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.

Neurocomputing 488 (2022) 457–469

Contents lists available at ScienceDirect

Neurocomputing

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

Comparison and ensemble of 2D and 3D approaches for COVID-19 detection in CT images ☆

Sara Atito Ali Ahmed [a,b,1], Mehmet Can Yavuz [a,1], Mehmet Umut Şen [a], Fatih Gülşen [c], Onur Tutar [c], Bora Korkmazer [c], Cesur Samancı [c], Sabri Şirolu [c], Rauf Hamid [c], Ali Ergun Eryürekli [c], Toghrul Mammadov [c], Berrin Yanikoglu [a,*]

[a] Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul 34956, Turkey
[b] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K.7XH, UK
[c] Istanbul University-Cerrahpaşa, Cerrahpaşa Faculty of Medicine, Istanbul 34096, Turkey

ARTICLE INFO

Article history:
Received 9 July 2021
Revised 6 November 2021
Accepted 3 February 2022
Available online 10 February 2022
Communicated by Zidong Wang

Keywords:
COVID-19
Computed Tomography
Detection
Deep Learning
Ensemble

ABSTRACT

Detecting COVID-19 in computed tomography (CT) or radiography images has been proposed as a supplement to the RT-PCR test. We compare slice-based (2D) and volume-based (3D) approaches to this problem and propose a deep learning ensemble, called IST-CovNet, combining the best 2D and 3D systems with novel preprocessing and attention modules and the use of a bidirectional Long Short-Term Memory model for combining slice-level decisions. The proposed ensemble obtains 90.80% accuracy and 0.95 AUC score overall on the newly collected IST-C dataset in detecting COVID-19 among normal controls and other types of lung pathologies; and 93.69% accuracy and 0.99 AUC score on the publicly available MosMedData dataset that consists of COVID-19 scans and normal controls only. The system also obtains state-of-art results (90.16% accuracy and 0.94 AUC) on the COVID-CT-MD dataset which is only used for testing. The system is deployed at Istanbul University Cerrahpaşa School of Medicine where it is used to automatically screen CT scans of patients, while waiting for RT-PCR tests or radiologist evaluation.

1. Introduction

COVID-19 is a highly contagious disease caused by the SARS-CoV-2 virus, which spread rapidly around the world starting early 2020 (Zhu et al. [1]). The definitive diagnosis of COVID-19 is based on real-time reverse transcriptase polymerase chain reaction (RT-PCR) positivity for the presence of coronavirus [2,3].

Due to the long duration to obtain the RT-PCR results and the prevalence of false negative results [4], the medical community has been in search of alternative or supplementary methods, including screening chest X-ray or Computed Tomography (CT) scans of patients for patterns of pneumonia caused by the COVID-19 infection. This work originated at Istanbul University-CerrahpaŞa Hospital, to automatically analyze CT scans while the patient is still in the tomography room, for successful containment of infected cases.

The chest X-ray consists of a single 2-dimensional, frontal image of the thorax, while a chest CT scan consists of a variable number of 2-dimensional axial slice images. The number of slices in a CT volume vary (typically 200–500) and the shape and size of lung tissue within the slice vary significantly between slices. Hence, detection of COVID-19 infection in a chest X-ray presents as a typical image classification problem, while the CT scan provides a richer, but also more challenging input.

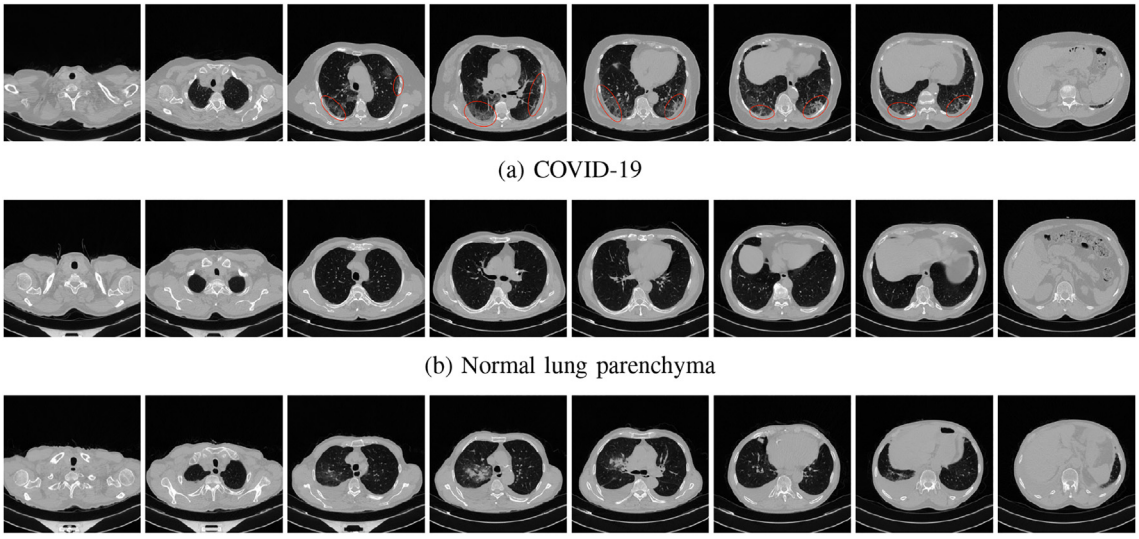
Detecting computed tomography or X-ray images has been studied widely since the beginning of the pandemic [5–14]. Some of these systems only address the 2-class problem: distinguishing between normal and COVID-19 infected parenchyma (e.g [11,12]), while others aim detect COVID-19 infection among all possible conditions (normal lung parenchyma and other lung pathologies, including other types of pneumonia). The latter, which is the

This work was supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) with project number 120E165.

* Corresponding author.
E-mail addresses: sara.atito@gmail.com (S.A. Ali Ahmed), mehmetyavuz@sabanciuniv.edu (M.C. Yavuz), umutsen@alumni.sabanciuniv.edu (M.U. Şen), fatih.gulsen@istanbul.edu.tr (F. Gülşen), onur.tutar@istanbul.edu.tr (O. Tutar), bora.korkmazer@istanbul.edu.tr (B. Korkmazer), cesur.samanci@istanbul.edu.tr (C. Samancı), sabri.sirolu@istanbul.edu.tr (S. Şirolu), rauf.hamid@istanbul.edu.tr (R. Hamid), ali.eryurekli@istanbul.edu.tr (A.E. Eryürekli), toghrul.mammadov@istanbul.edu.tr (T. Mammadov), berrin@sabanciuniv.edu (B. Yanikoglu).

[1] Have contributed equally to this work as first authors.

https://doi.org/10.1016/j.neucom.2022.02.018
0925-2312/



(a) COVID-19

(b) Normal lung parenchyma

(c) Others (including Non-COVID-19 pneumonia, tumors and emphysema.)

Fig. 1. IST-C dataset samples. The ground glass opacities can be observed in the COVID-19 images, marked with the ellipses.

problem addressed in this work, is a significantly more difficult problem as non-COVID-19 pneumonia presents similar patterns to COVID-19.

We developed a deep learning ensemble (IST-CovNet) for detecting COVID-19 infections in high resolution chest CT scans, where we compare and combine *slice-based* and *volume-based* approaches. The slice-based approach takes individual slices as input and outputs the COVID-19 probability for that slice. The system is based on transfer learning using the Inception-ResNet-V2 [15] network that is expanded with a novel attention mechanism [16]. To obtain the patient-level decision from slice-level predictions, we have evaluated different classifier combination techniques, including simple averaging and Long-Short Term Memory (LSTM) networks.

The volume-based approach is based on the DeCoVNet architecture of Wang et al. [9] with some modifications to the architecture. In both approaches, we make use of the pretrained U-Net [17] architecture to focus on the lung regions in the slice images. To combine 2D and 3D systems, we used ensemble averaging, multi-variate regression and Support Vector Machines (SVMs).

A new dataset (IST-C) is collected at Istanbul University-Cerrahpaşa, Cerrahpaşa Faculty of Medicine (IUC), consisting of 712 chest CT scans collected from 645 patients. It includes samples from COVID-19 infected patients, as well as normal lung parenchyma and Non-COVID-19 pneumonia, tumors and emphysema patients. Fig. 1 shows three samples from the IST-C dataset collected in this work, including a typical COVID-19 involvement pattern termed as *ground glass opacity*, along with normal lung parenchyma and other conditions including non-COVID-19 pneumonia, tumors and emphysema.

