAN) that can handle CCT and CXR images simultaneously, and present the diagnosis

\textsuperscript{*} Corresponding authors.

E-mail addresses: yudongzhang@ieee.org (Y.-D. Zhang), 973306782@qq.com (X. Zhang), shuhuawang@ieee.org (S.-H. Wang).

\textsuperscript{1} Yu-Dong Zhang & Zheng Zhang contributed equally to this paper.
output. The contributions of this study are itemized briefly as following five points:

- Attention mechanism, convolutional block attention module, is included in the proposed MIDCAN model to improve the performance;
- The proposed MIDCAN model can handle CCT and CXR images simultaneously;
- Multiple-way data augmentation is employed to overcome overfitting problem;
- This proposed MIDCAN model gives more accurate performances than individual modality-based approaches;
- The proposed MIDCAN model is superior to state-of-the-art COVID-19 diagnosis approaches.

2. Literature survey

From previous year, AI field has investigated ongoing researches on automatic COVID-19 diagnosis, which can save the workloads of manual labelling.

For the CCT image based COVID-19 diagnosis, Chen (2020) [7] employed gray-level occurrence matrix (GLCM) as feature extraction method. The authors then used support vector machine (SVM) as the classifier. Yao (2020) [8] combined wavelet entropy (WE) and biogeography-based optimization (BBO). Wu (2020) [9] presented a novel method—wavelet Renyi entropy (WRE) to help diagnose COVID-19. El-kenawy, Ibrahim (2020) [10] proposed a feature selection voting classifier (FSVC) approach for COVID-19 classification. Satapathy (2021) [11] combined DenseNet with optimization of transfer learning setting (OTLS). Saood and Hatem (2021) [12] explored two structurally-different deep learning (DL) methods—U-Net and SegNet—for COVID-19 CT image segmentation.

On the other side, there are several successful AI models for XR image based COVID-19 diagnosis. For example, Ismael and Sengur (2020) [13] presented a multi-resolution analysis (MRA) approach. Loey, Smarandache (2020) [14] combined generative adversarial network (GAN) with GoogleNet. Their method is abbreviated as GG. Togacar, Ergen (2020) [15] employed social mimic optimization (SMO) for feature selection and combination. Das, Ghosh (2021) [16] used weighted average ensembling technique with convolutional neural network (CNN) for automatic COVID-19 detection.

The main achievements of above approaches are three points: (i) They only consider individual modality, either CCT or CXR. (ii) Their AI models are either combined feature extraction plus classifier model, or modern deep neural network models. Nevertheless, their models lack attention mechanism. (iii) Efficient measures to resist overfitting are missing.

Algorithm 1 Data harmonization.

| Step | Action |
|------|--------|
| 1    | For CCT image $A_i(n)$ and CXR image $B_i(n)$ of subject $n$. |
| 2    | For CXR image $B_i(n)$ Central rectangle region is reserved, and outskirt pixels are removed. |
| 3    | For CXR image $B_i(n)$ Central rectangle region is reserved, and outskirt pixels are removed. |
| Output | $A_i(n)$ and CXR image $B_i(n)$. |

To solve or alleviate above three shortcomings, we proposed the multiple input deep convolutional attention network. The dataset and details of our method will be discussed in Sections 3 and 4, respectively.

3. Dataset

3.1. Data harmonization

This retrospective study was granted to exempt ethical approval. 42 COVID-19 patients and 44 healthy controls (HCS) were recruited. All the data were collected from local hospitals. Each subject $n$ takes a CCT scan a CXR scan, and thus generate a CCT image $A_i(n)$ and CXR image $B_i(n)$. Due to different chest sizes of different people and the different sources of scanning machines, the height of image $A_i(n)$ and the size of $B_i(n)$ vary.

To make a homogeneous dataset, data harmonization [17] is used. The central 64 slices of CCT image and the central rectangle region of CXR image are reserved. The height and width of CCT slices are resized to 1024 $\times$ 1024, and the CXR image is resized to 2048 $\times$ 2048. They are named $A_i(n)$ and $B_i(n)$. We choose 64 and 2048 because we find this can keep the lung part of images while get rid of unrelated body tissues. The details are displayed in Algorithm 1.

