Neural Computing and Applications
https://doi.org/10.1007/s00521-024-10326-8

ORIGINAL ARTICLE

A Siamese neural network-based diagnosis of COVID-19 using chest X-rays

Engin Tas[1,2] · Ayca Hatice Atli[1,2]

Received: 22 February 2024 / Accepted: 29 July 2024

Abstract

Radiological findings play an essential and complementary role in diagnosing Covid-19, assessing its severity, and managing its patients. Artificial intelligence technology based on medical imaging, which has made exciting developments by being applied in many areas, has become an area of interest for the rapid and accurate detection of the disease in the fight against the Covid-19 pandemic. The main difficulty is the inability to obtain a large dataset size with quality and standard images that neural networks need to perform well. Aiming at this problem, this study proposes a Siamese neural network-based deep learning framework for accurate diagnostics of Covid-19 using chest X-ray (CXR) images. The pre-trained VGG16 architecture, based on the transfer learning approach, forms the backbone of the Siamese neural network. The outputs of the backbones are joined together by a merging layer, and then the output passes through a fully connected layer. Based on this structure, category-aware Siamese-based models are produced for each class. The predictions from the models are combined using a voting mechanism to reduce the possibility of misclassification and to make better decisions. The framework was evaluated using a publicly available dataset for the 4-class classification task for Covid-19 pneumonia, lung opacity, normal, and non-Covid-19 viral pneumonia images. The findings reveal the high discrimination ability of the framework, trained using only 10 images per class in less training time, achieving an average test accuracy of 92%. Our framework, which learns a single Siamese-based pairwise model for each class, effectively captures class-specific features. Additionally, it has the potential to deal with data scarcity and long training time problems in multi-class classification tasks.

Keywords Chest X-ray images · Classification · Covid-19 · Siamese neural network

1 Introduction

One of the most major catastrophes of the twenty-first century, the Covid-19 pandemic, has resulted in more than 700 million cases and more than 6 million fatalities globally. To stop the spread of the virus, infected people must be isolated and treated as soon as possible after being identified through quick and accurate tests.

✉ Ayca Hatice Atli
aturkan@aku.edu.tr

Engin Tas
engintas@aku.edu.tr

1 Department of Statistics, Afyon Kocatepe University, ANS Campus, 03200 Afyonkarahisar, Turkey

2 IYAMER, Afyon Kocatepe University, ANS Campus, 03200 Afyonkarahisar, Turkey

The real-time reverse transcription polymerase chain reaction (RT-PCR) test, which analyzes the upper respiratory tract specimen, is extensively utilized to diagnose Covid-19. However, obtaining the results of the PCR test could take hours or even days, depending on the hospital's or health institution's workload. In addition to the various available test kits, the development of dependable and quick test kits continues, but the need for diagnostic kits and insufficient production are among the limitations. Another option for diagnosing infected individuals is to evaluate individuals' chest images using medical imaging techniques such as X-rays, computed tomography (CT) scans, and magnetic resonance imaging (MRI), which provide information to physicians in numerous medical fields. In addition to supporting the diagnosis, radiological findings play an essential and complementary role in determining the disease's severity, guiding patient management decisions and treatment, and assessing the patient's response to treatment. Because respiratory problems and abnormalities detected using imaging methods are among the primary symptoms of Covid-19 [1–3]. An advantage of using images is that most hospitals and laboratories already have the appropriate equipment and imaging systems. However, the fact that the interpretation of the images mainly depends on the personal experience of the relevant radiologist or specialist and that Covid-19 reveals similar radiological findings to other lung diseases make the clinical diagnosis of Covid-19 difficult [4–7]. In addition, the disease has a wide range of radiographic features [8, 9], and interpretation of chest images may be challenging. With this motivation, numerous artificial intelligence-based techniques with promising results have been proposed to detect Covid-19 cases and assist specialists in diagnosing using radiological images. Most studies have included classification tasks with two, three, or more categories focused on distinguishing individuals with Covid-19 pneumonia from individuals with other pneumonia or healthy individuals. X-ray and CT images constitute the two primary datasets used for classification. Limitations and difficulties such as data scarcity, non-standard datasets, and non-repeatability restrict the proposed models' theoretical and clinical applications. Therefore, efforts for generalization and improving robustness determine the direction of research.

