# Retraction Notice

The Editor-in-Chief and the publisher have retracted this article, which was submitted as part of a guest-edited special section. An investigation uncovered evidence of compromised peer review and determined the paper is unrelated to the special section. The Editor and publisher no longer have confidence in the results and conclusions of the article.

MFS disagrees with the retraction. AB and MS either did not respond or could not be reached.

# COVID-19 detection using machine learning: a large scale assessment of x-ray and CT image datasets

**Md. Fahimuzzman Sohan**[a,*] **Anas Basalamah**[b] **and Md. Solaiman**[a]
[a]Daffodil International University, Department of Software Engineering, Dhaka, Bangladesh
[b]Umm Al-Qura University, Department of Computer Engineering, Makkah, Saudi Arabia

**Abstract.** Millions of people are infected by the coronavirus disease 2019 (COVID-19) around the world. Within three months of its first report, it rapidly spread worldwide with thousands of deaths. Since that time, not only underdeveloped and developing countries, but also developed countries have suffered from insufficient medical resources and diagnoses. In this circumstance, researchers from medical and engineering fields have tried to develop automatic COVID-19 detection toolkits using machine learning (ML) techniques. The dataset is the fundamental element of any detection tool; therefore, most of the ML-based COVID-19 detection research was conducted used chest x-ray and computed tomography (CT) image datasets. In our study, we collected a series of publicly available unique COVID-19 x-ray and CT image datasets, then assessed and compared their performances using our proposed 22 layer convolutional neural network model along with ResNet-18 and VGG16. We investigated eight individual datasets known as Twitter, SIRM x-ray, COVID-19 Image Repository, EURORAD, BMICV, SIRM CT, COVID-CT, and SARS-CoV-2 CT. Our model obtained classification accuracy of 91%, 81%, 59%, 98%, 58%, 79%, and 97%, respectively. Our proposed model obtained the highest classification accuracy using four datasets (Twitter, COVID-19 Image Repository, COVID-CT, and SARS-CoV-2 CT). Similarly, ResNet-18 only utilized three (EURORAD, BMICV, and SIRM CT), whereas VGG16 only utilized the SIRM x-ray dataset. Results of this investigation indicate a significant comparison chart among the performance of the datasets. Indeed, our study is a large-scale assessment of existing COVID-19 x-ray and CT image datasets. And to the best of our knowledge, this is the first performance comparison study that includes all publicly available COVID-19 datasets. © *2022 SPIE and IS&T* [DOI: 10.1117/1.JEI.31.4.041212]

**Keywords:** COVID-19 imaging; deep learning; machine learning; COVID-19 detection; datasets; review.

Paper 210581SS received Sep. 1, 2021; accepted for publication Feb. 14, 2022; published online Mar. 2, 2022.

Retracted

## 1 Introduction

The severe acute respiratory syndrome coronavirus 2, also known as SARS-CoV-2 or novel coronavirus 2019 (COVID-19), was first reported in December 2019 in the Hubei Province of China.[1] This deadly virus attacked every country in the world within a few months of first appearing in Wuhan, China. At the time of writing, over 4.2 million deaths and 200 million confirmed cases have been reported worldwide by the World Health Organization due to the virus epidemic.[2] The impact of the COVID-19 outbreak is destructive in a comprehensive manner; not only the healthcare system, but also the economy, education, social sectors are damaged. At the initial stage of the pandemic, the most challenging step was rapidly identifying the infected individuals and isolating them to delay the spread of the pandemic.[3] Initially, the COVID-19 diagnosis toolkit utilized was reverse transcription polymerase chain reaction (RT-PCR), though later on, several antigens and antibody testing tool-kits were approved in many countries. Regardless, the RT-PCR testing process is expensive and takes several hours to yield results, whereas rapid testing kits have competitively lower accuracy.[4] Consequently, machine learning (ML) researchers along with medical scientists and radiologists have been

*Address all correspondence to Md. Fahimuzzman Sohan, fahimsohan2@gmail.com

1017-9909/2022/$28.00© 2022 SPIE and IS&T

data mining these diagnosis methods along with radiological images, such as x-ray and CT images, to find alternative ways to detect COVID-19 rapidly and accurately. Deep learning methods, more specifically convolutional neural network (CNN), is widely used to analyze the images.[5–7] Generally, a CNN model extracts deep and high-level features to train or learn from the raw image datasets. This learning model is used to build automatic tools that can identify positive COVID-19 cases. ML-based COVID-19 detection is automatic, easy to use, and when implemented in clinical settings only takes a few seconds to give the result.[8] Hence, x-ray or CT image data can significantly aid in COVID-19 detection.

We conducted this study following our initial investigation, made available in arXiv.[9] We initially investigated ∼100 articles to find trends in dataset usage for COVID-19 detection research. Santa Cruz et al.[10] presented a systematic literature review (SLR) article, where they presented nine datasets. The datasets were checked by CHARMS quality checklist and they used BIAS tool to conduct bias analysis. There are several other publications available on COVID-19 datasets,[10–14] however, the articles are mostly SLR and only represent the information from the datasets. In addition, large-scale assessments and competitive performance analyses among the existing datasets are lacking.

In this study, we focused on the open-access COVID-19 x-ray and CT image datasets, which are available since the beginning of the pandemic. First, we considered our previous study,[9] then we collected the information from COVID-19 datasets in Ref. 10 and other data repositories. We found a noticeable practice of using data for COVID-19 detection research. We termed "data blending," which indicates authors collected COVID-19 positive x-ray and CT images from various sources and prepared a blended dataset for their investigation. As an example, the Cohen JP dataset[15] was prepared by collecting COVID-19 positive images from four different sources.

However, we have used eight COVID-19 positive image datasets in our investigation and taken them from eight unique sources. The motive of using unique COVID-19 positive datasets is performance evaluation and quality checking of an individual dataset. To evaluate the performance of collected datasets, we used our proposed 22 layers CNN model and two other CNN models (ResNet-18 and VGG16), widely considered in image classification research. ResNet-18 is an 18-layer CNN model, mainly presented by He et al.[16] in 2015, whereas VGG16 is a 16-layer CNN model proposed by Simonyan and Zisserman[17] in 2014. However, we have preprocessed each dataset by resizing the images and applying a data normalization technique. We used training data to train our CNN model and test data to validate the learning of the model. The key contributions of this study are listed below.

- Followed by our initial work, we investigated around 100 articles to explore the trend of using datasets in COVID-19 detection research.
- In this study, we present the eight most commonly used public COVID-19 positive image datasets.
- We also introduce a deep learning-based COVID-19 detection model to extract features from the image datasets.
- We trained the model using the eight datasets to individually detect two classes: COVID-19 infected and normal. To ensure fair assessment, we used a uniform detection model and non-COVID image data for each dataset.
- We compared the performance of the eight datasets by our proposed CNN model and other two well-known CNN architectures: ResNet-18 and VGG16.