3.2. Data Preprocessing

Second, data preprocessing (See Fig. 2) is used since both CCT and XR image contain redundant/unrelated spatial information and their sizes are still too large. First, all the CCT and CXR images are grayscaled. Second, histogram stretch is carried to enhance the image contrast, where $(v_{min}, v_{max})$ stand for the minimum and maximum grayscale values of our images. Third, the margins at four directions are cropped (e.g., the text in the right side and the check-up bed in bottom side of CCT images, the neck part in the top side of CXR images, the background regions at four directions, etc.). Finally, CCT images are resized to $H_{CT} \times W_{CT} \times C_{CT}$ and CXR images are resized to $H_{XRX} \times W_{XRX} \times 1$.

Fig. 3 gives the examples of preprocessed images of a COVID-19 patient. Fig. 3(a) displays one slice out of 16 CCT slices, and Fig. 3(b) displays the CXR image.

4. Methodology

4.1. Convolutional block attention module

Table 1 gives the abbreviation list. DL has gained many successes in prediction/classification quests. Among all the DL structures, convolutional neural network (CNN) [18, 19] is particularly suitable for analyzing 2D/3D images. To help boost the performance of CNN, researches are proposed to modify CNN structures in terms of either depth, or cardinality, or width. Newly, scholars have studied on attention mechanism, and attempted to integrate attention to DL structures. For example, Hu, Shen (2020)
[20] proposed squeeze-and-excitation (SE) network. Woo, Park (2018) [21] presented a new convolutional block attention module (CBAM), that improves the traditional convolutional block (CB) by integrating attention mechanism. This study we choose CBAM because CBAM can provider both spatial attention and channel attention, compared to SE.

Take a 2D-image input as an example, Fig. 4(a) displays the structure of a traditional CB. The output of previous block was sent to m-repetitions of convolution layer, batch normalization (BN), and rectified linear unit (ReLU) layer. Finally, the m-repetitions is followed by a pooling layer. The output is named activation map (AM), symbolized as $G \in \mathbb{R}^{C \times H \times W}$, where $(C, H, W)$ stands for the sizes of channel, height, and width, respectively.

In contrast to Fig. 4(a), Fig. 4(b) adds the structure of CBAM, by which two modules: channel attention module (CAM) and spatial attention module (SAM) are added to refine the activation map $G$. Suppose the CBAM applies a 1D CAM $N_{CAM} \in \mathbb{R}^{1 \times 1}$ and a 2D SAM $N_{SAM} \in \mathbb{R}^{1 \times H \times W}$ in sequence to the input $G$. Hence, the channel-refined activation map can be obtained as:

$$H = N_{CAM}(D) \otimes G$$

(1)

And the final refined AM

$$I = N_{SAM}(E) \otimes H$$

(2)

where $\otimes$ means the element-wise multiplication. $I$ is the refined AM, which replaces the $G$ of traditional CB output, and it will be sent to the next block.

Note if the above two operands are not with the same dimension, the values are reproduced so that (i) the spatial attentional values are copied along the channel dimension, and (ii) the channel attentional values are copied along the spatial dimension.

4.2. Channel Attention Module

CAM is firstly defined. Both max pooling (MP) $z_{mp}$ and average pooling (AP) $z_{mp}$ are employed, making two features $J_{ap}$ and $J_{mp}$ as shown in Fig. 5(a).

$$\begin{align*}
J_{ap} &= z_{ap}(G) \\
J_{mp} &= z_{mp}(G)
\end{align*}$$

(3)

Both $J_{ap}$ and $J_{mp}$ are thenceforth sent to a shared multi-layer perceptron (MLP) to make the output AMs, that are then merged.
via element-wise summation $\oplus$. The merged sum $z_{mlp}[J_{ap}] \oplus z_{mlp}[J_{mp}]$ is lastly forwarded to $\sigma$. That is,

$$N_{\text{CAM}}(G) = \sigma \left( z_{mlp}[J_{ap}] \oplus z_{mlp}[J_{mp}] \right)$$  \hspace{1cm} (4)

where $\sigma$ is sigmoid function.

To decrease the parameter space, the hidden size of MLP is fixed to $\mathbb{R}^{C/r \times 1 \times 1}$ where $r$ stands for the reduction ratio. Assume $X_0 \in \mathbb{R}^{C \times r \times C}$ and $X_1 \in \mathbb{R}^{C \times C/r \times 1}$ mean the MLP weights (See Fig. 5a), respectively, equation (4) can be rewritten as:

$$N_{\text{CAM}}(G) = \sigma \left( [X_1]_{X_0[J_{ap}]} \oplus [X_1]_{X_0[J_{mp}]} \right)$$  \hspace{1cm} (5)

Note $X_0$ and $X_1$ are shared by both $J_{ap}$ and $J_{mp}$. Fig. 5(a) displays the diagram of CAM.