The contributions of this work are the following:

- We present a deep neural network ensemble (IST-CovNet) that combines 2D (slice-based) and 3D (volume-based) approaches and achieves state-of-art accuracies on the publicly available MosMedData [18] and the IST-C dataset collected in this work. The proposed system also obtains close to state-of-art results on the COVID-CT-MD [19] dataset which is not used for training, demonstrating the inter-operability of the proposed system.
- Rather than adopting a single approach as done commonly in the COVID-19 AI literature, we compare 2D and 3D approaches, along with relevant preprocessing, attention and combination alternatives on 3 different data sets, and combine the best systems to obtain the final ensemble classifier. Our approaches include novel aspects that contribute to improved performance, such as a new attention model and slice-level combination using LSTMs in the 2D system and an extended novel architecture in the 3D approach.
- We have collected a medium-size dataset consisting of 712 high resolution chest CT scans from 645 people, showing normal lung parenchyma, COVID-19 infections, as well as other pathologies (including non-COVID-19 pneumonia, tumors and emphysema). The IST-C dataset is made public along with our results as benchmark[3].
- The system is deployed at one of the biggest hospitals in Turkey (Istanbul University Cerrahpaşa School of Medicine), to screen for CT scans that show COVID-19 infections for timely containment of infected patients.

2. Related Works

Automatic COVID-19 detection research in literature have targeted both chest X-rays [5,11,6] and CT scans [7–10] as input and there have been many systems published in peer-reviewed venues or pre-print sites since the beginning of the pandemic. There are also systems that aim to leverage the potential of the two biomedical imaging modalities, taking as input both a chest CT and a chest X-ray [13,14,20].

Comprehensive literature reviews can be found in surveys about artificial intelligence (AI) based approaches to COVID-19 in [24–26]. Among these surveys, Ozsahin et al. [26] structure their survey into 3 groups: systems aiming to differentiate between i) COVID-19 versus normal lung parenchyma, ii) COVID-19 versus non-COVID-19 (sometimes called COVID-19 negative) consisting of both normal lung parenchyma and other types of pneumonia, and iii) COVID-19 versus other types of pneumonia. Systems included in this survey report the accuracy and/or the Area Under

[3] https://github.com/verimsu/IST-C-dataset

the Curve (AUC) score related to the Receiver Operating Characteristic (ROC) curve. State-of-art results are above 90% accuracy and 0.95 AUC for the first problem (i) and approximately 88% accuracy and 0.90 AUC for the second problem (ii).

AI based COVID-19 detection approaches are twofold: *2D* or *slice-based* approach, taking a single slice image as input and obtain a score for individual slices [11] and *3D* or *volume-based* approach, taking the whole volume (sequence of slices) as the input and produce a single score for the patient [7–9,6]. Note that while a patient may have more than one CT scan, we treat each CT scan as if it belongs to a unique patient and use the terms *CT-level* and *patient-level* interchangeably in this work.

In slice-based models, output scores of slices are often combined by averaging, to obtain the patient-level scores and decisions. Among volume-based approaches, most systems use adaptive-pooling operation for combining slice level features [8,9], while others use a more implicit combination using Recurrent Neural Networks (RNN) [6]. An advantage of 2D models is the direct interpretability while the 3D models is potentially more powerful as they leverage end-to-end optimization rather than a 2-stage process of obtaining patient-level scores after slice-level scores.

In the remainder of this section, we focus on a subset of the literature due to space limitations, reporting systems that analyze CT scans (not X-rays), address the problem of separating COVID-19 samples from all non-COVID-19 samples (not just normal lung parenchyma), and appear on peer-reviewed venues. While we include performance results reported in the referenced works, it should be kept in mind that most of the results cannot be directly compared, as the test datasets or experimental settings vary between systems.

Li et al. [8] developed a model called COVNet, that is based on the Resnet [15] backbone. The varying number of CT slices are input into parallel branches that use shared weights and the deep features extracted from each are combined by a max-pooling operation. They report 0.96 AUC score on the 3-class classification problem of distinguishing between normal lung parenchyma, COVID-19 and other lung pathologies.

Wang et. al. [9] use the pretrained U-Net [17] architecture to segment lung regions and obtain the lung mask volume. Then, the proposed DeCovNet takes the whole CT volume along with the corresponding lung mask volume as input, and outputs a patient-level probability for COVID-19. The variable number of slices is handled using adaptive maxpool operation. Authors report %0.91 accuracy and a 0.959 AUC score on the 2-class problem of separating COVID-19 positive cases from all others (non-COVID-19, including other pneumonia).

Hammoudi et al. [6] split a chest X-ray into patches and after obtaining patch-level predictions using deep convolutional networks, they use bidirectional recurrent networks to combine them to predict patient health status.

Liu et al. [10] fine-tune well-known deep neural networks for the primary task of detecting COVID-19 and the auxiliary task of identifying the different types of COVID-19 patterns (e.g. ground glass opacities, crazy paving appearance, air bronchograms) observed in the slice-image. They report that using the auxiliary task helps with the detection performance, which reaches 89.0% accuracy.

Harmon et al. [27] test the performance of a baseline deep neural network approach in a multi-center study. The approach consists of lung segmentation using AH-Net [28] and the classification of segmented 3D lung regions by pretrained DenseNet121 [29]. On a 1,337-patient test set, they report an accuracy of 0.908 and AUC score of 0.949.

Among systems that report on the MosMedData dataset, Jin et al. [30] propose a slice-based approach employing ResNet-152 [31] architecture. The developed model achieved comparable performance to experienced radiologists with an AUC score of 0.93. He et al. [32] propose a differentiable neural architecture search framework for 3D chest CT-scans classification with the Gumbel-Softmax technique [33] to improve the searching efficiency. The experimental results show that their automatically searched model outperforms three of the state-of-the-art 3D models achieving an accuracy of 82.29% on MosMedData dataset.

In a critical study, Maguolo and Nanni [34] show that some automatic COVID-19 detection systems achieve high accuracies even when the lung region is masked in chest X-rays, indicating that the underlying neural networks are learning patterns in the data that are not correlated to the presence of COVID-19. They also discuss how to construct a fair testing protocol. Our single-channel 3D system that achieves the best results in all datasets inputs CT scans that are *masked* with the lung mask (hence, we can assert that there is no information leakage outside of the lung region). Similarly, our 2D system attends to the lung areas, due to the PCA-based attention module.

In another recent and well-publicized survey, Roberts et al. [35] analyze all COVID-19 AI papers published in the first 9-months period of 2020, in terms of their potential potential biases, according to the criteria indicated in [36]. After filtering the 2,212 papers found in an initial search according to relevance and quality, the remaining 62 papers were analyzed in depth. Authors conclude that "none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases." While this work points out to some important biases in machine learning systems for Covid-19 detection, it is worth pointing out with this categorization, any system evaluated on a public dataset is directly categorized as having a high risk of "participant bias" (since the participants cannot be verified) and all deep learning approaches are categorized as having high risk of "predictor bias" (since deep features are deemed as "abstract and unknown imaging features"). In our work, we evaluate the proposed system on two large public datasets (one not used in training at all) and one private dataset collected in the scope of this work, to address participation and outcome biases. We also report cross-validation results for our final system, to eliminate analysis bias. The results obtained on the unseen data [19] are state-of-art (in AUC) and also close to the results obtained on the other two datasets, attesting to the generality of the system.

3. IST-C Dataset

While there are many works on automatic detection of COVID-19 infection on X-ray or CT images, there were only a handful publicly accessible COVID-19 CT scan datasets at the time of the preparation of this manuscript, shown in Table 1. Three of these datasets, CC-19 [21], MosMedData dataset [18] and BIMCV-COVID19 [22] only contain COVID-19 and normal lung parenchyma. On the other hand, in MosMedData, the COVID-19 samples are also labelled with the severity of the infection in 4 levels (CT-1 to CT-4). In addition to using two large public datasets [18,19] in evaluating the system developed in this study, we have also collected a new open-source dataset called IST-C, retrospectively from patients admitted to the Radiology department of Cerrahpaşa Faculty of Medicine from March 2020 to August 2020. The collected dataset consists of 336 chest CT scans that are positive for COVID-19, along with 245 scans showing normal lung parenchyma and 131 scans from Non-COVID-19 pneumonia, tumors and emphysema patients. The COVID-19 scans are selected by expert radiologists from among the patients to whom CT is performed with clinical suspicion of COVID-19 in the emergency department. These two last groups will be called simply as "Normal" and "Other" from here on. The detailed statistics of the dataset are shown in Table 2.

Table 1
Some of the publicly available COVID-19 CT scan datasets. The first four datasets contain scans of only COVID-19 infected patients and those with normal lung parenchyma. IST-C dataset collected in this work includes non-COVID-19 pneumonia, tumors and emphysema as well.

Dataset	Description	Resolution	# CT Scans	# Slices	# COVID-19	# Normal	# Others
CC-19 [21]	CT scans collected from 3 different hospitals and 6 different scanners	High	89	34,006	68	21	0
MosMedData [18]	CT scans with indicated COVID-19 severity level (4 levels)	High	1,110	46,411	856	254	0
BIMCV-COVID19[22]	COVID-19 and Normal only	High	2,068	314,056	1,141	927	0
COVID-CT-MD [19]	COVID-19, Normal and Other	High	305	45,471	170	77	61
HKBU-HPML-COVID-19 [23]	COVID-19, Normal and Other Collected from different hospitals	High	6,878	406,449	2,513	1,927	2,435
IST-C (this work)	COVID-19, Normal, Other CT scans from one hospital	High	712	200,647	336	245	131

Table 2
Overview of the IST-C dataset: COVID-19 infections are all people diagnosed with the infection; "Normal" is everyone with no infection whatsoever; "Other" is all other types, including pneumonia, tumors and emphysema.