One approach to dealing with the large training set requirement and long training time problem is to apply to the Siamese neural network (SNN) architecture, also known as metric-based few-shot learning [10]. Examples of distance metrics include cosine similarity, Euclidean, and Manhattan distance. The goal of the network is not to directly recognize or classify the input. The network aims to determine the similarity (or difference) between the input with the class label and the input whose class label is unknown by learning their good encodings that provide representations of the images. Extracting features based on similarity and difference helps to train the model with fewer examples [11]. For this reason, despite being limited, SNNs were also discussed in Covid-19 researches. Li et al. [12] provided a Siamese neural network-based algorithm using DenseNet121 to measure COVID-19 disease severity in chest radiographs. Shalu et al. [13] proposed an approach for binary and multi-class classification scenarios related to Covid-19. Their approach included a Siamese network and achieved a testing accuracy of over 98%. Jadon [11] applied to the Siamese network to detect Covid-19 using chest radiographs and achieved an accuracy of over 96%. Jiang et al. [14] proposed a Siamese network-based method for the Covid-19 diagnostic task using CT scans, addressing the domain shift issue. They used the Xception network as the feature extractor and achieved around 80% accuracy for the 5-shot image classification task. Li et al. [15] used dual-Siamese channels consisting of four encoders (Res2Net) to extract image features of lesion regions while evaluating four clinical stages of Covid-19 patients based on CT images and clinical metadata. Their method achieved 86.7% accuracy. Shorfuzzaman and Hossain [16] reached over 95% accuracy with a deep Siamese network model they proposed to diagnose Covid-19 from chest radiographs. Abugabah et al. [17] proposed a Siamese convolutional neural network model using X-ray images to classify images with Covid-19, non-Covid-19, and pneumonia. They achieved 96.70% test accuracy. Al Rahhal et al. [18] proposed an approach based on a vision transformer that uses a Siamese encoder to separate the categories of normal, Covid-19, and non-Covid-19 pneumonia. Nneji et al. [19] used the Siamese network in their proposed approach for binary and four-class classification tasks for Covid-19 identification.

Although the relevant literature is growing rapidly, using non-public data or an unknown subset of publicly available data without a specific set of rules or not disclosing experimental details causes the inability to reproduce published results, and the proposed methods suffer from generalizability to new data. We considered these issues sensitively, presented the proposed framework's data organization scheme in detail, and tried to provide experimental details meticulously to ensure that the experiments in the study were reproducible. To the best of our knowledge, earlier SNN-based Covid-19 studies mainly dealt with two- or three-class classification tasks considering Covid-19, healthy/normal, and other pneumonia categories. However, there is still a need to enhance disease diagnostic systems based on medical images for other datasets with more challenging categories. This paper examines categories different from those considered in previous Siamese network-based approaches regarding Covid-19. We propose a framework containing category-aware Siamese-based models combined by majority voting for a four-class classification task related to the categories Covid-19 pneumonia, lung opacity, normal, and non-Covid-19 viral pneumonia. For this purpose, we employ chest X-rays from an open-source dataset. The framework uses four pairwise models, each containing a Siamese neural network. As the backbones of the Siamese network, we use the VGG16 deep learning architecture, pre-trained on images from ImageNet, and then fine-tuned on the task-specific dataset after a few modifications. The training of traditional Siamese neural networks mainly includes distance metric and distance-based triplet loss and contrastive loss. We incorporate structural changes to the Siamese neural network, such as concatenation instead of the commonly used distance metric and binary entropy loss instead of triplet and contrastive loss. In the testing phase, we combine the predictions of the models included in the framework with the majority vote. In order to demonstrate the main contribution of the proposed framework, we first compare it with a traditional VGG16 and a single Siamese-based pairwise model adapted to the four-class classification task and its distance metric-used version. Next, we compare it with ten state-of-the-art models to show its superiority.

The main contributions of the article are summarized as follows:

- The proposed framework, which can also be trained on small amounts of data, involves using feature conjunctions across the two examples via concatenation of image pairs examples' feature vectors.
- The proposed approach assigns the final class label using class-specific Siamese-based pairwise models and majority voting.
- When the proposed framework is applied by concatenating the feature vectors of the images, an average of 92% test accuracy is reached in the four-class classification.
- The suggested approach distinguishes Covid-19 from other viral pneumonia and lung opacity.