### 4.3. Spatial Attention Module

Next, SAM is defined as in Fig. 5(b). The spatial attention module $N_{\text{SAM}}$ is a complementary procedure to the previous CAM $N_{\text{CAM}}$. The average pooling $z_{ap}$ and max pooling $z_{mp}$ are harnessed again to the channel-refined activation map $H$,

$$\begin{cases} K_{ap} = z_{ap}(H) \\ K_{mp} = z_{mp}(H) \end{cases}$$  \hspace{1cm} (6)

Both $K_{ap}$ and $K_{mp}$ are two dimensional AMs: $K_{ap} \in \mathbb{R}^{1 \times H \times W}$ and $K_{mp} \in \mathbb{R}^{1 \times H \times W}$. They are concatenated via concatenation function $z_{con}$ together along the channel dimension as

$$K = z_{con}(K_{ap}, K_{mp})$$  \hspace{1cm} (7)

Afterwards, the concatenated AM is passed into a standard convolution $z_{con}$ with the size of $7 \times 7$, followed by sigmoid function $\sigma$. Overall, we attain:

$$N_{\text{SAM}}(H) = \sigma \left( z_{con} \left( [K] \right) \right)$$  \hspace{1cm} (8)

The $N_{\text{SAM}}(H)$ is then element-wisely multiplied by $H$ to get the final refined AE $I$. See Equation (2). The diagram of SAM is portrayed in Fig. 5(b).

### 4.4. Single Input and Multiple Input Deep Convolutional Attention Networks

In this study, we proposed a novel multiple-input deep convolutional attention network (MIDCAN) based on the ideas of CBAM and multiple-input. The structure of this proposed MIDCAN is determined by trial-and-error method. The variable $m$ in each varies, and we found the best values are chosen in the range from $[1, 3]$. We tested values larger than 3, which increase the computation burden, but the performances do not increase.

The structure of proposed shown below in Fig. 6(a), which is composed of two inputs. The left input is “Input-CCT” where CCT images are passed into the network. The right input is “Input-CXR” where CXR images are passed into the network. Suppose $N_{\text{CCT}}$ and $N_{\text{CXR}}$ stand for the number of CBAM blocks individually. We set $N_{\text{CCT}} = N_{\text{CXR}} = 5$ in this study by trial-and-error.

For the left branch, the CCT input $A_1$ goes through $N_{\text{CCT}}$ 3D-CBAMs and generates the output AM $A_2$, which is then flattened into $A_3$. Similarly, the CXR input at right branch $B_1$ goes through $N_{\text{CXR}}$ 2D-CBAMs, generates the output AM $B_2$, which is flattened into $B_3$. The deep CCT features and deep CXR features are then concatenated via concatenation function $z_{con}$ as

$$\text{MIDCAN} : F_1 = z_{con}(A_3, B_3)$$  \hspace{1cm} (9)

Here note than in our experiments, we use ablation studies, where we set two models: single-input deep convolutional atten-
tion network (SIDCAN) models, which remove the left and right branches, respectively. The first SIDCAN model, shown in Fig. 6(b), will only use CCT features, i.e.,

**SIDCAN – CCT :** $F_1 = A_B$  

This model is given a short name as **SIDCAN-CCT**.

The second SIDCAN model will only use CXR features, i.e.,

**SIDCAN – CXR :** $F_1 = B_B$  

This model is named as **SIDCAN-CXR**. Its flowchart is displayed in Fig. 6(c). Those two models will be used as comparison method in our experiments.

The feature $F_2$ is then passed to two fully-connected layers [22]. The first FCL contains 500 neurons, and the last FCL contains $N_C$ neurons, where $N_C$ stands for the number of classes. In this study $N_C = 2$. Finally, a softmax layer [23] turns the $F_2$ into probability. The loss function of this MIDCAN is cross entropy [24] function.

Table 2 gives the details of proposed MIDCAN. For the kernel parameter in Table 2, $[3 \times 3 \times 3, 16] \times 3$, $[2]/[2]/2$ stands for 3 repetitions of 16 filters with each size of $3 \times 3 \times 3$, following by a pooling with pooling factor of 2, 2, and 2 along three dimensions, respectively. In FCL stage, the kernel parameter gives the size of weight matrix and bias vector, respectively.