	# Patients	# CT volumes	Total # slices	Avg # slices/person
COVID-19	300	336	92,905	276 ± 83
"Normal"	245	245	67,712	277 ± 67
"Other"	131	131	40,030	306 ± 98
Overall	645	712	200,647	282 ± 82

The collected CT scans in DICOM format consists of 16-bit gray scale images of size 512×512. Each scan is accompanied with a set of personal attributes, such as patient ID, age, gender, location, date, etc. (not used in this work). The average age of the patients is 52 ± 17 years, in which 405 of the patients are male and 274 patients are female.

The annotation of this dataset is at CT scan level: the CT of a patient as a whole is labelled as COVID-19, "Normal", or "Other" by expert radiologists at Istanbul University-Cerrahpaşa, Cerrahpaşa Faculty of Medicine.

Sample images extracted from COVID-19, "Normal" and "Other" classes are shown in Fig. 1. The anonymized dataset is now shared publicly at http://github.com/verimsu.

4. Preprocessing

Pixel values of images in a CT scan are in Hounsfield Unit (HU), which is a radiodensity measurement scale that maps distilled water to 0 and air to -1000. The HU values range between -1024 and 4096, with higher values being obtained from bones and metal implants in the body and lung regions typically ranging in $[-1024, 0]$. Similar to literature, we process chest CT scans such that values higher than $u_{max} = 600$ are mapped to u_{max} and the range $[-1024, u_{max}]$ is normalized to the $[0, 1]$ linearly.

Slice images that are originally 512×512 are resized to match the input size of the respective deep networks, namely 299×299 for slice-based system and 256×256 for the volume-based system. For the 3D approach, we have also reduced the slice count by half, so that the whole CT volume consisting of up to around 500 slice images fits in the GPU memory. We compared two alternatives for this: interpolation of two subsequent slices and skipping every other slide. We found that the latter results in higher accuracy, even though interpolation is commonly used in many biomedical applications. This reduction is done for only the IST-C dataset where the number of slices per CT scan is high (Table II).

5. Lung Segmentation

Lung shapes vary greatly within a chest CT scan, as can be seen in Fig. 1. With the aim of focusing on the lung areas, we make use of the pretrained U-Net network to segment lung regions from non-lung areas. Focusing to lung areas is possible by masking the input with the lung mask as done in the 3D system or guiding the attention of the network to the lung areas. This step is found to be quite important in reducing overfitting [37], as well as information leakage found in some previous COVID-19 detection systems [34].

The U-Net architecture was first proposed by Ronneberger et al. [17] for biomedical image segmentation in general and trained specifically for lungs by Hofmanninger et al. [38]. Since then it has been used in detecting lung regions extensively in the diagnosis of lung health [7–9]. The U-Net network, shown in Fig. 2, is named after the U-shape formed by the encoder branch consisting of convolutional layers and the decoder branch consisting of deconvolution operations. The network also has skip connections in each layer, carrying the output of earlier layers to later layers.

Lung segmentation is applied to individual slices in the CT volume. The output for each slice is the corresponding binary segmentation mask, separating lung areas (including air pockets, tumors and effusions in lung regions) from background or other organs, as shown in Fig. 3. The segmentation extracts left and right lungs separately, although this information is not used in our model.

Lung segmentation with U-Net is very successful, as reported in [38] and also observed in our case. Nonetheless, in order not to miss infected regions, we dilated the masks with a 10-pixel structuring disk. Sample slices from the IST-C dataset and corresponding lung masks obtained by U-Net and the dilated masks are shown in Fig. 3.

6. Slice-based Approach

In the 2D approach, CT slices are analyzed independently, before combining them to obtain patient-level predictions.

6.1. Base Model

To construct the base network architecture, we employed Inception-ResNet-V2 architecture [15], one of the top-ranked architectures of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2014 [39]. The network architecture was used successfully in various image classification and object detection tasks [40,41].

Inception-ResNet-V2 network is an advanced convolutional neural network that combines the inception module with ResNet [31] to increase the efficiency of the network. The network is 164 layers deep with only 55.9 million parameters. It consists of three main reduction modules with 10, 20 and 10 inception blocks, respectively. The size of the output feature maps of the three reduction modules are $35 \times 35, 17 \times 17$, and 8×8, respectively.

Training a large deep learning network from scratch is time consuming and requires a tremendous amount of training data.

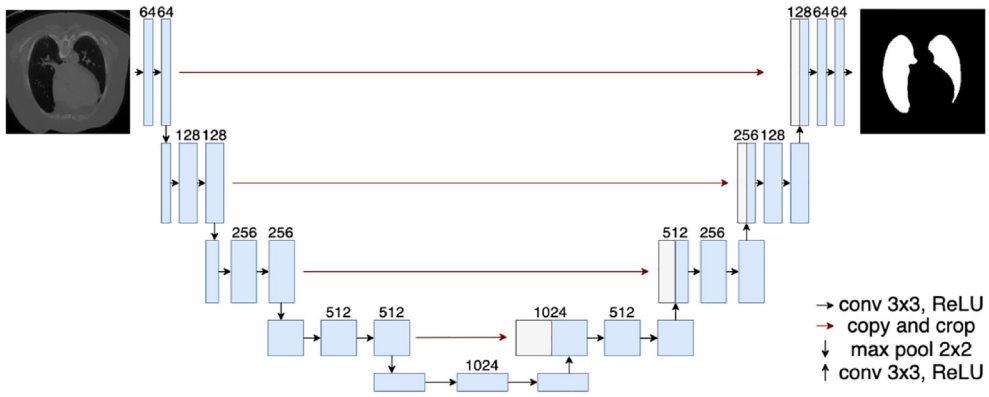


Fig. 2. Segmentation network U-Net [17]: input is a slice image and the output is the corresponding lung mask.

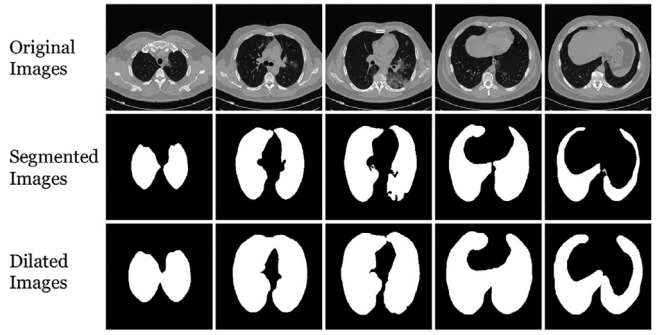


Fig. 3. Sample slice images along with their segmentation masks as obtained by U-Net and dilated masks.

Therefore, our approach is based on fine-tuning a pre-trained Inception-ResNet-V2 model, that is originally trained on the ImageNet dataset with 1.2 million hand-labeled images of 1,000 different object classes.

6.2. Attention Mechanism

To investigate the predictions of the trained base model, we applied Class Activation Mapping (CAM) [42] on some of the images from the validation set. Observing that the attention of the network is not always directed to the area of interest (lung tissues) in misclassified images, we decided to use attention maps and thereby guide the network to the regions that are important to the problem at hand. Attention mechanism has been successfully applied in many computer vision tasks, including fine-grained image recognition [43] and face attributes classification [44].

We add an attention map block inserted to the backbone of our base network, as shown in Fig. 4. The input to the attention layer is a convolutional feature map $\mathbf{F} \in R^{H\times W\times C}$, where H, W, and C are the height, width, and the number of channels, respectively. The output of the attention module is the masked feature map $\mathbf{F}' = \mathbf{F} \odot \sigma(\Phi(\mathbf{F}))$, obtained via element-wise multiplication of the feature maps $\mathbf{F}$ and sigmoid (σ) attenuated attention layer output, $\Phi(\mathbf{F}) \in R^{H\times W}$.

Unlike the standard approach of learning the attention layer fully within the network, the approach used in this work is suggested to be an explainable and modular approach [16]. It makes the assumption that an attention map can be represented using the linear combination of a set of basis vectors, as:

$$\Phi(\mathbf{F}) = \overline{\mathbf{M}} + \mathbf{B} \times \boldsymbol{\alpha}$$

where $\overline{\mathbf{M}} \in R^{H\times W}$ is the average segmentation map; H and W are the height and width of the images; $\mathbf{B} \in R^{H\times W\times n}$ is the matrix of the n basis vectors; and $\boldsymbol{\alpha} \in R^{n\times 1}$ are the coefficients.