The organization of the remainder of the paper is as follows. The second section introduces the proposed framework with all its components. The third section provides the CXR dataset used, the pre-processing applied to the CXR images, the data organization scheme, and other experimental details and results. In the last section, the article concludes with potential directions for further research and conclusions.

2 Materials and methods

In this section, we tried to explain the proposed methodology in four separate sections in accordance with the progress of the main framework. We will first consider the training of the pre-trained models used, then, based on pre-trained models, we detail the pairwise model based on a Siamese network, and finally, a specific majority vote system that combines all these.

2.1 Transfer learning

Transfer learning is the use of knowledge gained by a model from a task with a large amount of labeled training data in a new task that contains less data or, equivalently, the reuse of the pre-trained model on a new task. This paper includes transfer learning using the VGG16 model trained on the ImageNet-1K dataset [20], commonly used for pre-training deep learning models. ImageNet-1K is a subset of the ImageNet dataset designed for visual object recognition research. It is an extensive visual database containing 1.2 million labeled images of 1000 object classes we encounter daily. The VGG16 network [21], whose parameters are initialized with the weights obtained from the ImageNet-1K training, is used as the baseline network. Only the upper layers of the VGG16 network, that is, the fully connected layers, are removed, and the fully connected layer with the sigmoid activation function is added. Since the pooling layer produces a matrix as output, a flatten layer is also added, transforming the matrix into a vector that will be the input for the dense layer. The output layer in the network is revised to have two nodes. Within the proposed framework, VGG16, modified as described above, is trained for binary classifications of Covid-19 vs. others, lung opacity vs. others, normal vs. others, and non-Covid-19 viral pneumonia vs. others. In this way, it is aimed to learn each category separately and, therefore, more accurately. The network is optimized based on accuracy and Adam optimizer, which minimizes the cross-entropy loss. The training is conducted for 140 epochs with a batch size of 16. The learning rate is fixed to 10^{-5}. An example of a pre-trained VGG16 network for the class Covid-19 is given in Fig. 1. Models obtained after fine-tuning take place in a separate Siamese-based pairwise architecture for each of the four classes labeled Covid-19, lung opacity, normal, and viral pneumonia.

2.2 Siamese-based pairwise model (SbPM)

This study introduces a novel pairwise classifier that draws inspiration from the Siamese model but differs significantly in structure. To understand this concept, we must first define positive and negative pairs. For a reference class like Covid-19, a positive pair is a pair in which both images belong to the reference class (Covid-19), while a negative pair comprises one image from the reference class and the other from a different class. The Siamese-based pairwise classifier (SbPC) aims to identify whether a pair is a positive pair based on the reference class.

Figure 2 illustrates this idea; let (x_i, x_j) be defined as the pair of x_i and x_j images where $x_i \in \mathbb{R}^{d\times h\times w}$ $i \in \mathcal{M} := \{1, 2, \ldots, m\}$, d is the number of channels, h the height, and w the width, respectively. The image pairs, $c_{ij} = (x_i, x_j) \epsilon \mathbb{R}^{d\times h\times w} \times \mathbb{R}^{d\times h\times w}$, are fed to the proposed Siamese neural network to get feature pairs, (f_i, f_j). The Siamese neural network comprises two backbones with the same configuration and weights. Each backbone consists of a pre-trained model. The last layers' outputs of the pre-trained architectures are fed into a concatenation layer. Next, a fully connected layer with 128 neurons, sigmoid function, L2 regularization, and an output layer with