### 4.5. 18-way data augmentation

Data augmentation (DA) [25] is an important utensil over the training set to avoid overfitting of classifiers when applied to test set. Meanwhile, DA can overcome the small-data-set problem. Recently, Wang (2021) [26] proposed a novel 14-way data augmentation (DA), which used seven different DA techniques to the preprocessed image $v(k)$ and its horizontal mirrored image $v^H(k)$, respectively. Cheng (2021) [27] presented a 16-way DA, and used PathShuffle technique to avoid overfitting.

This study enhances the 14-way DA method [26] to 18-way DA, by adding two new DA methods: salt-and-pepper noise (SAPN) and speckle noise (SN) on both $v(k)$ and $v^H(k)$. Use $v(k)$ as an example, the SAPN altered image is symbolized as $A_{\text{SAPN}}(k)$ with its values are set as

$$
\begin{align*}
\text{R}(\rho_{\text{SAPN}} = I) &= 1 - \gamma_{d}^{\text{SAPN}}, \\
\text{R}(\rho_{\text{SAPN}} = I_{\text{min}}) &= \frac{\gamma_d^{\text{SAPN}}}{2}, \\
\text{R}(\rho_{\text{SAPN}} = I_{\text{max}}) &= \frac{1}{\gamma_d^{\text{SAPN}}},
\end{align*}
$$

where $\gamma_d^{\text{SAPN}}$ stands for noise density, and $\text{R}$ the probability function. $I_{\text{min}}$ and $I_{\text{max}}$ correspond to black and white colors, respectively.

On the other side, the SN altered image is defined as

$$v^{\text{SN}}(k) = v(k) + \gamma \times v(k),$$

(13)

where $\gamma$ is a uniformly distributed random noise, of which the mean and variance are symbolized as $U_m^{\text{SN}}$ and $U_v^{\text{SN}}$, respectively. Take Fig. 3(b) as the example, Fig. 7(a-b) display the SAPN and SN altered images, respectively. Due to the page limit, the results of other DA are not shown in this paper.

Let $M_d$ stands for the number of DA techniques to the preprocessed image $v(k)$, and $M_b$ stands for the number of new generated images for each DA. This proposed $(2 \times M_d)$-way DA algorithm is a four-step algorithm depicted below:

First, $M_b$ geometric/photometric/noise-injection DA transforms are utilized on preprocessed train image $v(k)$. We use $w_m^{\text{DA}}$ index to denote each DA operation. See each DA operation $w_m^{\text{DA}}$ yields $M_b$ new images. Thus, for a given image $v(k)$, we yield $M_b$ different data set $w_m^{\text{DA}}(v(k))$, $m = 1, \ldots, M_b$, and each dataset contains $M_b$ new images.

Second, horizontally mirrored image is generated as

$$v^H(k) = z_{\text{H}}[v(k)],$$

(14)

where $z_{\text{H}}$ means horizontal mirror function. In this study, all the $M_b$ DA methods are carried out on the mirror image $v^H(k)$, and generate $M_b$ different dataset $w_m^{\text{DA}}(v^H(k))$, $m = 1, \ldots, M_b$.

Four, the raw image $v(k)$, the horizontally mirrored image $v^H(k)$, all $M_b$-way results of preprocessed image $w_m^{\text{DA}}(v(k))$, $m = 1, \ldots, M_b$, and $M_b$-way DA results of horizontally mirrored image $w_m^{\text{DA}}(v^H(k))$, $m = 1, \ldots, M_b$, are fused together using concatenation function $z_{\text{CON}}$, as defined in Eq. (9).

The final combined dataset is defined as

$$|\Lambda| = 2 \times M_d \times M_b + 2$$

(16)

images (including the original image $v(k)$). Note in our dataset, different $M_b$ will be assigned to CCT training images and CXR images since CCT images are 3D and CXR images are 2D. That means for each DA, we have $M_b^{\text{CCT}}$ new images for each CCT image and $M_b^{\text{CXR}}$ new images for each CXR image.

### 4.6. Implementation and evaluation

K-fold cross validation is employed on both datasets. Suppose confusion matrix $J$ over $r$-th $(1 \leq r \leq R)$ run and $k$-th $(1 \leq k \leq K)$
fold is defined as
\[ J(r, k) = \begin{bmatrix} j_{11}(r, k) & j_{12}(r, k) \\ j_{21}(r, k) & j_{22}(r, k) \end{bmatrix} \]  
where \((j_{11}, j_{12}, j_{21}, j_{22})\) stand for TP, FN, FP, and TN, respectively. 
P stands for positive class, i.e., COVID-19, and N means negative class, i.e., healthy control. \(k\) represents the index of trial/fold, and \(r\) stands for the index of run. At \(k\)-th trial, the \(k\)-th fold is used as test, and all the left folds are used as training.