The average lung map $\overline{\mathbf{M}}$ and the 12 basis vectors $\mathbf{B}$ are obtained by applying Principal Component Analysis (PCA) to lung masks obtained by U-Net segmentation network. The 12 basis vectors that retain approximately 75% of the variance are shown in Fig. 5 and U-Net is explained in Section 5.

To obtain the attention map coefficients $\boldsymbol{\alpha}$, an additional convolutional block is inserted to the network getting the input from the feature maps $\mathbf{F}$, as shown in Fig. 4. The convolutional block consists of a separable convolutional layer which is a depth-wise convolution performed independently over each channel of an input, followed by a pointwise convolution, batch normalization, and ReLU activation function. The output of the convolutional block (or attention coefficient block) are the weights $\boldsymbol{\alpha}$ which form the coefficients in the linear basis vector representation.

6.3. Implementation Details

The Inception-ResNet-V2 network used as the base model in the slice-based approach is chosen due to its relatively small size and good performance. The network has an RGB image input size of 299×299. The output layer of the model is replaced with a fully connected layer with 2 units to represent the given classes: COVID-19 vs Non-COVID-19 (including "Normal" and "Other" samples). All the layers in the classification network are finetuned and optimized using categorical cross-entropy loss function.

For the attention based model, we added the attention layer after the first reduction block as shown in Fig. 4. As for the attention loss function, we trained the network in unsupervised manner. Even in the absence of the attention map supervision, we found that the attention module is able to learn the discriminative regions automatically.

The implementation is done using the Inception-ResNet-V2 model provided in the Matlab deep learning toolbox. Several commonly used data augmentation techniques are applied during training, such as rotation ($[-5, 5]$ degrees), x and y translation ($[-5, 5]$ pixels), and x and y scaling ($[0.9, 1.1]$).

For all 2D systems, we set the batch size equal to 64 and the initial learning rate as 1e-5 with a total of 50 epochs using the Adam optimizer. The training process takes around 100 min per epoch for the IST-C dataset and 40 min per epoch for the MosMedData dataset using an 8 GB Nvidia GeForce RTX 2080 GPU.

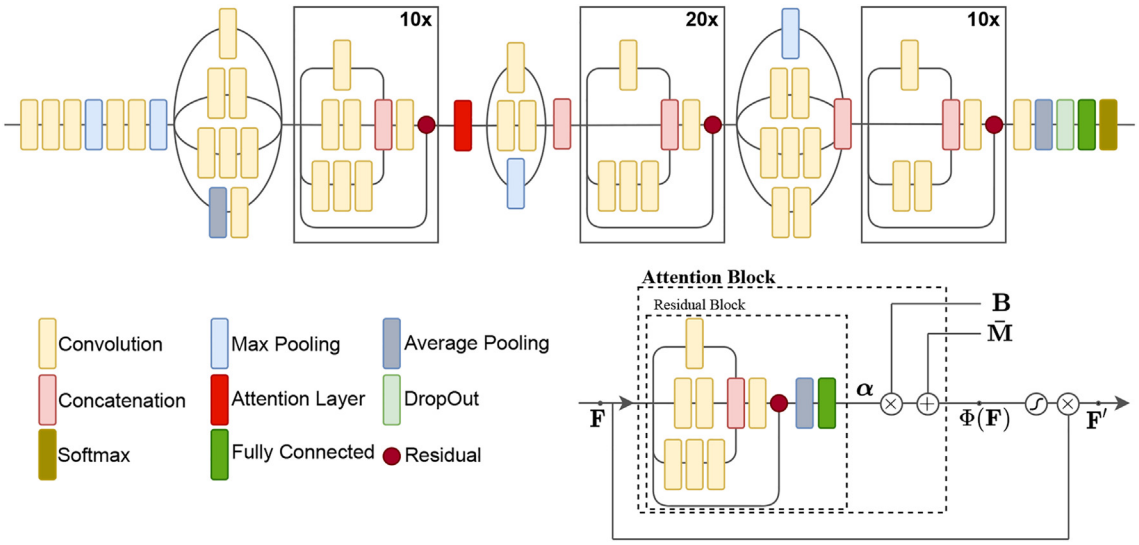


Fig. 4. The base network and the inserted attention-based layer. Attention layer takes the feature maps **F** as an input and estimate the attention map $\Phi(\mathbf{F})$, which is then used to attend to the original features after a sigmoid activation.

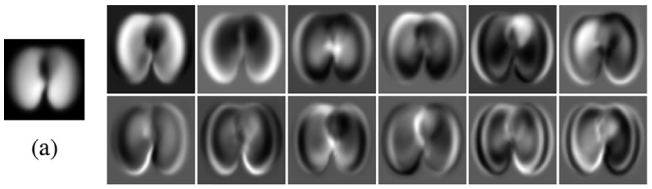


Fig. 5. (a) Mean mask $\overline{M}$ and (b) The first 12 eigenvectors.

6.4. Combining Slice-level Predictions

The straightforward approach to obtain patient-level decision is to combine the predictions of the slice based model using simple averaging of slice-level predictions. This is evaluated as the base model, to obtain the patient-level score.

However, simple averaging does not take into account the information about the characteristics of COVID-19 infection, such as the fact that the patterns are often seen in the lower parts of the lungs. To learn this type of information about the slice sequence and to also handle the variable length of the slice sequence, we also used Recurrent Neural Networks (RNNs) as an alternative [45].

We used Long Short-Term Memory (LSTM) network [46] that is the most powerful type of recurrent network. The input to the network consists of deep features corresponding to each slice in the CT volume. The features are extracted from the last pooling layer of the slice-based CNN model with the attention module, discussed in Section 6.2. The LSTM learns to combine the slice-level features to obtain patient-level predictions.

The LSTM architecture consists of 3 layers: i) a bidirectional LSTM layer with 1024 hidden units and a dropout layer to reduce overfitting; ii) another bidirectional LSTM layer with 512 hidden units; and iii) a fully connected layer with an output size corresponding to the number of classes (2 or 3 in our case). It is important to note that the number of slices in the CT volumes varies substantially which can introduce lots of padding into the training process of the LSTMs and consequently negatively impact the classification accuracy. To overcome this issue, we normalized each CT sequence into 282 slices (the mean slice count across the IST-C dataset), by either dropping or replicating slices depending on the length of the volume. After normalization, each slice of the CT volume is passed to the trained CNN model for feature extraction. Then, the LSTM model is trained using the sequence of the feature vectors corresponding to the slices.

7. Volume-base Approach

The 3D volume-based approach takes as input the whole CT volume and outputs patient-level decision (COVID-19 positive and negative probabilities), based on a single step processing of the input. It uses the lung segmentation volume obtained by U-Net (described in Section 5), followed by a classification network based on DeCoVNet [9].

The segmentation network (U-Net) takes as input a single slice of the chest CT and outputs a binary mask indicating the lung region. The classification network subsequently takes the segmented CT volume and outputs the patient-level scores.

7.1. Classification network

The classification network used in our work is based on DeCoVNet that has been proposed by Wang et al. [9]. We have made some modifications to this network, without significantly changing its architecture. The network consists of three consecutive blocks, (1) Stem (2) ResBlocks (3) Classifier, as shown in Fig. 6 and detailed in Table 3.

The stem block consists of a convolutional layer with a receptive field size $5\times7\times7$ (depth, height, width), as used in well-known networks AlexNet [47] and Resnet [31]. The convolutional layer is followed by a batchnorm layer and a pooling layer. We evaluated using both a single channel input, consisting of the slice image with the lung mask applied, as well as the 2-channel input, consisting of the input slice and its lung mask, as in the original network. As we expected, the 2-channel approach led to less efficient training and did not bring accuracy gains.

The second block of the network is adopted without any modification. It consists of two 3D residual blocks (ResBlocks), with maxpool operation in between, to reduce the volume depth by half to $64\times T/2\times64\times64$. In each residual block, there are 2 kernels: $3\times1\times1$, $1\times3\times3$ (depth,height,width) with a stride of 1 in each dimension and padding of 1 wherever needed. The output volume is of size $128\times T/2\times32\times32$. This block is adopted without any modification.

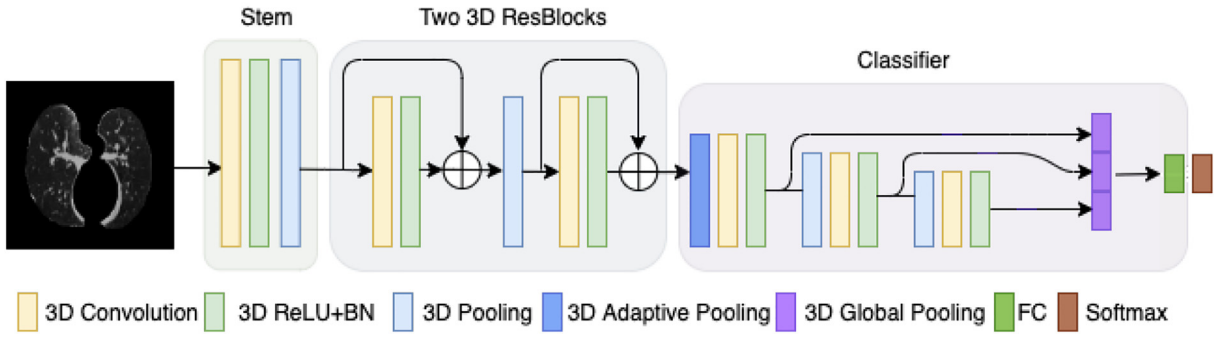


Fig. 6. Architecture of the classification network which is based on DeCoVNet [9].