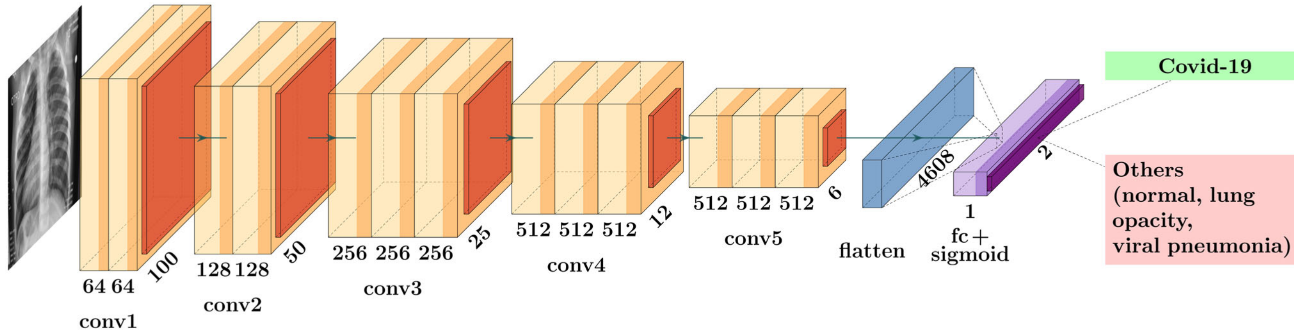
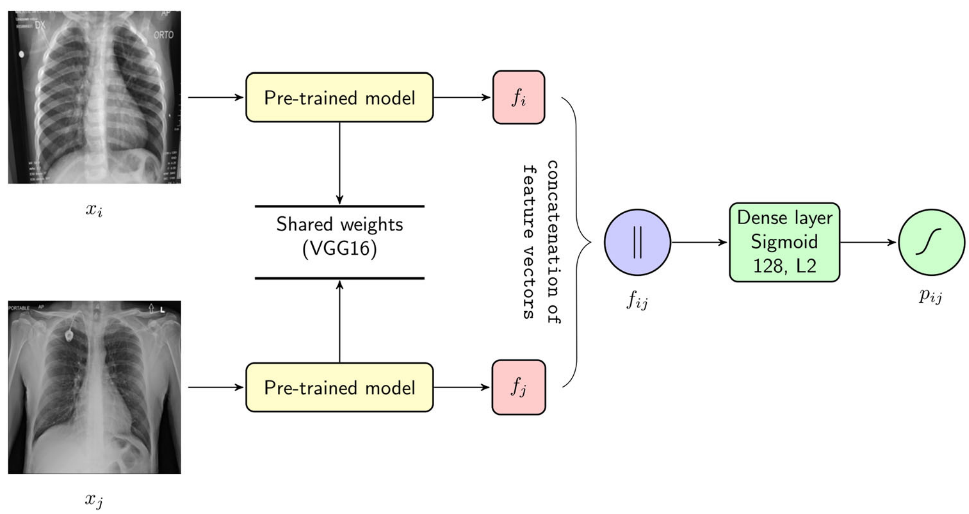


Fig. 1 Architecture of a pre-trained VGG16 model for Covid-19 class

Fig. 2 A simple schema of the proposed Siamese-based pairwise classifier (SbPC) for a specific class (e.g., Covid)



sigmoid activation are added. The proposed architecture aims to obtain pairwise predictions for each class. Based on the output of the network, p_{ij}, pairs are labeled according to whether or not the examples of the input images given in pairs belong to the same class. If the examples of the pair belong to the same class $\left(p_{ij} > 0.5\right)$, the pair, c_{ij}, is called a positive pair, and the corresponding binary label $\hat{y}_{ij}$ is 1. On the contrary, if the examples of the pair, c_{ij}, do not belong to the same class $\left(p_{ij} < 0.5\right)$, this pair is called a negative pair, and $\hat{y}_{ij}$ is 0. With these goals, the described network (Fig. 3) is optimized by minimizing the binary cross-entropy loss using training and validation pairs. We used the latex code [22] to illustrate the convolutional neural network architectures in Figs. 1 and 3.

2.3 Loss function

The traditional Siamese network branches result in different encodings as the parameters of the conventional Siamese network change. Therefore, the target is to learn the network parameters that provide good encodings. Distance-based loss functions are widely implemented during training to learn the parameters of the network that provide better encoding of images. Since we try to solve the multi-class classification problem by transforming it into binary classification problems for each class separately, the loss function used in the proposed Siamese-based pairwise model is the binary cross-entropy loss function:

$$L = -y_{ij}\log\left(p_{ij}\right) + \left(1 - y_{ij}\right)\log\left(1 - p_{ij}\right), \tag{1}$$

where y_{ij} is the true label and p_{ij} the probability of ith and jth examples being from the same class.