Note that \(J(r, k)\) is calculated based on each test fold, and are then summarized across all \(K\) trials, as shown in Fig. 8. Afterwards, we get the confusion matrix at \(r\)-th run \(J(r)\) as
\[ J(r) = \sum_{k=1}^{K} J(r, k) \]  
Seven indicators \(\bar{\eta}(r)\) are computed based on the confusion matrix over \(r\)-th run \(J(r)\).
\[ \bar{\eta}(r) = [\eta_1(r), \eta_2(r), \ldots, \eta_7(r)] \]
where first four indicators are: \(\eta_1\) sensitivity, \(\eta_2\) specificity, \(\eta_3\) precision, and \(\eta_4\) accuracy. Those four indicators are commonly used. Their definitions can be found easily. \(\eta_5\) is F1 score.
\[ \eta_5(r) = \frac{2 \times j_{11}(r)}{2 \times j_{11}(r) + j_{12}(r) + j_{21}(r)} \]  
\(\eta_6\) is Matthews correlation coefficient (MCC)
\[ \eta_6(r) = \frac{j_{11}(r) \times j_{22}(r) - j_{12}(r) \times j_{21}(r)}{\sqrt{(j_{11}(r) + j_{12}(r)) \times (j_{11}(r) + j_{21}(r)) \times (j_{22}(r) + j_{21}(r)) \times (j_{22}(r) + j_{12}(r))}} \]
and \(\eta_7\) is the Fowlkes–Mallows index (FMI),
\[ \eta_7(r) = \sqrt{\frac{j_{11}(r)}{j_{11}(r) + j_{12}(r) + j_{21}(r) + j_{22}(r)}} \]  
There are two indicators \(\eta_4\) and \(\eta_6\) using all the four basic measures \((j_{11}, j_{12}, j_{21}, j_{22})\). Considering the range of \(\eta_4\) is \(0 \leq \eta_4 \leq 1\), and the range of \(\eta_6\) is \(-1 \leq \eta_6 \leq 1\), we finally choose \(\eta_6\) as the most important indicator. Besides, Chicco, Totsch (2021) [28] stated that MCC is more reliable than many other indicators.

Above procedure is one run of K-fold cross validation. We run the K-fold cross validation \(R\) runs. The mean and standard deviation (MSD) of all seven indicators \(\eta_m(m = 1, \ldots, 7)\) are calculated over all \(R\) runs.
\[ \mu(\eta_m) = \frac{1}{R} \sum_{r=1}^{R} \eta_m(r) \]
\[ \sigma(\eta_m) = \sqrt{\frac{1}{R-1} \sum_{r=1}^{R} [\eta_m(r) - \mu(\eta_m)]^2} \]  
where \(\mu\) stands for the mean value, and \(\sigma\) stands for the standard deviation. The MSDs are reported in the format of \(\eta = \mu(\eta) \pm \sigma(\eta)\).

5. Experiments, results, and discussions

5.1. Parameter setting

Table 3 itemizes the parameter setting. Here the minimum value and maximum value of our images are set to 0 and 255, respectively. The size of preprocessed CCT images and CXR images are set to \(256 \times 256 \times 16\) and \(256 \times 256\), respectively. The number of CBAM blocks for CCT and CXR branches are set to 5. The noise density of SAPN is set to 0.05. The mean and variance of uniform distributed noise in SN are set to 0 and 0.05, respectively.

Nine different DA methods are used, so we have an 18-way DA if we consider both raw training image and its horizontal mirrored image. For each DA, 30 new images are generated for each CXR image, and 90 new images are generated for each CCT image. The number of K-fold is set to \(K = 10\). We run our model \(R = 10\) runs.

5.2. Statistics of proposed MIDCAN

We use two modalities, CCT and CXR, in this experiment. The structure of our model is shown in Fig. 6(a). The statistical results of proposed MIDCAN are shown in Table 4. As it shows, the sensitivity, specificity, precision and accuracy are 98.10±1.88, 97.95±2.26, 97.92±2.24, and 98.02±1.35, respectively. Moreover, the F1 score is 97.93±1.37, the MCC is 96.09±2.68, and FMI is 97.99±1.36.