Table 3
The 3D-classification network architecture. The residual blocks have two kernels.

	Operation	Output	Penultimate
Stem	Conv3d@5×7×7	16×T×64×64	
ResBlocks	ResBlock@3×1×1&1×3×3	64×T×64×64	
	MaxPool3d	64×T/2×64×64	
	ResBlock@3×1×1&1×3×3	128×T/2×64×64	
Progressive Classifier	AdaptiveMaxPool3d	128×16×32×32	
	Conv3d@3×3×3	96×16×32×32	
	GlobalPool3d		96×1×1×1
	———— 2nd Block	————	
	MaxPool3d	96×4×16×16	
	Conv3d@3×3×3	48×4×16×16	
	Dropout3d (p = 0.5)	48×4×16×16	
	GlobalPool3d		48×1×1×1
	———— 3rd Block	————	
	MaxPool3d	48×4×16×16	
	Conv3d@3×3×3+ReLU	48×4×16×16	
	GlobalPool3d		48×1×1×1
	FullyConnected	2	

The third block, called the Progressive classifier, starts with an adaptive maxpool operation that handles the variable number of slices and outputs 128×16 feature maps of size 32×32. It is followed by 3 convolution layers and pooling operations, followed by a fully connected output layer with softmax activation. The main modification in this block is to enrich the feature representation. The original DeCoVNet had a global max pooling layer with 32×1×1×1 nodes, in the penultimate layer. We extended the Progressive classifier block by adding a new layer of concatenated features obtained using global max pool operation after each of the 3D convolutional layers. More specifically, from a convolutional layer with F×D×H×W output volume, the global max pooling operation outputs a vector of size *F*. The resulting 192-dimensional (96 + 48 + 48) feature vector is fully connected to the output layer (2 nodes with softmax activation), as shown in Fig. 6. We thus increased the penultimate layer size from 32 to 192. This feature representation was inspired by the work in [48], where authors proposed to approximate a deep learning ensemble by replicating the output layer with connections from earlier layers and extending the loss function to include all the loss terms [48].

The modified classification network architecture is given in Table 3.

7.2. Implementation Details

In training the system, the settings are the same and as follows: the loss function is the categorical cross-entropy; the optimizer is the Adam optimizer used with 1e-5 learning rate. Since the graphical card Nvidia 2080 can only process a single volume at a time, the batch size is one. We also used data augmentation exactly the same way as DeCoVNet: scaling ($[1, 1.2]$), rotation ($[-10, 10]$ degrees) and translation ($[-10, 10]$ pixels). All 3D systems were run for a fixed number of 200 epochs, observing validation set accuracy at each epoch. The optimal weights were chosen as those giving the highest validation set results. The training process takes around 8 min for an epoch of the IST-C dataset and 4 min for an epoch of MosMedData dataset, using an 8 GB Nvidia GeForce RTX 2080 GPU.

8. Combining Multiple Systems

After training the 2D and 3D systems, we combine output of the systems (*patient-level* predictions) to obtain the final prediction. In contrast, please note that Section 6.4 discusses the combination of *slice-level* predictions to obtain patient-level predictions for the 2D approach.

The 2D (slice-based) approach is realized with or without the attention mechanism and using different combination mechanisms to obtain the patient-level decision. Similarly, the 3D (volume-based) approach is realized with 1-channel input where the input is masked with the lung mask, or with 2-channel input as in the original DeCovNet [9].

The combination methods that were evaluated were averaging, multivariate linear regression and Support Vector Machines (SVM). However, we only report ensemble averaging results because multi-variate regression essentially assigned the same weights to

the two combined systems and the SVM did not bring noticeable improvements to justify the more complex combination method.

9. Experimental Evaluation

We have trained and evaluated the proposed 2D and 3D approaches along with considered submodules, with the IST-C collected in this work (Section 3) and and the MosMedData dataset [18]. These results are given in Tables 4 and 5, respectively. Furthermore, we report results of the above trained models on the COVID-CT-MD dataset [19], to evaluate inter-operability performance. These results are given in Table 6.

We have done extensive evaluation comparing different preprocessing, segmentation, architecture and ensemble methods. However for the sake of clarity, we report only the most important experiments, using accuracy and AUC scores, in line with the literature. The accuracy values are given together with 95% confidence intervals that are computed using the Wilson score interval method [49] for the number of test samples in each dataset.

We split the IST-C database into training/validation/testing data. For "COVID-19" class, 100 volumes are used for testing and the rest are used of the training and the validation. For "Normal" and "Others" classes, 100 and 50 volumes are used for testing, respectively. In total, we assigned 250 volumes for testing and 462 for training and validation. The MosMed dataset was split randomly as train-test, with a 80–20% split, resulting in a 222 test samples. The full COVID-CT-MD dataset was used only for testing.

9.1. 2D vs 3D

We first compared the effectiveness of 2D and 3D approaches in identifying COVID-19 positive samples in IST-C and MosMedData datasets. Specifically, we evaluated the 2D approach with or without using the attention module and using simple averaging or the LSTM architecture for combining slide-level features/predictions. For the 3D approach, we compared using a single channel as input (the masked CT scan), two-channel (CT scan and segmentation masks separately). Only the best configurations were evaluated for MosMedData due to long training times needed.

For MosMedData, the systems were trained *only* on the training portion of MosMedData to separate COVID-19 positive samples from the Normal class and tested on the MosMedData test portion, with results given in Table 5. For IST-C dataset, the systems were pretrained with all of the 1,110-sample MosMedData and fine-tuned on the IST-C training set.

The state-of-art results from the literature are also included whenever available [30,32]. We have also implemented DeCovNet [9], that our 3D approach is based on, using the code supplied by the authors[4], following the same training procedure used for our 3D model.

Considering the results given in Tables 4 and 5, we see that the best 2D and 3D approach have the same accuracy on the IST-C datasets (87.20%), while the 3D system is slightly better for the MosMedData dataset (93.24% vs 91.89%) and slightly better in AUC score in both datasets. However it should be noted that training was faster for the 3D dataset per epoch thanks to the Python environment (vs. Matlab) and the smaller 3D network afforded longer training times (200 vs 50).

The 2D system on the other hand can be said to be more explainable, since it is possible to view slice-level decisions to identify where COVID-19 infection patterns are detected by the system; this information can be displayed to the attending physicians in the deployed system.

[4] https://github.com/sydney0zq/covid-19-detection

Table 4
Performance results for the IST-C test set with $n = 250$ samples from 3 classes. The 2D systems are trained with only IST-C and the 3D systems were trained with MosMedData and IST-C training subsets. DeCoVNet results are obtained with author supplied code. Bold figures indicate the best accuracy in slice-based or volume-based approaches.

Model	Accuracy (%)	AUC
2D - Base Network + Averaging	80.80 ± 4.88	0.87
2D - Base + Attention + Averaging	85.60 ± 4.35	**0.90**
2D - Base + Attention + LSTM	**87.20 ± 4.14**	0.89
3D - DeCoVNet [9]	78.00 ± 5.14	0.78
3D - single channel - interpolation	82.80 ± 4.68	0.86
3D - single channel - skipping	**87.20 ± 4.14**	**0.90**
3D - two channels - skipping	81.45 ± 4.82	0.86
Ensemble - Averaging (IST-CovNet)	**90.80 ± 3.58**	**0.95**

Table 5
Performance results for the MosMedData [18] test set with $n = 222$ samples from only 2 classes (COVID-19 and Normal). Our approaches are trained using only MosMedData training subset. DeCoVNet results are obtained with author supplied code. Bold figures indicate the best results in the literature and among our two different approaches.

Model	Accuracy (%)	AUC
Jin et al. (2D) [30]	-	**0.93**
He et al. (3D) [32]	82.29	-
3D - DeCoVNet [9]	**82.43**	0.82
2D - Base + Attention + Averaging	90.09 ± 3.93	**0.96**
2D - Base + Attention + LSTM	**91.89 ± 3.59**	0.95
3D - single channel - skipping	**93.24 ± 3.30**	**0.96**
Ensemble - Averaging (IST-CovNet)	**93.69 ± 3.20**	**0.99**

Table 6
Inter-operability results using the COVID-CT-MD [19] dataset with $n = 305$ samples from 3 classes. Our ensemble system was trained using *only* MosMedData and IST-C datasets to measure the inter-operability of the developed system. Bold figures indicate the best results in the literature and among our approaches.