The remaining part of this paper contains: Sec. 2 provides a description of the previous investigation on the COVID-19 detection using ML. Then Sec. 3 addresses the methodology of this study, which includes our approach for considering the datasets. Section 4 describes the results of this study. Finally, Sec. 5 provides a conclusion.

## 2 Literature Review

As a global crisis, COVID-19 research using ML techniques motivated the development of automated detection tools. As a result, hundreds of articles have been published on COVID-19 detection by the classification of medical image data[18] within just one year. In this section,

Retracted

**Table 1** Summary of used dataset and performance of recent researches related to COVID-19 detection using deep learning approaches.

| References | Dataset | Performance |
|---|---|---|
| 19 | COVID*x* dataset (358 COVID-19, 8066 normal, and 5538 pneumonia) | Accuracy 93.3% |
| 20 | Cohen JP dataset and ChestX-ray8 (127 COVID-19, 500 normal, and 500 pneumonia) | Accuracy of two-class 98.08%<br>Accuracy of three-class 87.02% |
| 21 | Cohen JP dataset, RSNA, radiopaedia, SIRM, and Mendeley dataset (dataset 1 = 224 COVID-19, 504 normal, and 700 pneumonia, dataset 2 = 224 COVID-19, 504 normal, and 714 pneumonia) | Accuracy 96.78% |
| 22 | Cohen JP dataset and JSRT(105 COVID-19, 80 normal, and 11 SARS) | Accuracy of 93.1% |
| 23 | Cohen JP dataset, Mendeley dataset, chest x-ray images pneumonia, snapshots (dataset 1 = 453 COVID-19 and 497 non-COVID, dataset 2 = 71 COVID-19 and 7 non-COVID) | Accuracy of first dataset 91.16%<br>Accuracy of second dataset 97.44% |
| 24 | COVID-19 radiology database (219 COVID-19, 1341 normal, and 1345 pneumonia) | Accuracy of 98.97% |
| 25 | COVID-19 radiology database (219 COVID-19, 1341 normal, and 1345 pneumonia) | Accuracy of 98.70% |
| 26 | Cohen JP dataset (69 COVID-19, 79 normal, and 158 pneumonia) | Accuracy of two-class 100% |
| 27 | Cohen JP and RSNA dataset (180 COVID-19, 8851 normal, and 6054 pneumonia) | Average accuracy of 99.50% |
| 28 | Cohen JP and others (122 COVID-19, 150 normal, and 300 pneumonia) | Accuracy of two-class 100%<br>Accuracy of four-class 76% |
| 29 | Cohen JP, Qatar-Dhaka COVID-19 data (dataset 1 = 200 COVID-19 and 1675 non-COVID, dataset 2 = 219 COVID-19 and 1341 non-COVID) | Accuracy of dataset 98.7%<br>Accuracy of second dataset 99.6% |
| 30 | Private dataset (5634 COVID, 6000 normal, and 5000 others) | Highest accuracy 93.48% |
| 31 | Qatar-Dhaka COVID-19 data (219 COVID-19, 1341 normal, and 1345 viral pneumonia) | Average accuracy of 97.20% |
| 32 | COVID CT dataset, Cohen JP, and SARS-COV-2 (dataset 1 = 349 COVID-19 and 397 normal; dataset 2 = 125 COVID-19 and 1000 normal; dataset 3 = 1252 COVID-19 and 1230 normal) | Accuracy of first dataset 89.41%<br>Accuracy of second dataset 99.02%<br>Accuracy of third dataset 98.11% |
| 33 | Cohen JP, SIRM database, radiopaedia, and others (2300 COVID-19 and 2300 normal) | Highest accuracy 98.91% |

Retracted

we review several published COVID-19 research articles using ML in an informative manner. Additionally, Table 1 summarizes a list of previously published relevant articles with the information from the used datasets and their performance.

A dataset named COVID*x* and a deep CNN learning model, COVID-Net, was introduced by Wang et al.[19] to diagnose COVID-19 patients from chest x-ray images. The x-ray images were collected from three perspectives: COVID positive (358), normal (8066), and pneumonia viral/bacterial (5538). Their proposed COVIDx dataset is a combined and modified form of five publicly available databases, and they updated the data repository regularly in GitHub.

Their proposed COVID-Net model achieved an overall accuracy of 92.4%. Ozturk et al.[20] proposed a deep learning model that was developed under both binary and multiclass classification. They have used two sources to collect image data for the classification model where COVID-19 positive x-ray images were collected from the Cohen JP data set. Additionally, normal and pneumonia images from the ChestX-ray8 database were used. The dataset is comprised of 125 COVID-19 positive, 500 normal, and 500 pneumonia chest x-ray images. Their proposed CNN model achieved 98.08% and 87.02% accuracy according to binary and multiclass cases, respectively. Tuncer et al.[21] also collected data from Cohen JP and ChestX-ray8 datasets for their proposed three-class CNN-based transfer learning models. Their data set contains three types of x-ray images: 224 COVID-19 positives, 504 normal, and 700 bacterial pneumonia images. One of their pretrained models, MobileNetV2, achieved a height accuracy of 96.78% and an overall accuracy of 98.75%.

Khan et al.[14] proposed a CNN-based COVID-19 detection model by x-ray and CT scan images from Cohen JP and chest x-ray images pneumonia datasets. Their experiment was conducted by binary (COVID-19/normal), three-class (normal/COVID-19/Pneumonia), and four-class (viral pneumonia/COVID-19/bacterial pneumonia/normal) classifications. The detection model used the Xception model and was pretrained on the ImageNet dataset. The results show that the overall accuracy of binary, three-class, and four-class were 99%, 95%, and 89.6%, respectively. Maghdid et al.[34] also proposed a CNN-based deep learning and AlexNet network-based transfer learning model for COVID-19 detection using x-ray and CT images. Their pretrained network achieved an accuracy of 98% and modified CNN achieved 94.1% accuracy. In Ref. 26, Loey et al. collected x-ray image data from four perspectives: COVID-19 positive, normal, bacterial pneumonia, and viral pneumonia images. They used three pretrained models: AlexNet, Googlenet, and Restnet18. Googlenet achieved an accuracy of 80.6%, Alexnet 85.2%, and Googlenet 100% in terms of four, three, and binary class classifications, respectively, among the three deep transfer models. Rahimzadeh and Attar[27] focused on the unbalanced dataset to create a better learning environment. They collected image data from two different sources with three classes (COVID-19 positive, normal, and pneumonia) They also proposed a neural network using Xception and ResNet50V2 networks. The model achieved 91.4% average accuracy according to three class classification.

Most of the articles in Table 1 used the "Cohen JP dataset"[15] to collect COVID-19 positive samples and train their detection modes. This dataset is widely used and is described as the first COVID-19 image dataset. This open-access dataset contains chest x-ray and CT images collected from different hospitals in different countries. All data are released under the GitHub repository and updated continuously by the authority. In addition, the "chest x-ray images pneumonia" dataset is also used in many of the articles. This dataset includes normal and pneumonia (bacterial and viral) images and is used as a source of non-COVID-19 samples. Moreover, most of the articles used CNN architecture for deep learning[35] and of them many used transfer learning techniques[36] for diagnosing COVID-19 patients.