2.4 Final class prediction

Once we have SbPC for each class, the classification of a new image x from the test set follows a specific process. Firstly, a query set is generated by randomly selecting a predetermined number of (r) examples from all classes in the training set. Next, a paired query set is created by matching the image x with the examples in the query set. This pairing process is repeated for each example in the test set, resulting in a pairwise test set.

We then present these pairwise test images to all trained SbPCs and obtain the pairwise predictions. A pairwise

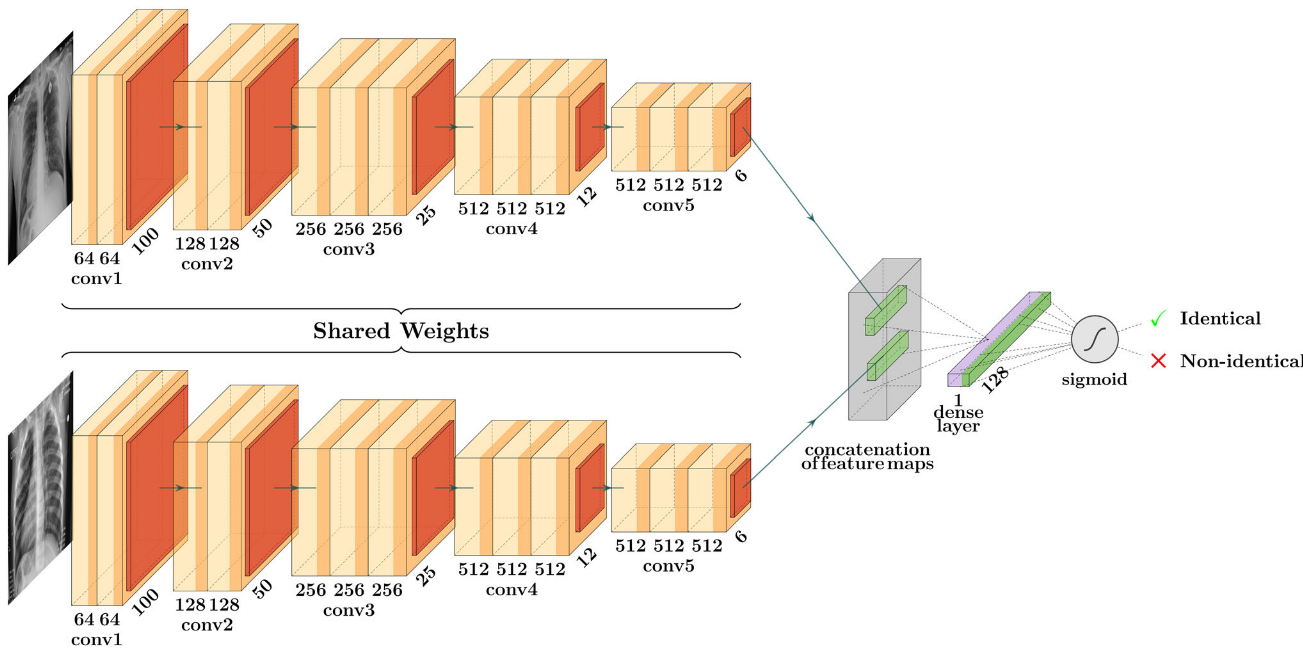


Fig. 3 A detailed view of a Siamese-based pairwise model for a specific class (e.g., Covid)

prediction determines whether the images that make up the pair come from the same class. In other words, each training image we match with the new (test) image gives a vote for this new image as to whether it belongs to the class of the training image or not.