Model	Accuracy (%)	AUC
COVID-FACT [50]	**91.83**	-
CT-CAPS [51]	89.80	**0.930**
Deep-CT-Net [52]	86.00	0.886
2D - Base + Attention + Averaging	75.41 ± 4.84	**0.838**
2D - Base + Attention + LSTM	**79.34 ± 4.55**	0.819
3D - single channel	**87.87 ± 3.67**	**0.931**
Ensemble Averaging (IST-CovNet)	**90.16 ± 3.35**	**0.942**

9.2. Comparison to the Results in Literature

Our best results obtained on the IST-C dataset is 90.80% accuracy and 0.95 AUC score with ensemble averaging of the best 2D and and best 3D system (Table 4).

The results obtained on the MosMedData dataset with only COVID-19 and Normal classes are better as expected (93.69% accuracy and 0.99 AUC), given the relatively simpler problem with two classes (Table 5). In comparison to the best results in the literature, our ensemble accuracy (93.69%) is 10% points higher compared to the state-of-art and the AUC score (0.99) is also very high, exceeding the state-of-art.

9.3. Evaluating Novel Sub-Modules

Considering the results in Table 4, we see that the attention layer in the 2D approach increases the accuracy significantly

(85.60% vs 80.80% in IST-C), in line with other problems where attention brings performance increase in literature.

The use of LSTM to obtain the patient-level predictions from slice-level features brings another 1.6 and 1.8% points improvements in accuracy, for IST-C and MosmedData, compared to averaging the slice-level predictions. The CT sequence size normalization in LSTM training is an important aspect for this improvement. On the other hand, the LSTM achieves lower AUC scores compared to averaging; we expect that this is due to LSTM outputs being close to 0 or 1.

For the 3D approach, we observed that the 2-channel input also used in DeCoVNet achieves significantly lower accuracy (81.45% vs 87.20%), probably due to the difficulty in training the first layer weights and the success in obtaining good segmentation masks (see Table 4).

The model trained with the author supplied DeCoVNet [9] also achieved lower results (Table 4) compared to our extended version (78.00% vs 81.45% obtained by the two-channel system that is basically the same as DeCovNet except for the added skip connections), showing the benefits of extending the network to deal with the rich information present in the CT scan.

Additionally, we found that the interpolation done to halve the large CT volume in the case of the IST-C dataset, leads to significantly lower performance (%87.20 vs %82.80) compared to skipping every other slice, presumably due to the loss of the fine details in the images. This is something to be aware of when dealing with this or similar problems, as interpolation is commonly used in many biomedical applications.

9.4. Inter-Operability

To study the inter-operability of systems with respect to different datasets collected from different patient populations and tomography equipment and settings, we tested the accuracy of the systems trained using the MosMedData and IST-C datasets, on the COVID-CT-MD dataset [19]. As the COVID-CT-MD dataset was not used in training at all, we used the whole dataset for testing. Hence our results are obtained on the whole dataset, while others are obtained on the testing portion of the dataset. COVID-CT-MD dataset comprises 305 CT scans from 3 classes, as indicated in Table 2.

The results shown in Table 6 (90.16% accuracy and 0.9418 AUC) are in line with results reported in literature, even though our systems were not trained or finetuned at all for this dataset. In particular, the AUC of the ensemble is highest and accuracy value is only slightly behind the best reported results in literature for this dataset [50].

Furthermore, while the results are not directly comparable, our results on COVID-CT-MD dataset show only a slight decrease compared to the IST-C dataset results (90.80% accuracy and 0.95 AUC vs 90.16% accuracy and 0.942 AUC), indicating the generality of the proposed system.

9.5. Error Analysis

The confusion matrix of the ensemble that obtained 90.8% accuracy on the IST-C dataset (4) is given in Table 7. The system predicted 9 false negatives (9/100 COVID-19 samples) and 14 false positives (9/100 Normal and 5/50 Other samples) in total. Hence the error rates were almost the same in each group.

An analysis of the errors by expert physicians revealed that the majority (6/9) of the false negatives were due to minimal lung involvement or respiratory motion artifacts. Respiratory motion artifacts were also observed alone or with atelectasis in 4/9 false positives with normal parenchyma.

Table 7
Confusion matrix for the IST-C dataset.

Actual | Predicted	Covid-19	Non-Covid-19
Covid-19	**91**	9
Normal	9	**91**
Other	5	**45**

9.6. Prediction Scores Distribution

The system is designed to alert the attending physicians in case of sufficiently high COVID-19 probability. Hence, we also considered the COVID-19 prediction distribution of the ensemble, shown in Fig. 7. An adjustable threshold (e.g. 0.3 or 0.4) can be set to alert the attending physician, at the risk of some increased false positives.

At 0.3 threshold, we obtain 95.0% sensitivity (true positive rate) and 84.0% specificity (true negative rate) on the IST-C test data set. ROC figures corresponding to IST-C and MosMedData datasets are given in Fig. 8.

9.7. Lung Segmentation Results

Regarding lung segmentation accuracy, Hofmanninger et al. [38] report 97–98% Dice similarity scores measuring how much the mask generated by U-Net and ground-truth overlaps, on different test datasets involving multiple lung pathologies. While their tested datasets also included ground glass opacities observed in COVID-19 cases, we evaluated the segmentation network's performance specifically for the COVID-19 detection problem by visually checking the segmentation results of 5 slices from sampled at regular intervals from 1,156 CT scans (all covid patients from IST-C and MosMedData datasets), for a total of 5,783 slice images. We found around 11 serious segmentation errors, corresponding to roughly %0.19, which is in line with [38]. Samples of these images are given in Fig. 9, where lung areas that are considered as background and are highlighted by ellipses. Noting that the errors occur only in some of the slices within one CT scan, we conclude that U-Net provides a successful segmentation, suitable for COVID-19 detection.

9.8. Discussion

While our 3D approach is based on DeCoVNet [9], we were able to outperform its results on both datasets, thanks to the changes made to the model. In particular, using only one input channel leads to more efficient training, especially since the U-Net lung segmentation is very accurate, while enriching the network architecture also contributed to higher accuracy.

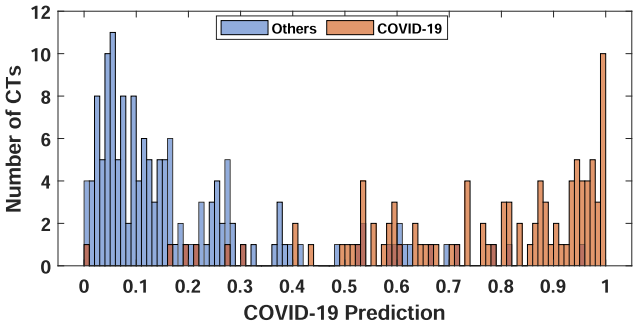


Fig. 7. COVID-19 predicted probability distribution for the IST-C dataset, using the ensemble.

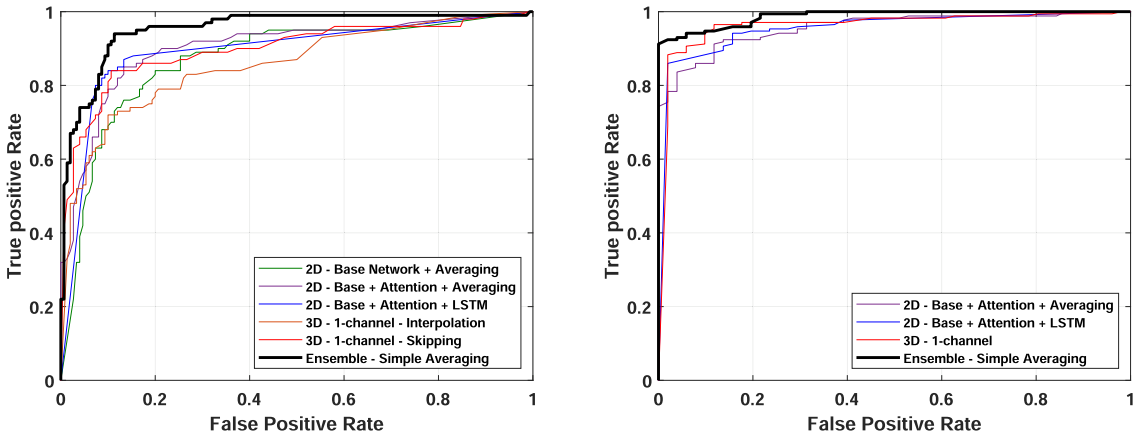


Fig. 8. ROC curves of the trained models on (a) IST-C dataset and (b) MosMedData dataset.