## 3 Methodology

### 3.1 *Dataset*

Nowadays, diagnosis using medical images is a popular research field and it has implementation in many areas; for example, implemented in pneumonia detection, cancer detection and prediction, tumor diagnosis, and more.[37–39] Medical image data played a key role in COVID-19 diagnosis using ML methods. As usual, such respective detection models have used x-ray and CT images of COVID-19 infected patients, and two-class, three-class, and four-class classifications have been used to classify pneumonia and normal images. Therefore, we found three flows for using medical image data in COVID-19 detection research: such as COVID-19 positive, pneumonia infected, and normal images. Conventional COVID-19 datasets are prepared by aggregating primary sources[40–44] and the majority of research was driven using compilations of various sources' data.[10] On the other hand conversely, several non-COVID datasets have been widely used in different areas of radiographic image analysis.[45,46]

Retracted

**Table 2** Details of used datasets in this investigation for two-class classification.

| Dataset name | COVID | Normal | Total |
|---|---|---|---|
| Twitter | 114 | 342 | 456 |
| SIRM x-ray | 119 | 357 | 476 |
| COVID-19 image repository | 243 | 729 | 972 |
| EURORAD | 258 | 774 | 1032 |
| BMICV | 450 | 1350 | 1800 |
| SIRM CT | 270 | 810 | 1080 |
| COVID-CT | 349 | 047 | 1396 |
| SARS-CoV-2 CT | 1230 | 1230 | 2460 |

In this study, we mainly worked with the primary sources of COVID-19 x-ray and CT image datasets that are available online. In continuation of this, we summarized our findings with eight COVID-19 image datasets (five x-ray and three CT) after investigating hundreds of articles and our previous work a thorough literature review. Additionally, we used the "chest x-ray images of pneumonia" dataset as a source of non-COVID data for our detection model. We considered two features in terms of collecting the raw datasets: the dataset had to have more than 100 images[47] was excluded since it only has 57 images and was openly accessible. Table 2 shows the statistics of the eight datasets, including dataset name and amount of data. Figure 1 shows the visual representation of all the data samples. Note that we used an equal number of non-COVID data samples in each COVID-19 dataset.

### 3.1.1 *COVID-19 positive datasets*

*Twitter*. These image data are available on Twitter, where a cardiothoracic radiologist from Spain has shared high-quality COVID-19 positive subjects. A total 128 images of 50 cases were uploaded in 50 different tweets from the Twitter account. We considered 114 images for our investigation after manually checking the quality of the images (link to data for Ref. 48).

*SIRM*. Italian Society of Medical and Interventional Radiology (SIRM) posted images of 115 COVID-19 positive cases on their website. The repository includes more than 600 x-ray images and CT scans. We collected the images and prepared two individual datasets: SIRM x-ray

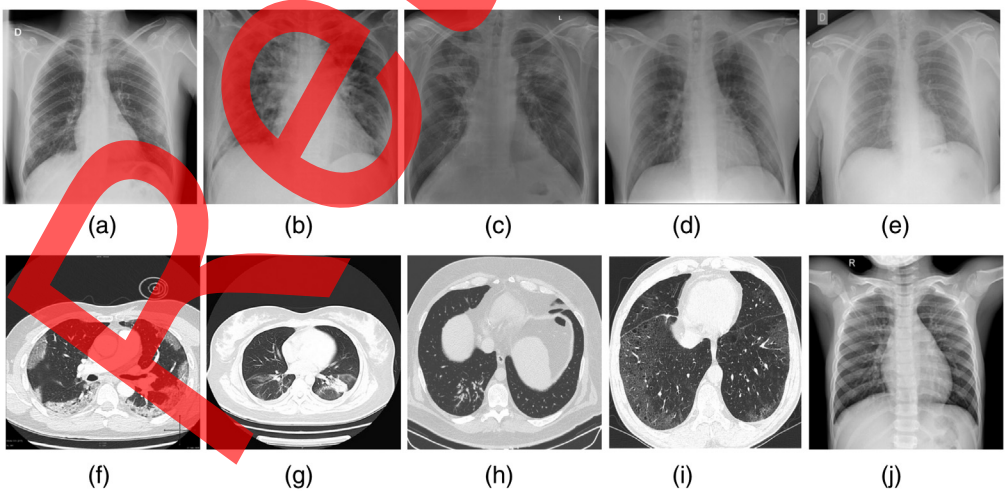


**Fig. 1** Samples of the used images with respective dataset: (a) Twitter, (b) SIRM x-ray, (c) COVID-19 Image Repository, (d) EURORAD, (e) BMICV, (f) SIRM CT, (g) COVID-CT, (h) SARS-CoV-2 CT, (i) normal CT, and (j) normal x-ray.

Retracted

includes the x-ray images and SIRM CT includes the CT images. We manually selected the x-ray and CT images. In this investigation, we have used 114 images for the SIRM x-ray dataset and 270 images for SIRM CT images (link to data for Ref. 49).

*COVID-19 Image Repository*. These COVID-19 positive images were collected from Hannover Medical School, Hannover, Germany. The dataset has 243 x-ray images (link to data for Ref. 50).

*EURORAD*. It is the learning platform of the European Society of Radiology for radiologists, radiology residents, and students. Since the COVID-19 outbreak arose, the platform has been publishing chest x-rays and CT scans of COVID-19 patients. We collected 258 of these COVID-19 positive x-ray images (link to data for Ref. 51).

*BMICV*. These data were originally collected from the BMICV website by Chowdhury et al.,[52] we used a portion of these data from there. The BMICV dataset is a large-scale dataset from Spain; the dataset includes chest x-ray CXR (CR, DX) and CT images of COVID-19 positive patients. The main archive contains 7377 CR, 9463 DX, and 6687 CT images. The article[52] used 2473 CXR images, and for our study, we randomly selected subset of these images (450) (link to data for Ref. 53).

*COVID-CT*. This global dataset is divided into two parts: one for COVID-19 positive known as "COVID-CT" and non-COVID-19 known as "non-COVID-19 CT." This dataset has been prepared by combining data from different COVID-19-related papers. The contributors collected 760 COVID-19 research preprints from two different platforms. They did an interesting job, then they extracted the CT images from the PDF files. Extracted images were manually preprocessed and their metadata collected metadata of each image. The positive COVID-19 part of the dataset contains 349 images from 216 patients, and we were used as COVID-19 data source [link to data (https://github.com/UCSD-AI4H/COVID-CT)].

*SARS-CoV-2 CT*. A research team built a public COVID-19 detection CT dataset with 1252 COVID-19 positive and 1230 non-COVID-19 CT images.[54] They collected real patient CT images in hospitals from Sao Paulo, Brazil. In our study, we only used 1230 COVID-19 positive images from the full volume (link to data for Ref. 55).