5.3. Effect of multimodality and attention mechanism

We compare multiple-modality against single-modality. Two models, viz., SIDCAN-CCT and SIDCAN-CXR, shown in Fig. 6(b-c)
are used. Meanwhile, using attention and not using attention are compared.

The comparison results are shown in Table 5, where NA means no attention. Fig. 9 presents the error bar comparison of all the six setting. Comparing using attention against without using attention, we can observe the attention mechanism does help improve the classification performance, which is coherent with the conclusion of Ref. [21].

Meanwhile, if we compare MDCAN with two SIDCAN models, we can conclude that multimodality has the better performance than single modalities (both CT and CXR).

Fig. 10 displays the ROC curves of six settings. The blue patch corresponds to the lower bound and upper bound. For the first three models using attention, we can observe their AUCs are 0.9855, 0.9695, and 0.9567 for MDCAN, SIDCAN-CCT, and SIDCAN-CXR, respectively. If removing the CBAM module, we can observe from the bottom part of Fig. 10, that the corresponding AUCs decrease to 0.9512, 0.9361, and 0.9262, respectively. In addition, multimodality is proven to give better performance than using single-modality.

5.4. Explainability of proposed model

Fig. 11 presents the manual delineation and heatmap results of one patient. The heatmap images are generated via Grad-CAM method [29]. From Fig. 11, we can observe the proposed MDCAN model is able to capture the lesions of both CT image and CXR image accurately. This explainability via Grad-CAM can help the doctors, radiologists, and patients to better understand how our AI model works.
Table 5
Comparison of different settings.

| Method         | $\eta_1$ | $\eta_2$ | $\eta_3$ | $\eta_4$ | $\eta_5$ | $\eta_6$ | $\eta_7$ |
|----------------|----------|----------|----------|----------|----------|----------|----------|
| MIDCAN         | 98.10    | 97.95    | 97.92    | 98.02    | 97.98    | 96.09    | 97.99    |
|                | ±1.88    | ±2.26    | ±2.24    | ±1.35    | ±1.37    | ±2.66    | ±1.36    |
| SIDCAN         | 96.19    | 95.91    | 95.76    | 96.05    | 95.96    | 92.11    | 95.97    |
| -CCT           | ±1.23    | ±1.44    | ±1.40    | ±0.60    | ±0.60    | ±1.20    | ±0.60    |
| SIDCAN-CXR     | 93.81    | 93.86    | 93.70    | 93.84    | 93.69    | 87.78    | 93.72    |
| ±3.01          | ±2.64    | ±2.48    | ±0.78    | ±0.85    | ±1.56    | ±0.84    |
| MIDCAN         | 94.29    | 94.32    | 94.10    | 94.30    | 94.17    | 88.65    | 94.18    |
| (NA)           | ±2.30    | ±1.61    | ±1.45    | ±0.86    | ±0.94    | ±1.69    | ±0.93    |
| SIDCAN-CCT(NA) | 92.86    | 93.64    | 93.36    | 93.26    | 93.08    | 86.55    | 93.10    |
| ±1.59          | ±2.09    | ±2.02    | ±0.74    | ±0.71    | ±1.49    | ±0.72    |
| SIDCAN-CXR(NA) | 89.52    | 90.68    | 90.29    | 90.12    | 89.85    | 80.31    | 89.88    |
| ±2.30          | ±2.92    | ±2.63    | ±0.82    | ±0.76    | ±1.64    | ±0.76    |

Table 6
Comparison with SOTA approaches (Unit: %).