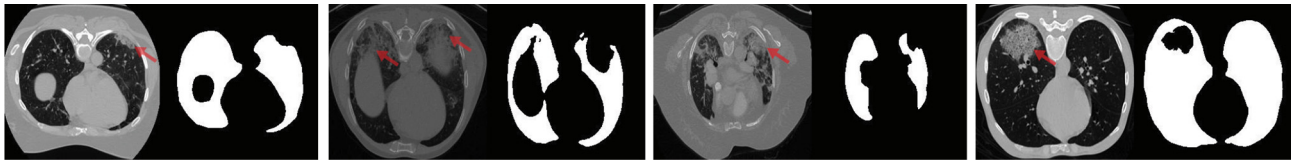


Fig. 9. Samples of segmentation errors (a) slice image (b) corresponding lung masks. Problematic areas are indicated with red arrows and are often missed lung tissue due to infection or tumors.

Similarly, even though the 2D system is based on fine-tuning a pretrained deep network, the use of the novel attention mechanism and LSTMs to combine slice-level features bring significant improvements over the base network and the standard approach of averaging slice predictions. We are aware of only one other work that also combines a deep network with LSTMs, related to COVID-19 predictions: Hammoudi et al. [6] use bidirectional LSTMs to predict patient health status by combining the predictions made by a deep network for *image-patches* of an X-ray.

Considering the results in Table 4, we see that our contributions improve accuracies by 6.4 and 9.20 percentage points, in 2D and 3D models respectively (%87.20 vs %80.80 and %87.20 vs %78.00). Furthermore, we gain another 3.6 percentage points when we combine the 2D and 3D systems (%90.80 vs %87.20). Hence, while the main contributions of our work are in the network architectures, the ensemble approach also brings significant improvements.

10. Conclusion

In addition to presenting a state-of-art system, we provide an evaluation of different 2D and 3D approaches on two datasets and discuss the effects of relevant preprocessing, segmentation and classifier combination steps on performance. A third large and public dataset is used to show inter-operability results.

The collected dataset (IST-C) is made public to contribute to the literature as a challenging new dataset that consists of high resolution chest CT scans from a variety of conditions.

This work was motivated to help combat the pandemic and the developed system (IST-CovNet) is deployed and in use at Istanbul University Cerrahpaşa School of Medicine, to flag suspected COVID-19 cases when the patient is still at the tomography room.

In future work, we plan to study how to best use semi-supervised approaches (especially with 3D volumes) to leverage larger collections of unlabelled chest CTs.

Credit authorship contribution statement

Sara Atito Ali Ahmed: Methodology, Software, Validation, Writing - review & editing. **Mehmet Can Yavuz:** Methodology, Software, Validation, Writing - review & editing. **Mehmet Umut Şen:** Methodology, Software. **Fatih Gülşen:** Conceptualization, Funding acquisition, Validation, Supervision. **Onur Tutar:** Investigation, Validation. **Bora Korkmazer:** Investigation, Validation. **Cesur Samancı:** Investigation, Validation. **Sabri Şirolu:** Conceptualization, Investigation, Validation, Data curation. **Rauf Hamid:** Conceptualization, Investigation, Validation, Data curation. **Ali Ergun Eryürekli:** Investigation, Data curation. **Toghrul Mammadov:** Investigation, Data curation. **Berrin Yanikoglu:** Conceptualization, Funding acquisition, Methodology, Validation, Supervision, Writing - review & editing.

Declaration of Competing Interest

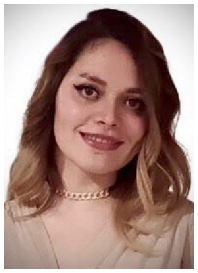
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, et al., A novel coronavirus from patients with pneumonia in China, 2019, New England Journal of Medicine (2020).

[2] V.M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D.K. Chu, T. Bleicker, S. Brünink, J. Schneider, M.L. Schmidt, et al., Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR, Eurosurveillance 25 (3) (2020) 2000045.

[3] G.D. Rubin, C.J. Ryerson, L.B. Haramati, N. Sverzellati, J.P. Kanne, S. Raoof, N.W. Schluger, A. Volpi, J.-J. Yim, I.B. Martin, et al., The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the Fleischner Society, Chest 158 (1) (2020) 106–116.
[4] Q.-X. Long, X.-J. Tang, Q.-L. Shi, Q. Li, H.-J. Deng, J. Yuan, J.-L. Hu, W. Xu, Y. Zhang, F.-J. Lv, et al., Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections, Nature Medicine 26 (8) (2020) 1200–1204.
[5] L. Wang and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," arXiv preprint arXiv:2003.09871, 2020.
[6] K. Hammoudi, H. Benhabiles, M. Melkemi, F. Dornaika, I. Arganda-Carreras, D. Collard, and A. Scherpereel, "Deep learning on chest X-ray images to detect and evaluate pneumonia cases at the era of COVID-19," arXiv preprint arXiv:2004.03399, 2020.
[7] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Q. Ni, Y. Chen, J. Su, et al., A deep learning system to screen novel coronavirus disease 2019 pneumonia, Engineering 6 (10) (2020) 1122–1129.
[8] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, et al., Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT, Radiology (2020).
[9] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, C. Zheng, A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT, IEEE Transactions on Medical Imaging 39 (8) (2020) 2615–2625.
[10] B. Liu, X. Gao, M. He, L. Liu, G. Yin, A fast online COVID-19 diagnostic system with chest CT scans, in: Proceedings of KDD, 2020.
[11] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using X-ray images and deep convolutional neural networks," arXiv preprint arXiv:2003.10849, 2020.
[12] X. Yu, S. Lu, L. Guo, S.H. Wang, Y.D. Zhang, Resgnet-c: A graph convolutional neural network for detection of COVID-19, Neurocomputing 452 (2021) 592–605.
[13] Y.-D. Zhang, Z. Zhang, X. Zhang, S.-H. Wang, MIDCAN: A multiple input deep convolutional attention network for COVID-19 diagnosis based on chest CT and chest X-ray, Pattern Recognition Letters 150 (2021) 8–16.
[14] A. Chaddad, L. Hassan, C. Desrosiers, Deep CNN models for predicting COVID-19 in CT and X-ray images, Journal of Medical Imaging 8 (S1) (2021) 014502.
[15] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, "Inception-V4, Inception-Resnet and the impact of residual connections on learning," in Thirty-First AAAI Conf. on, Artificial Intelligence (2017).
[16] H. Dang, F. Liu, J. Stehouwer, X. Liu, A.K. Jain, On the detection of digital face manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5781–5790.
[17] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation, in: Int. Conf. on Medical Image Computing and Computer-assisted Intervention, Springer, 2015, pp. 234–241.
[18] S. Morozov, A. Andreychenko, N. Pavlov, A. Vladzymyrskyy, N. Ledikhova, V. Gombolevskiy, I.A. Blokhin, P. Gelezhe, A. Gonchar, and V.Y. Chernina, "Mosmeddata: Chest CT scans with COVID-19 related findings dataset," arXiv preprint arXiv: 2005.06465, 2020.
[19] P. Afshar, S. Heidarian, N. Enshaei, F. Naderkhani, M.J. Rafiee, A. Oikonomou, F. B. Fard, K. Samimi, K.N. Plataniotis, and A. Mohammadi, "COVID-CT-MD: COVID-19 computed tomography (CT) scan dataset applicable in machine learning and deep learning," arXiv 2009.14623, 2020.
[20] P.K. Chaudhary, R.B. Pachori, FBSED based automatic diagnosis of COVID-19 using X-ray and CT images, Computers in Biology and Medicine 134 (2021) 104454.
[21] R. Kumar, A.A. Khan, S. Zhang, W. Wang, Y. Abuidris, W. Amin, and J. Kumar, "Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging," arXiv preprint arXiv:2007.06537, 2020.
[22] M. de la Iglesia Vayá, J.M. Saborit, J.A. Montell, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. García-García, M. Caparrós, G. Gonzáclez, and J.M. Salinas, "BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients," arXiv 2006.01174, 2020.
[23] X. He, S. Wang, S. Shi, X. Chu, J. Tang, X. Liu, C. Yan, J. Zhang, and G. Ding, "Benchmarking deep learning models and automated model design for COVID-19 detection with chest CT scans," medRxiv, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/06/17/2020.06.08.20125963.
[24] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, D. Shen, Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19, IEEE Reviews in Biomedical Engineering (2020) 1.
[25] M.M. Islam, F. Karray, R. Alhajj, J. Zeng, A Review on Deep Learning Techniques for the Diagnosis of Novel Coronavirus (COVID-19), 9, IEEE Access, 2021, pp. 30551–30572.
[26] I. Ozsahin, B. Sekeroglu, M.S. Musa, M.T. Mustapha, D.U. Ozsahin, Review on diagnosis of COVID-19 from chest CT images using artificial intelligence, Computational and Mathematical Methods in Medicine 9756518 (2020).
[27] S.A. Harmon, T.H. Sanford, S. Xu, E.B. Turkbey, H. Roth, Z. Xu, D. Yang, A. Myronenko, V. Anderson, A. Amalou, et al., Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets, Nature Communications 11 (1) (2020) 1–7.
[28] S. Liu, D. Xu, S.K. Zhou, O. Pauly, S. Grbic, T. Mertelmeier, J. Wicklein, A. Jerebko, W. Cai, D. Comaniciu, 3D anisotropic hybrid network: Transferring convolutional features from 2D images to 3D anisotropic volumes, in: Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 851–858.
[29] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[30] C. Jin, W. Chen, Y. Cao, Z. Xu, Z. Tan, X. Zhang, L. Deng, C. Zheng, J. Zhou, H. Shi, et al., Development and evaluation of an artificial intelligence system for COVID-19 diagnosis, Nature Communications 11 (1) (2020) 1–14.
[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[32] X. He, S. Wang, X. Chu, S. Shi, J. Tang, X. Liu, C. Yan, J. Zhang, and G. Ding, "Automated model design and benchmarking of 3D deep learning models for COVID-19 detection with chest CT scans," arXiv preprint arXiv:2101.05442, 2021.
[33] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-softmax," arXiv preprint arXiv:1611.01144, 2016.
[34] G. Maguolo, L. Nanni, A critic evaluation of methods for COVID-19 automatic detection from X-ray images, Information Fusion 76 (2021) 1–7.
[35] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A.I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al., Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nature Machine Intelligence 3 (3) (2021) 199–217.
[36] R.F. Wolff, K.G. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett, PROBAST: a tool to assess the risk of bias and applicability of prediction model studies, Annals of Internal Medicine 170 (1) (2019) 51–58.
[37] N. Gupta, A. Kaul, D. Sharma et al., "Deep learning assisted COVID-19 detection using full CT-scans," TechRxiv 10.36227/techrxiv.13162049.v1, 2020.
[38] H. Johannes, P. Jeanny, R. Sebastian, P. Helmut, and L. Georg, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," European Radiology Experimental, vol. 4, no. 1, 2020.
[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, Int. Journal of Computer Vision (IJCV) 115 (3) (2015) 211–252.
[40] S.A.A. Ahmed, B. Yanikoglu, C. Zor, M. Awais, J. Kittler, Skin lesion diagnosis with imbalanced ECOC ensembles, in: Int. Conf. on Machine Learning, Optimization, and Data Science, Springer, 2020.
[41] W. Lee, J. Na, G. Kim, Multi-task self-supervised object detection via recycling of bounding box annotations, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 4984–4993.
[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929.
[43] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in Proceedings of the IEEE Int. Conf. on Computer Vision (ICCV), 2017, pp. 5209–5217.
[44] S.A. Aly and B. Yanikoglu, "Multi-label networks for face attributes classification," in 2018 IEEE Int. Conf. on Multimedia & Expo Workshops (ICMEW). IEEE, 2018, pp. 1–6.
[45] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.
[46] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.
[47] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems 25 (2012) 1097–1105.
[48] S.A.A. Ahmed, B. Yanikoglu, Within-network ensemble for face attributes classification, in: Int. Conf. on Image Analysis and Processing, Springer, 2019, pp. 466–476.
[49] E.B. Wilson, Probable inference, the law of succession, and statistical inference, Journal of the American Statistical Association 22 (158) (1927) 209–212.
[50] S. Heidarian, P. Afshar, N. Enshaei, F. Naderkhani, A. Oikonomou, S.F. Atashzar, F.B. Fard, K. Samimi, K.N. Plataniotis, A. Mohammadi, and M.J. Rafiee, "COVID-FACT: A fully-automated capsule network-based framework for identification of COVID-19 cases from chest CT scans," arXiv 2010.16041, 2020.
[51] S. Heidarian, P. Afshar, A. Mohammadi, M.J. Rafiee, A. Oikonomou, K.N. Plataniotis, and F. Naderkhani, "CT-CAPS: Feature extraction-based automated framework for covid-19 disease identification from chest CT scans using capsule networks," arXiv 2010.16043, 2020.
[52] M. Dialameh, A. Hamzeh, H. Rahmani, A.R. Radmard, and S. Dialameh, "Screening COVID-19 based on CT/CXR images & building a publicly available CT-scan dataset of COVID-19," arXiv 2012.14204, 2020.