### 3.1.2 *Non-COVID dataset*

*Normal CT images*. We already mentioned the SARS-CoV-2 CT dataset has 1230 non-COVID CT images that we used along with our two CT image datasets as non-COVID data sources.

*Normal x-ray images*. This chest x-ray image dataset is known as chest x-ray images pneumonia and is a part of the Mendeley dataset.[46] The authority prepared the dataset by screening and checking raw images to ensure quality at Guangzhou Women and Children's Medical Center, Guangzhou, China. It is a two-class image dataset, with a total of 5856 images, 4273 pneumonia images, and 1583 normal images. In our study, we used this dataset as a non-COVID-19 data source. Additionally, for each experiment of a COVID-19 x-ray image dataset, we used the same number of normal images from this pneumonia dataset (link to data for Ref. 56).

## 3.2 *CNN Architecture*

CNN models are commonly used in COVID-19 imaging research articles; with some using simple CNN architectures,[20,57] others using CNN architectures along with transfer learning,[21,58] and multiclassifier models.[33] In our study, we used a 22-layer typical CCN model including a convolutional layer, pooling layer, flatten layer, dense layer, and normalization layer. As Fig. 2 shows, the model is separated into five different blocks containing three layers each, blocks 2, 4, and 5 each has a fourth layer (the dropout layer). An elaborate description of the blocks and their layers is given below.

- *Convolutional layer*. The learning process of the model begins with the convolutional layer. The convolution layer is the most important part of a CNN model as it extracts features for learning using different filters, kernels, and convolution layers.[57] In our model, we used five Conv2D layers and all of their filter matrices were $3 \times 3$. This feature extraction layer extracts feature(s) from input images and produces a feature map. For our study, we used a $244 \times 244$ size image and this image size was multiplied by the $3 \times 3$ filter

Retracted

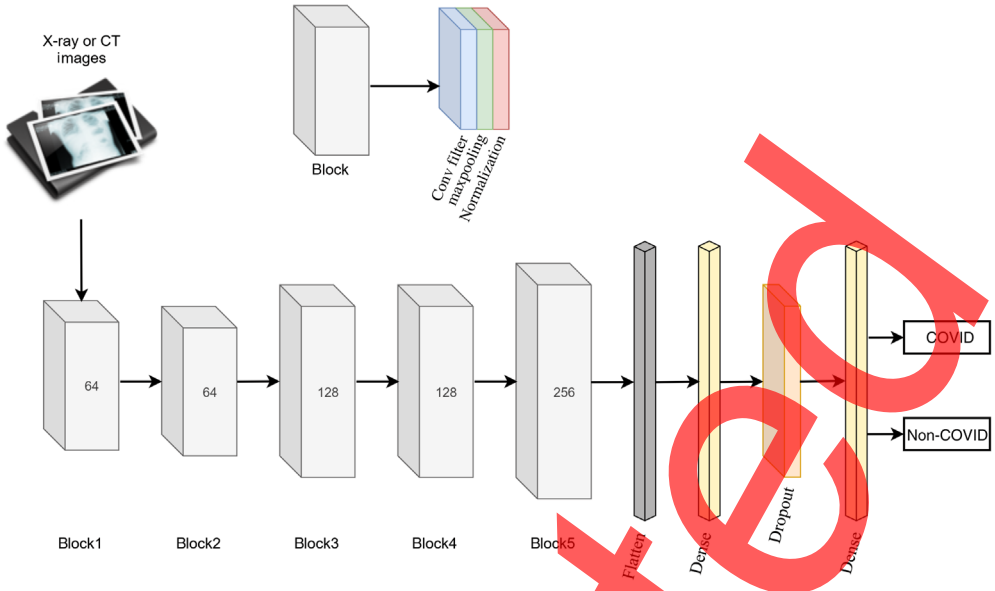


**Fig. 2** CNN architecture of this study.

matrix ($244 \times 244 \times 3$). We used a stride size of 1, which means the filter moves from left to right one pixel over an image. So the ultimate goal of this layer is to extract features from the input images and produce a learning map. In addition, we used the ReLU activation function in the convolution layer to transfer output to the next layer.[31]

- *Pooling layer.* This layer is commonly used in the CNN models; it reduces the size of the feature map and stores only necessary information. It is obvious that after reducing the size of the image and quantity of parameters, the layer helps increase the speed of the computation, reduce the memory required, and control overfitting.[33,58] Max pooling and average pooling are two commonly used pooling layers, and our study consisted of five max-pooling layers. We have used a $2 \times 2$ size filter and a stride size of 2 in each max-pooling layer.
- *Normalization layer.* We have used batch normalization to standardize the input size of our model. In addition, this layer reduces training time and makes the model more stable. In our model, we used one normalization layer for each Conv2D layer, five layers in total.
- *Dropout layer.* The dropout layer is used to deal with the problem of overfitting the model. We used four dropout layers in the model with a threshold of 0.2.
- *Flatten layer.* This layer is important when we use Conv2D and pooling layers. This layer converts the 2D convolutions into a one-dimensional array. Flatten layer summarizes the previous layers and sends them to the next layer for further processing.
- *Dense layer.* This layer performs like a neural network, where neurons of the dense layer are connected to the output of the previous layer's neurons. The dense layer is also known as the connected layer because this process connects the neurons from two layers. We have used two dense layers in our CNN model; one is 128 and the other is 1 unit in size with ReLU and sigmoid activation function, respectively.

## 3.3 *Training Model*

In Fig. 3, we present the workflow of our proposed model. Our study started with collecting datasets as discussed in Sec. 3.1, then we selected and prepared them to execute in our model. Then the data were preprocessed, where we resized the data shape, used data normalization, and data labeling techniques. We collected data from various repositories where the size of the data was highly variable. Moreover, unexpected text and symbols were on the image data, so it was important to remove these unnecessary elements from the objects.[31] As a result of this unequal distribution of image sizes, we resized all images with a $244 \times 244 \times 1$ pixel format. Thus we used data normalization to prevent overfitting of our model.[59] This study conducted two-class classification, labeled as 0 for COVID-19 positive and 1 for normal images. The next step was

Retracted

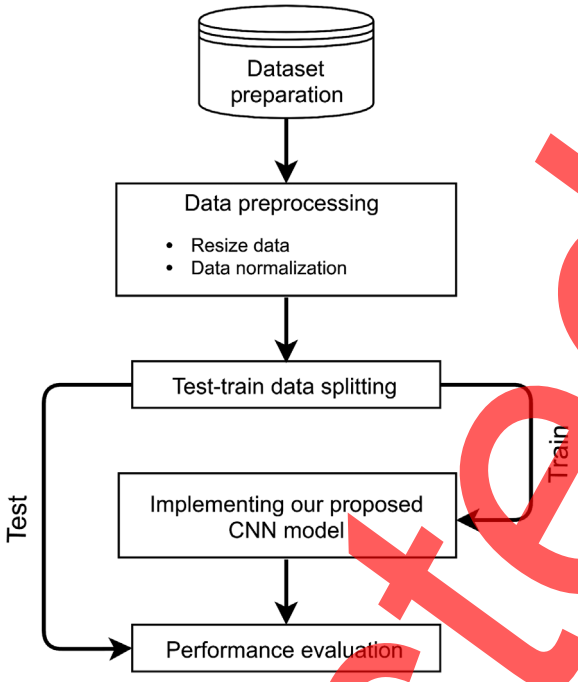


**Fig. 3** Experiment procedure of our COVID-19 detection model.

the test–train splitting of the datasets. All datasets used in this study were divided into an 80% training and 20% test ratio. Training data travel to the CNN model for learning the model, whereas the test data were used to validate the model.