| Approach              | $\eta_1$ | $\eta_2$ | $\eta_3$ | $\eta_4$ | $\eta_5$ | $\eta_6$ | $\eta_7$ |
|-----------------------|----------|----------|----------|----------|----------|----------|----------|
| GLCM [7]              | 71.90    | 78.18    | 76.04    | 75.12    | 73.80    | 50.35    | 73.89    |
| ±4.02                 | ±3.89    | ±2.41    | ±0.98    | ±1.49    | ±1.91    | ±1.19    |
| WE-BBO [8]            | 74.05    | 74.77    | 73.83    | 74.42    | 73.81    | 48.98    | 73.88    |
| ±4.82                 | ±3.93    | ±1.84    | ±0.78    | ±1.65    | ±1.65    | ±1.67    |
| WRE [9]               | 86.43    | 86.36    | 86.01    | 86.40    | 86.12    | 72.95    | 86.17    |
| ±3.18                 | ±3.86    | ±3.13    | ±0.56    | ±0.39    | ±0.40    |
| FSVC [10]             | 91.90    | 90.00    | 89.85    | 90.93    | 90.82    | 81.97    | 90.85    |
| ±2.56                 | ±2.44    | ±1.99    | ±0.49    | ±0.55    | ±0.99    | ±0.56    |
| OTLS [11]             | 95.95    | 96.59    | 96.45    | 96.28    | 96.17    | 92.60    | 96.19    |
| ±2.26                 | ±1.61    | ±1.56    | ±1.07    | ±1.13    | ±2.09    | ±1.12    |
| MRA [13]              | 86.43    | 90.45    | 89.71    | 88.49    | 87.98    | 77.09    | 88.02    |
| ±3.00                 | ±2.79    | ±2.63    | ±2.08    | ±2.27    | ±4.17    | ±2.26    |
| GG [14]               | 93.33    | 90.00    | 90.13    | 91.63    | 91.61    | 83.49    | 91.67    |
| ±2.70                 | ±4.44    | ±3.81    | ±1.53    | ±1.35    | ±2.84    | ±1.30    |
| SMO [15]              | 93.10    | 95.23    | 94.99    | 94.19    | 93.99    | 88.45    | 94.02    |
| ±2.37                 | ±2.50    | ±2.45    | ±1.10    | ±1.13    | ±2.16    | ±1.11    |
| MIDCAN (Ours)         | 98.10    | 97.95    | 97.92    | 98.02    | 97.98    | 96.10    | 97.99    |
| ±1.88                 | ±2.26    | ±2.24    | ±1.35    | ±1.37    | ±2.66    | ±1.36    |

Table 6
Comparison with SOTA approaches (Unit: %).

Fig. 12. 3D bar plot of approach comparison.

5.5. Comparison to State-of-the-art approaches

We compare the proposed method “MIDCAN” with 8 state-of-the-art methods: GLCM [7], WE-BBO [8], WRE [9], FSVC [10], OTLS [11], MRA [13], GG [14], SMO [15]. Those methods were carried out on single modality dataset depending on their original paper reported (either CCT or CXR), so we test those methods in the corresponding SIDCAN models and single-modality dataset.

All the methods were evaluated via 10 runs of 10-fold cross validation. The MSD results $\eta$ of all approaches on ten runs of 10-fold cross validation are pictured in Fig. 12, which sorts all the methods in terms of $\eta_6$, and itemized in Table 6.

From Table 6, we can observe that this proposed MIDCAN outperforms all the other 8 comparison baseline methods in terms of all indicators.

The reason why our MIDCAN method is the best lie in following three facts: (i) we propose to use multiple modality instead of traditional single modality; (ii) CBAM is used in our network that attention mechanism can help our AI model focuses on the lesion region; (iii) multiple-way data augmentation is employed to overcome overfitting.

6. Conclusion

This paper proposed a novel multiple input deep convolutional attention network (MIDCAN) model for diagnosis of COVID-19. The results show our method achieves a sensitivity of 98.10±1.88%, a specificity of 97.95±2.26%, and an accuracy of 98.02±1.35%.

In the future researches, we shall carry out several attempts: (i) expand our dataset; (ii) include other advanced network strategies, such as graph neural network; (iii) collect IoT signals of subjects.

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.
Acknowledgment

This paper is partially supported by Medical Research Council Confidence in Concept Award, UK (MC_PC_17171); Hope Foundation for Cancer Research, UK (RM60GO680); British Heart Foundation Accelerator Award, UK; Sino-UK Industrial Fund, UK (RP202G0289); Global Challenges Research Fund (GCRF), UK (P202PF11); Royal Society International Exchanges Cost Share Award, UK (RP202G0230).