Sara Atito Ali Ahmed received her Bsc. in Computer Science from Ain Shams University, Egypt (2011). Her Msc. degree was a collaboration between Nile University, Egypt and TU Berlin, Germany (2014), working on vehicle detection and tracking in crowded scenes and sensitivity analysis of deep neural networks. She received her Ph.D. degree in Computer Science from Sabancı University, Turkey (2021), with a thesis on deep learning ensembles for image understanding. Currently, she is a Research Fellow in the CVSSP group in University of Surrey, UK, working on developing self-supervised learning techniques for object detection and classification.

Mehmet Can Yavuz is a PhD student at Sabanci University, Istanbul. He received his BS degrees in Electrics-Electronic Engineering and Physics (double major) from Işık University in 2010 and MS degree from the Boğaziçi University Physics Department in 2016. His research interests are applications of machine learning in biometrics and biomedical engineering. He has worked as a Research Assistant in various European Union ERC, TÜBİTAK and scientific research projects.

Mehmet Umut Sen completed his PhD degree in Electronics Engineering Department of Sabancı University, Istanbul. He holds MSc and BSc degrees also from the same department, received in 2009 and 2011 respectively. His research interests include text categorization, machine learning and speech processing.

Fatih Gulsen received the B.S. degree from University of Eskisehir Osmangazi, Eskisehir, Turkey, 2001 and the M. S. degree in Radiology Istanbul University-Cerrahpaşa, Cerrahpaşa Medical Faculty, Istanbul, Turkey, in 2006. Now, he is a Professor in the Interventional Radiology Department, Istanbul University-Cerrahpaşa, Cerrahpaşa Medical Faculty. He is a European Board Certified interventional radiologist. His research interests focus on interventional radiology, angiographic techniques, interventional oncology, medical imaging, medical image processing, and disease diagnosis.

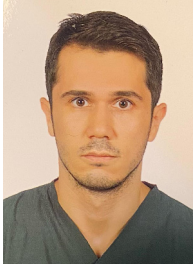
Onur Tutar is an Associate Professor of Radiology specialist at Istanbul University-Cerrahpaşa, Cerrahpaşa Medical Faculty, Istanbul, Turkey. He earned his medical degree at Ankara University, in 2003 and received his radiology residency degree at Istanbul University-Cerrahpaşa, Cerrahpaşa Medical Faculty in 2009. His research interests are abdominal radiology and imaging informatics.

Dr. Bora Korkmazer is a Radiologist, at Istanbul University Cerrahpaşa Medical Faculty, Istanbul, Turkey. He graduated from Marmara University Faculty of Medicine in 2009. He is a European Board Certified interventional neuroradiologist. He is interested in interventional neuroradiology, angiographic techniques, angioplasty, stenting and new devices.

Dr Cesur Samancı is an associate professor of Radiology, at Istanbul University Cerrahpaşa Mediacal Faculty, Istanbul, Turkey. He graduated from Ankara University Faculty of Medicine in 2011.He is a European Board-Certified interventional radiologist. He is interested in interventional radiology, interventional oncological procedures, angiographic techniques, angioplasty, stenting and new devices. He received bioistatistic courses.

Sabri Şirolu works as a radiology specialist at the University of Health Sciences, Sisli Hamidiye Etfal Education and Research Hospital, Istanbul, Turkey. He earned his medical degree at Hacettepe University, Ankara, in 2014 and received his radiology residency degree at Istanbul University-Cerrahpaşa, Cerrahpasa Medical Faculty in 2021. His research interests are abdominal radiology and imaging informatics.

Rauf Hamid received the B.S. degree from University of Hacettepe, Ankara, Turkey, 2013 and is a 5th year residency student at Cerrahpaşa Faculty of Medicine, Department of Radiology. His research interests focus on interventional radiology, angiographic techniques, interventional oncology, medical imaging, medical image processing, and disease diagnosis.



Ali Ergun Eryürekli is a 5th year residency student at Cerrahpaşa Faculty of Medicine, Department of Radiology. He graduated from Cerrahpaşa Faculty of Medicine in 2015. His interests are in artificial intelligence, machine learning and interventional radiology.

Toghrul Mammadov is a 5th year residency student at Cerrahpaşa Faculty of Medicine, Department of Radiology. He graduated from Istanbul University Faculty of Medicine in 2016. His interests are in artificial intelligence, machine learning, chest radiology and cardiovascular radiology.

Berrin Yanikoglu is Professor of Computer Science and Director of the Center of Excellence in Data Analytics (VERIM), at Sabanci University, Istanbul, Turkey. She received a double major in Computer Science and Mathematics from Bogazici University, Turkey in 1988 and her Ph.D. degree in Computer Science from Dartmouth College, USA in 1993. Prof. Yanikoglu worked at Rockefeller University, Xerox Imaging Systems and IBM Almaden Research Center, before joining Sabanci University in 2000. Her research interests lie in machine learning with applications to image/video understanding, in particular medical image understanding, biometric verification and privacy, and handwriting recognition. She is an Editor for the Turkish Journal of Electrical Engineering and Computer Science.