We have used several hyperparameters in our experiment and presented them in Table 3. Then we executed 22 layers of CNN architecture, which is the key part of the proposed model. In Fig. 2, the first 18 layers are divided into five blocks; the first two blocks have 64 filters, the third and fourth blocks have 128, and the last one has 256. We used Conv2D, MaxPool2D, and batch normalization in each block, and a dropout layer used with 64, 128, and 256 filters. Then we used a flatten, a 128-unit dense layer, a dropout layer, and lastly, a 1-unit dense layer. Table 4 summarizes all of the 22 layers and their parameters. After implementing the CNN layers, we trained our model for 10, 20, and 100 epochs. The batch size of our experiment was 32 with a learning rate of 0.00001. Performance measures are a fundamental function of the ML model,

**Table 3** Values of used hypermeters in this study.

| Hypermeters | Values |
|---|---|
| Batch size | 32 |
| Train-test ratio | 80:20 |
| Zoom range | 0.2 |
| Width shift range | 0.1 |
| Height shift range | 0.1 |
| Optimizer | Adam |
| Rotation range | 30 |
| Input size | $244 \times 244$ pixels |
| Dropout | 0.2 |
| Activation | ReLU/sigmoid |
| Epochs | 100 |

Retracted

**Table 4** Used 22 layers CNN architecture with parameters.

| Layer number | Layer type | Output shape | Parameter |
|---|---|---|---|
| 1 | conv2d (Conv2D) | (None, 224, 224, 64) | 640 |
| 2 | batch_normalization (BatchNormalization) | (None, 224, 224, 64) | 256 |
| 3 | max_pooling2d (MaxPooling2D) | (None, 224, 224, 64) | 0 |
| 4 | conv2d_1 (Conv2D) | (None, 224, 224, 64) | 36,928 |
| 5 | dropout (Dropout) | (None, 224, 224, 64) | 0 |
| 6 | batch_normalization_1 (BatchNormalization) | (None, 224, 224, 64) | 256 |
| 7 | max_pooling2d_1 (MaxPooling2d) | (None, 112, 112, 64) | 0 |
| 8 | conv2d_2 (Conv2D) | (None, 112, 112,128) | 73,856 |
| 9 | batch_normalization_2 (BatchNormalization) | None, 112, 112,128) | 512 |
| 10 | max_pooling2d_2 (MaxPooling2d) | (None, 56, 56, 128) | 0 |
| 11 | conv2d_3 (Conv2D) | (None, 56, 56, 128) | 147,584 |
| 12 | dropout_1 (Dropout) | (None, 56, 56, 128) | 0 |
| 13 | batch_normalization_3 (BatchNormalization) | (None, 56, 56, 128) | 512 |
| 14 | max_pooling2d_3 (MaxPooling2d) | (None, 28, 28, 128) | 0 |
| 15 | conv2d_4 (Conv2D) | (None, 28, 28, 256) | 295,168 |
| 16 | dropout_2 (Dropout) | (None, 28, 28, 256) | 0 |
| 17 | batch_normalization_4 (BatchNormalization) | (None, 28, 28, 256) | 1024 |
| 18 | max_pooling2d_4 (MaxPooling2d) | (None, 14, 14, 256) | 0 |
| 19 | flatten (Flatten) | (None, 50176) | 0 |
| 20 | dense (Dense) | (None, 128) | 6,422,656 |
| 21 | dropout_3 (Dropout) | (None, 128) | 0 |
| 22 | dense_1 (Dense) | (None, 1) | 129 |

commonly used in relevant fields.[60,61] In our experiment, four performance measures have been used to assess the performances of the learning model. The four performance measures are accuracy, precision, recall, and $F1$-score. Performance measures are calculated from the confusion matrix, which includes four different combinations of actual versus predicted values: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

We implemented the whole operation of our proposed model in Google Colab Pro, it is a Jupyter notebook-based cloud service that provides excellent and uninterrupted service for ML tasks.[33]

## 4 Results

We experimented with three CNN models: our proposed 22 layer model, ResNet-18, and VGG16. In this section, we present the results, including the respective confusion matrix (Table 5) and four performance measures. First, we review the results and performance of our proposed CNN model, next we compare our results with the performance of two notable CNN models. Finally, we present a list of the eight datasets according to their performance achieved by the three CNN models as the mainspring of this study.

Retracted

**Table 5** Confusion matrix for testing accuracy of eight datasets.

| Dataset name | Confusion matrix ([TP, FP] [FN, TN]) | | |
|---|---|---|---|
| | Presented model | ResNet-18 | VGG16 |
| Twitter | [23, 0] [4, 19] | [20, 4] [13, 9] | [22, 1] [5, 17] |
| SIRM x-ray | [24, 0] [9, 15] | [21, 3] [0, 18] | [21, 3] [5, 19] |
| COVID-19 Image Repository | [49, 0] [2, 47] | [47, 2] [2, 47] | [48, 1] [12, 37] |
| EURORAD | [42, 14] [30, 18] | [43, 9] [10, 42] | [51, 1] [24, 28] |
| BMICV | [90, 0] [ 6, 84] | [90, 0] [1, 89] | [90, 0] [4, 86] |
| SIRM CT | [53, 1] [43, 11] | [31, 23] [17, 37] | [29, 25] [20, 34] |
| COVID-CT | [64, 6] [24, 46] | [51, 19] [15, 55] | [43, 27] [20, 50] |
| SARS-CoV-2 CT | [201, 42] [80, 169] | [53, 27] [32, 48] | [63, 17] [46, 34] |

## 4.1 *Performance of Proposed Model*

*Twitter*. The Twitter dataset is the smallest of the eight datasets used in this study. This dataset has 228 x-ray images (114 COVID-19 positive and 114 normal images). We trained our 22 layer CNN model with 182 images (80%) and tested the trained model with 46 images (20%). Table 5 shows the confusion matrix, including the Twitter dataset. In this confusion matrix, the total number of test data is 46, correctly detected samples are 42, and wrongly predicted samples are 4, which indicates the accuracy of the model is 91%.

*SIRM x-ray*. This dataset has 238 x-ray images (119 COVID-19 positives and 119 normal images), and the number of test data is 48. Under our CNN model achieved 81% accuracy with this dataset; in Table 5, 39 samples were detected correctly and 9 predicted incorrectly.