References

[1] K. Turgutlu, et al., Determinants of mortality in a large group of hemodialysis patients hospitalized for COVID-19, BMC Nephrol. 22 (1) (2021) 10 Article ID. 29.
[2] A.J. Hall, The United Kingdom joint committee on vaccination and immunisation, Vaccine 28 (2010) AS4–A57.
[3] D. Sakashita, et al., Comparative evaluation of nasopharyngeal swab and saliva specimens for the molecular detection of SARS-CoV-2 RNA in Japanese patients with COVID-19, J. Infect. Chemother. 27 (1) (2021) 126–129.
[4] C. Giannitro, et al., Chest CT in patients with a moderate or high pretest probability of COVID-19 and negative swab, Radiol. Med. (Torino) 125 (12) (2020) 1260–1270.
[5] R.L. Draelos, et al., Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes, Med. Image Anal. 67 (2021) 12 Article ID. 101857.
[6] A. Braga, et al., When less is more: regarding the use of chest X-ray instead of computed tomography in screening for pulmonary metastasis in post-molar gestational trophoblastic neoplasia, Br. J. Cancer (2021), doi:10.1038/s41416-020-01209-5.
[7] Y. Chen, Covid-19 classification based on gray-level co-occurrence matrix and support vector machine, in: COVID-19: Prediction, Decision-Making, and its Impacts, Springer Singapore, Singapore, 2020, pp. 47–55. K.C. Santosh and A. Joshi, Editors.
[8] X. Yao, COVID-19 detection via wavelet entropy and biogeography-based optimization, in: in COVID-19: Prediction, Decision-Making, and its Impacts, Springer Singapore, Singapore, 2020, pp. 69–76. K.C. Santosh and A. Joshi, Editors.
[9] X. Wu, Diagnosis of COVID-19 by wavelet Renyi entropy and three-segment biogeography-based optimization, Int. J. Comput. Intell. Syst. 13 (1) (2020) 1332–1344.
[10] E.S.M. El-Kenawy, et al., Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images, IEEE Access 8 (2020) 179317–179335.
[11] S.C. Satapathy, Covid-19 diagnosis via DenseNet and optimization of transfer learning setting, Cognitive Comput. (2021), doi:10.1007/s12559-020-09776-8.
[12] A. Saood, et al., COVID-19 lung CT image segmentation using deep learning methods: U-Net versus SegNet, BMC Med. Imaging 21 (1) (2021) 10 Article ID. 19.
[13] A.M. Ismael, et al., The investigation of multiresolution approaches for chest X-ray image based COVID-19 detection, Health Inf. Sci. Syst. 8 (1) (2020) Article ID. 29.
[14] M. Loey, et al., Within the lack of chest COVID-19 X-ray dataset: a novel detection model based on GAN and deep transfer learning, Symmetry-Base12 (4) (2020) Article ID. 651.
[15] M. Topcaci, et al., COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches, Comput. Biol. Med. 121 (2020) 12 Article ID. 103805.
[16] A.K. Das, et al., Automatic COVID-19 detection from X-ray images using ensemble learning with convolutional neural network, Pattern Anal. Appl. (2021) 14, doi:10.1007/s10044-021-00970-4.
[17] R. Susukida, et al., Data management in substance use disorder treatment research: Implications from data harmonization of National Institute on Drug Abuse-funded randomized controlled trials, Clin. Trials (2021) 11, doi:10.1177/1740774520972687.
[18] K. Kumari, et al., Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization, Future Generat. Comput. Syst. Int. J. Esience 118 (2021) 187–197.
[19] A.M. Haner, et al., Replacing human interpretation of agricultural land in Afghanistan with a deep convolutional neural network, Int. J. Remote Sens. 42 (8) (2021) 3017–3038.
[20] J. Hu, et al., Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42 (8) (2020) 2011–2023.
[21] S. Woo, et al., CBAM: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), Munich, Germany, Springer, 2018, pp. 3–19.
[22] H. Sindl, et al., Random fully connected layered 1D CNN for solving the Z-bus loss allocation problem, Measurement 171 (2021) 8 Article ID. 108794.
[23] A. Kumar, et al., Topic-document inference with the gumbel-softmax distribution, IEEE Access 9 (2021) 1313–1320.
[24] P.D. Sathy, et al., Color image segmentation using Kapur, Otsu and minimum cross entropy functions based on exchange market algorithm, Expert Syst. Appl. 172 (2021) 30 Article ID. 114636.
[25] S. Kim, et al., Synthesis of brain tumor multi-contrast MR images for improved data augmentation, Med. Phys. (2021) 14, doi:10.1002/mp.14701.
[26] S.-H. Wang, Covid-19 classification by FGCCNet with deep feature fusion from graph convolutional network and convolutional neural network, Inf. Fusion 67 (2021) 208–229.
[27] X. Cheng, PSSPNN: PatchShuffle stochastic pooling neural network for an explainable diagnosis of COVID-19 with multiple-way data augmentation, Comput. Math. Methods Med. 2021 (2021) 663755.
[28] D. Chicco, et al., The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, Biodata Mining 14 (1) (2021) 22 Article ID. 13.
[29] R.R. Selvaraju, et al., Grad-CAM: visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vision 128 (2) (2020) 336–359.