*COVID-19 Image Repository*. A total of 486 x-ray images were used in this dataset including 243 COVID-19 positive and 243 normal images. The dataset achieved 98% detection accuracy; among the 98 test data, 96 were detected accurately, and 2 samples were detected incorrectly.

*EURORAD*. EURORAD dataset has 516 images, from which 104 images were used to test the model. According to our proposed model, the accuracy of this dataset was lowest, 58%. The correctly and incorrectly identified images were 60 and 44, respectively.

*BMICV*. It is the second largest dataset used in this study. The dataset has 900 x-ray images, 450 COVID-19 positive and 450 normal. This dataset also produced a high accuracy of 97%. Accounting to Table 5, the confusion matrix of the proposed model indicates that 174 images are classified correctly, and only 6 images are classified incorrectly.

*SIRM CT*. SIRM repository has x-ray and CT images of COVID-19 positive patients, and we considered 270 CT images for our dataset. The dataset achieved the second-lowest accuracy of only 59% among the eight participants. Table 5 shows that 64 images were correctly detected, in contrast, 44 images were incorrectly predicted.

*COVID-CT*. We included 349 COVID-19 positive and 349 normal CT images in the dataset. From all of the 698 images, we used 558 images to train our CNN model, the rest of the data were used for testing the model. Table 5 shows that 110 images were correctly detected images, and 30 images were incorrectly detected by the trained model. The confusion matrix shows our model achieved 79% detection accuracy.

*SARS-CoV-2 CT*. This is the largest dataset used in this experiment. We used 2460 CT images to train and test the models. Our 22-layer CNN model achieved 75% accuracy from this dataset.

## 4.2 *Result Comparison and Analysis*

As previously mentioned, we have used two well-known, widely used, build-in CNN models: ResNet-18 and VGG16. The main objective of using these two models was to assess the performance of the eight datasets on a large scale. Additionally, we compared the results of the two models with our proposed 22-layer CNN model. Table 6 shows the accuracy, precision, recall,

Retracted

Retracted

**Table 6** Data of performance measures following the eight datasets.

| Dataset name | Accuracy | | | Precision | | | Recall | | | *F*1-score | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Presented model | ResNet-18 | VGG16 | Presented model | ResNet-18 | VGG16 | Presented model | ResNet-18 | VGG16 | Presented model | ResNet-18 | VGG16 |
| Twitter | **0.91** | 0.62 | 0.87 | 0.85 | 0.89 | **0.96** | **1** | 0.59 | 0.81 | **0.92** | 0.7 | 0.88 |
| SIRM x-ray | 0.81 | 0.81 | **0.83** | 0.73 | **0.87** | **0.87** | **1** | 0.77 | 0.8 | **0.84** | 0.82 | **0.84** |
| COVID-19 Image Repository | **0.98** | 0.95 | 0.87 | **0.96** | 0.95 | 0.87 | **1** | 0.95 | 0.87 | **0.98** | 0.95 | 0.87 |
| EURORAD | 0.58 | **0.82** | 0.76 | 0.6 | 0.83 | **0.98** | 0.72 | **0.81** | 0.68 | 0.66 | **0.82** | 0.8 |
| BMICV | 0.97 | **0.98** | 0.97 | 0.94 | 0.97 | **1** | **1** | 0.98 | 0.95 | 0.97 | **0.98** | 0.97 |
| SIRM CT | 0.59 | **0.62** | 0.58 | **0.98** | 0.57 | 0.53 | 0.55 | **0.64** | 0.59 | **0.71** | 0.6 | 0.56 |
| COVID-CT | **0.79** | 0.76 | 0.66 | 0.72 | **0.73** | 0.61 | **0.91** | 0.77 | 0.68 | **0.81** | 0.75 | 0.65 |
| SARS-CoV-2 CT | **0.75** | 0.69 | 0.63 | **0.82** | 0.72 | 0.7 | **0.71** | 0.68 | 0.61 | **0.77** | 0.7 | 0.65 |

and $F1$-score of all eight datasets according to the three different CNN models. In Fig. 4, we present the comparison chart of the four performance measures from the three CNN models. From Table 6 and Fig. 4, our observations and interpretations over the results of the eight datasets are given individually below according to their performances:

*Accuracy.* We used bold font in Table 6 to indicate the highest accuracy achieved by the datasets. Our proposed model has the highest accuracy in four datasets, ResNet-18 in three, and VGG16 in only one dataset. The highest accuracy of Twitter, COVID-19 Image Repository, COVID-CT, and SARS-CoV-2 CT datasets achieved by the proposed model are 91%, 98%, 79%, and 75%, respectively. By the ResNet-18 model, EURORAD, BMICV, and SIRM CT datasets achieved an accuracy of 82%, 98%, and 62%, respectively. VGG16 model has the highest accuracy of 83% for the SIRM x-ray dataset.

*Precision*. As Table 6 shows, VGG16 achieved the highest precision in four datasets, the proposed model in three, and ResNet-18 in only one dataset. For the VGG16 model, Twitter, SIRM x-ray, EURORAD, and BMICV have the highest precision of 96%, 87%, 98%, and 100%, respectively. The highest precision of COVID-19 Image Repository, SIRM CT, and SARS-CoV-2 datasets are 96%, 98%, and 82% from our proposed model. Finally, the ResNet-18 model has the highest precision of COVID-CT and SIRM x-ray datasets at 73% and 87%, respectively.

*Recall*. Twitter, SIRM X-ray, COVID-19 Image Repository, BMICV, COVID-CT, and SARS-CoV-2 are the highest performing datasets under our proposed model in term of recall; their values are 100%, 100%, 100%, 100%, 0.91%, and 0.71%, respectively. EURORAD (81%) and SIRM CT (64%) achieved the highest recall under the ResNet-18 model.

*F1-score*. The performance of our proposed model maintained dominance in $F1$-score also. Most of the datasets performed high under our 22-layer CNN model, such as Twitter (92%), SIRM x-ray (84%), COVID-19 Image Repository (98%), SIRM CT (71%), COVID-CT (81%), and SARS-CoV-2 CT (77%). In addition, two datasets EURORAD and BMICV achieved the highest $F1$-score of 82% and 98% under the ResNet-18 model. Finally, the VGG16 model performs high (84%) with only on the SIRM x-ray dataset.

## 4.3 *Result and Discussion*

The fundamental objective of this study was to identify the public x-ray and CT datasets used to detect COVID-19 patients using ML techniques, then assess the quality of the datasets by analyzing their performance with the proposed and other two established CNN architectures. To summarize these results, the important question is as follows.

How well or poorly did the eight datasets perform in this investigation?

To answer this question, Fig. 4 illustrates a bar chart for each of the four different performance metrics. Every chart includes the eight datasets and their performance with each of the

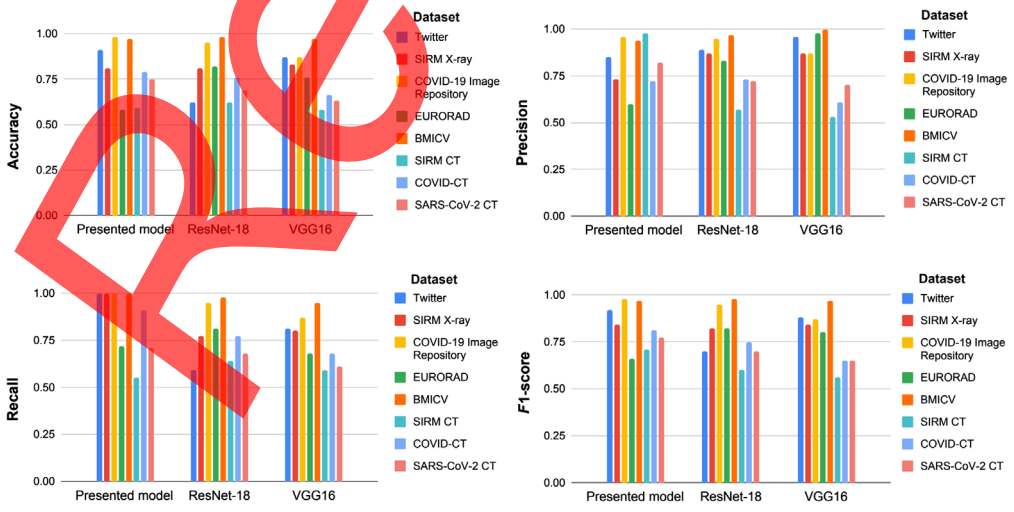


**Fig. 4** Performance comparison chart among the datasets.

Retracted

three models. It is quickly apparent that the BMICV dataset outperforms for every performance measure. The accuracy of this dataset under the proposed model, ResNet-18, and VGG16 was 97%, 98%, and 97%, respectively. Next is the COVID-19 Image Repository; this dataset also has high performance, 98% for the proposed model, 95% for ResNet-18, and 87% for VGG16. SIRM x-ray has a fair accuracy for the proposed (81%), ResNet-18 (81%), and VGG16 (83%) models. Additionally, the SIRM CT dataset achieved the worst accuracy with all three CNN models. Here the accuracy of the proposed model, ResNet-18, and VGG16 was only 59%, 62%, and 58%, respectively. The reason for the low accuracy is the high FP value in the confusion matrix, which indicates a large number of normal CT images were detected as COVID19 positive. More specifically, the quality of normal CT images is questionable; additional investigation is needed on the source data. Our collected datasets were versatile in the number of data samples, some datasets utilize a small number of samples, for instance, Twitter with only 114 and SIRM x-ray with 119 data. One of the established concepts in the ML-based data analysis model is the machine learning model needs a sufficient amount of data for learning, and a small amount of data can cause overfitting.[59] Consequently, we did not expect high accuracy from these datasets, however, they did achieve acceptable accuracy, precision, recall, and $F1$-score.

## 5 Conclusion

From the beginning of the COVID-19 pandemic, RT-PCR has been the most reliable diagnostic tool globally. In addition, researchers from artificial intelligence and ML have contributed to the COVID-19 crisis by developing automatic detection tools with advanced image analysis techniques. They used radiologic images such as x-ray and CT to train ML models and develop COVID-19 diagnosis tools. Though while the RT-PCR test continues to be the most popular, identifying COVID-19 infected patients using ML-based radiologic image analysis can help medical teams when there are insufficient test kits. In this study, we presented a deep learning-based COVID-19 detection model's performance assessment of existing open-access COVID-19 x-ray and CT image datasets. Specifically, we collected eight datasets from various open access repositories; all datasets carried unique (without repetition) COVID-19 positive x-ray or CT images. We trained the datasets individually with our proposed CNN model. Our proposed model outperformed using four datasets; their accuracies are Twitter 91%, COVID-19 Image Repository 98%, COVID-CT 79%, and SARS-CoV-2 CT 75%. While the datasets performed differently, this helps better explain the existing COVID-19 image datasets for researchers. We applied our CNN model on the collected datasets, but there are many deep and transfer learning architectures available in the field. In the future, we will apply other ML, transfer learning, quality, and bias analysis techniques architectures (such as Xception, Alexnet, and Googlenet) to the datasets to assess their quality and performance as well.

## Acknowledgments

The authors declare that they have no financial interests and no other potential conflicts of interest in the manuscript.

## References

1. A. J. Rodriguez-Morales et al., "Clinical, laboratory and imaging features of COVID-19: a systematic review and meta-analysis," *Travel Med. Infectious Disease* **34**, 101623 (2020).
2. WHO, "WHO coronavirus (COVID-19) dashboard," https://covid19.who.int/ (accessed 08 Aug. 2021).
3. L. Wynants et al., "Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal," *BMJ* **369**, m1328 (2020).
4. M. Ahishali et al., "Advance warning methodologies for COVID-19 using chest x-ray images," *IEEE Access* **9**, 41052–41065 (2021).
5. M. A. Elaziz et al., "New machine learning method for image-based diagnosis of COVID-19," *PLoS One* **15**(6), e0235187 (2020).

Retracted

6. J. Lin, J.-Y. Pai, and C.-C. Chen, "Applied patent RFID systems for building reacting HEPA air ventilation system in hospital operation rooms," *J. Med. Syst.* **36**(6), 3399–3405 (2012).
7. A. Basalamah and S. Rahman, "An optimized CNN model architecture for detecting coronavirus (COVID-19) with x-ray images," *Comput. Syst. Sci. Eng.* **40**, 375–388 (2022).
8. A. Z. Khuzani, M. Heidari, and S. A. Shariati, "COVID-classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images," *Sci. Rep.* **11**(1), 1–6 (2021).
9. M. F. Sohan, "So you need datasets for your COVID-19 detection research using machine learning?," https://arxiv.org/abs/2008.05906 (2020).
10. B. G. Santa Cruz et al., "Public Covid-19 X-ray datasets and their impact on model bias–a systematic review of a significant problem. medical image analysis," *Med. Image Anal.* **74**, 102225 (2021).
11. O. Albahri et al., "Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: taxonomy analysis, challenges, future solutions and methodological aspects," *J. Infection Pub. Health* **13**, 1381–1396 (2020).
12. W. T. Li et al., "Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis," *BMC Med. Inf. Decis. Making* **20**(1), 247 (2020).
13. J. Shuja et al., "COVID-19 open source data sets: a comprehensive survey," *Appl. Intell.* **51**(3), 1296–1325 (2021).
14. A. I. Khan, J. L. Shah, and M. M. Bhat, "CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," *Comput. Methods Prog. Biomed.* **196**, 105581 (2020).
15. J. P. Cohen et al., "COVID-19 image data collection: prospective predictions are the future," https://arxiv.org/abs/2006.11988 (2020).
16. K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
17. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Biol. Learn. Soc.*, https://arxiv.org/abs/1409.1556, 1–14 (2015).
18. A. Rehman et al., "COVID-19 detection empowered with machine learning and deep learning techniques: a systematic review," *Appl. Sci.* **11**(8), 3414 (2021).
19. L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images," *Sci. Rep.* **10**(1), 1–12 (2020).
20. T. Ozturk et al., "Automated detection of COVID-19 cases using deep neural networks with x-ray images," *Comput. Boil. Med.* **121**, 103792 (2020).
21. T. Tuncer, S. Dogan, and E. Akbal, "A novel local senary pattern based epilepsy diagnosis system using EEG signals," *Aust. Phys. Eng. Sci. Med.* **42**(4), 939–948 (2019).
22. A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of COVID-19 in chest x-ray images using DeTraC deep convolutional neural network," *Appl. Intell.* **51**(2), 854–864 (2021).
23. B. Abraham and M. S. Nair, "Computer-aided detection of COVID-19 from x-ray images using multi-CNN and BayesNet classifier," *Biocybern. Biomed. Eng.* **40**(4), 1436–1445 (2020).
24. M. Nour, Z. Cömert, and K. Polat, "A novel medical diagnosis model for COVID-19 infection detection based on deep features and Bayesian optimization," *Appl. Soft Comput.* **97**, 106580 (2020).
25. M. F. Aslan et al., "CNN-based transfer learning–BiLSTM network: a novel approach for COVID-19 infection detection," *Appl. Soft Comput.* **98**, 106912 (2021).
26. M. Loey, F. Smarandache, and N. E. M Khalifa, "Within the lack of chest COVID-19 x-ray dataset: a novel detection model based on GAN and deep transfer learning," *Symmetry* **12**(4), 651 (2020).
27. M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest x-ray images based on the concatenation of Xception and ResNet50v2," *Inf. Med. Unlocked* **19**, 100360 (2020).


Retracted

28. N. Tsiknakis et al., "Interpretable artificial intelligence framework for COVID-19 screening on chest x-rays," *Exp. Ther. Med.* **20**(2), 727–735 (2020).
29. A. T. Sahlol et al., "COVID-19 image classification using deep features and fractional-order marine predators algorithm," *Sci. Rep.* **10**(1), 1–15 (2020).
30. B. Nigam et al., "COVID-19: automatic detection from x-ray images by utilizing deep learning methods," *Expert Syst. Appl.* **176**, 114883 (2021).
31. C. Ouchicha, O. Ammor, and M. Meknassi, "CVDNet: a novel deep learning architecture for detection of coronavirus (COVID-19) from chest x-ray images," *Chaos, Solitons Fractals* **140**, 110245 (2020).
32. A. Saygılı, "A new approach for computer-aided detection of coronavirus (COVID-19) from ct and x-ray images using machine learning methods," *Appl. Soft Comput.* **105**, 107323 (2021).
33. P. Saha, M. S. Sadi, and M. M. Islam, "EMCNet: automated COVID-19 diagnosis from x-ray images using convolutional neural network and ensemble of machine learning classifiers," *Inf. Med. Unlocked* **22**, 100505 (2021).
34. H. S. Maghdid et al., "Diagnosing COVID-19 pneumonia from x-ray and ct images using deep learning and transfer learning algorithms," *Proc. SPIE* **11734**, 117340E (2021).
35. H. Yasar and M. Ceylan, "A novel comparative study for detection of COVID-19 on ct lung images using texture analysis, machine learning, and deep learning methods," *Multimedia Tools Appl.* **80**(4), 5423–5447 (2021).
36. A. Jaiswal et al., "Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning," *J. Biomol. Struct. Dyn.* **39**, 5682–5689 (2020).
37. B. J. Erickson et al., "Machine learning for medical imaging," *Radiographics* **37**(2), 505–515 (2017).
38. M. J. Willemink et al., "Preparing medical imaging data for machine learning," *Radiology* **295**(1), 4–15 (2020).
39. B. Heinrichs and S. B. Eickhoff, "Your evidence? machine learning algorithms for medical diagnosis and prediction," *Hum. Brain Mapp.* **41**(6), 1435–1444 (2020).
40. www.eurorad.org.
41. www.sirm.org.
42. www.radiopedia.org.
43. www.twitter.com/ChestImaging.
44. https://bimcv.cipf.es.
45. X. Wang et al., "ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2097–2106 (2017).
46. D. Kermany et al., "Labeled optical coherence tomography (OCT) and chest x-ray images for classification," *Mendeley Data* **2**(2) (2018).
47. https://radiopaedia.org/articles/covid-19
48. https://twitter.com/ChestImaging
49. https://sirm.org/category/senza-categoria/covid-19/.
50. https://figshare.com/articles/dataset/COVID-19_Image_Repository/12275009.
51. https://www.eurorad.org/.
52. M. E. Chowdhury et al., "Can AI help in screening viral and COVID-19 pneumonia?" *IEEE Access* **8**, 132665–132676 (2020).
53. https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/#1590858128006-9e640421-6711&450.
54. P. Angelov and E. Almeida Soares, "SARS-CoV-2 CT-scan dataset: a large dataset of real patients ct scans for SARS-CoV-2 identification," MedRxiv (2020).
55. https://www.kaggle.com/plameneduardo/sarscov2-ctscan-dataset.
56. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.
57. E. Hussain et al., "CoroDet: a deep learning based classification for COVID-19 detection using chest x-ray images," *Chaos, Solitons Fractals* **142**, 110495 (2021).
58. H. Panwar et al., "A deep learning and GRAD-CAM based color visualization approach for fast detection of COVID-19 cases using chest x-ray and CT-scan images," *Chaos, Solitons Fractals* **140**, 110190 (2020).


Retracted

59. H. Zhang, L. Zhang, and Y. Jiang, "Overfitting and underfitting analysis for deep learning based end-to-end communication systems," in *11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, IEEE, pp. 1–6 (2019).
60. M. F. Sohan et al., "Prevalence of machine learning techniques in software defect prediction," in *Int. Conf. Cyber Secur. and Comput. Sci.*, Springer, pp. 257–269 (2020).
61. M. F. Sohan and A. Basalamah, "A systematic literature review and quality analysis of Javascript malware detection," *IEEE Access* **8**, 190539–190552 (2020).

**Md. Fahimuzzman Sohan** received his BSc degree in software engineering from Daffodil International University, Dhaka, Bangladesh, in 2019. He has published journal article in *IEEE Access* and several conference proceedings. His research interests include machine learning, computer vision, and image processing.

**Anas Basalamah** received his MSc and PhD degrees from Waseda University, Tokyo, Japan, in 2006, 2009, respectively. He worked as a postdoctoral researcher at the University of Tokyo and the University of Minnesota in 2010 and 2011, respectively. He is an associate professor in the Department of Computer Engineering of Umm Al Qura University. His areas of interest include embedded networked sensing, smart cities, ubiquitous computing, participatory, and urban sensing.

**Md. Solaiman** received his BSc degree in software engineering from Daffodil International University, Dhaka, Bangladesh in May 2019. His research interest includes data science, AI, machine learning, and image processing.

Retracted