Pairwise predictions from all SbPCs are combined via the majority voting rule. The set of voting pairs is $Q(x) = \{q_i := (x, x_i^k) : i = 1, \ldots, r; k = 1, \ldots, l\}$. The voting method described below [23] is used to make the final prediction about the class of the new image:

$$M_k(x) = \sum_{q_i \in Q} v(p^k(q_i)) \tag{2}$$

$$v(t) = \begin{cases} 1, & t > 0.5; \\ -1, & \text{otherwise.} \end{cases} \tag{3}$$

Here, $p^k(q_i)$ represents the probability that the unknown test example x in the pair q_i belongs to the kth class, and $v(t)$ maps $p^k(q_i)$ to $\{-1, 1\}$ for each test pair. Since the class label of the training example in each pair is known, it generates a vote for the test example to which it is matched. If p^k decides that the examples of a pair are not from class k ($p^k < 0.5$), it contributes -1 as its vote to $M_k(x)$, otherwise, it contributes $+1$. After collecting votes from all classifiers and completing the voting process, the class label of image x is decided as follows:

$$\text{class of } x = \text{argmax}_{k=1,\ldots,l} M_k(x). \tag{4}$$

That is, x is assigned to the class with the highest number of votes.

We have provided the details of the approach we suggested in the above four sections. We call this multi-class classification technique “MultiCOVID”, which combines pre-trained models, SbPMs, and a special majority voting scheme.

3 Experiments

In this section, we evaluate MultiCOVID for the multi-class classification task on the dataset described in Sect. 3.2. We compare the suggested framework with three methods to highlight its contributions. The first is a version of MultiCOVID where we use the elementwise absolute difference of feature vectors as the merge layer in SbPMs instead of concatenation. For simplicity, we call this version MultiCOVID_AbsDiff and our proposed model MultiCOVID_Concat. The second is VGG16, whose architecture for two classes is shown in Fig. 1, and the experimental setup is provided in Sect. 2.1. Only the classification layer was changed to adapt the architecture to the 4-class classification problem. In our suggested framework, each SbPM learns to discriminate one of the four classes from the others. As the third model, a single SbPM trained for four classes is considered, which is compatible with the traditional similarity learning task of Siamese networks. After relevant evaluations, the experimental results are concluded by comparing the proposed framework with ten available state-of-the-art models.

When comparing models, two essential points were considered. First, depending on the initial values of the parameters, we observed that the trained network converges to different solutions. Second, since the objective of the suggested approach is to enable the network to learn with a limited number of examples, at the beginning of the training, 10 examples from each class are chosen randomly and used to train the models. Therefore, the solution to which the network converges at the end of the training varies depending on the selected examples at the beginning. Considering these two critical points, network training and testing were performed 10 times for each model to compare the techniques accurately. In each trial, the network was initialized with distinct initial parameters. Furthermore, ten different examples were drawn randomly from each class, provided that the chosen example was not re-selected in another trial. These examples were then used to generate pairwise training and validation sets. The provided results include the mean and standard deviation of 10 trials for relevant performance metrics when comparing models. Thus, the models are compared in a manner that ensures fairness.

When the studies in the literature were examined, it was seen that the processes related to the experiments were not explained enough, and the critical points we discussed above were not handled with sensitivity. However, we have observed in our experiments that these details significantly impact the training. Therefore, since the MultiCOVID technique we recommend has its data organization, it is necessary to outline the training and testing process step by step as follows:

- First, the data set is separated into three categories: training, validation, and testing.
- To create pairwise training data sets, the first step is to randomly select ten examples from each class in the training set. Since there are four classes, a total of 40 examples are selected. Then, each example is paired with one positive (from the same class) and one negative (from a different class) example from the training set. As a result, 40 x 2 = 80 training pairs are created for all classes, that is, pairwise training sets for four classes.
- All examples in the validation set are utilized for constructing pairwise validation sets. Each example in each class is paired with one positive (from the same class) and one negative (from a different class) example from the validation set. As a result, *the number of examples in that class* x 2 pairs are formed for each class, and *the total number of examples in the validity set* x 2 validity pairs are created for all classes. This meticulous process ensures a balanced representation of positive and negative pairs within the pairwise training and validation datasets across the Covid-19 pneumonia, lung opacity, normal, and non-Covid-19 viral pneumonia classes.
- An individual SbPM (Fig. 3) is trained for each class using these training and validation pairs.
- Once the training of SbPMs is completed, a pairwise test set is formed. Each example in the test dataset is matched with a specified number ($r = 10$) of examples (query set) from each class in the training dataset while constructing the pairwise test dataset. Therefore, *the number of examples in the test set* x 10 test pairs is generated.
- The created pairwise test dataset is fed into SbPMs, each specially trained for only one class, to yield pairwise predictions.
- Finally, the class of the test example is determined by pairwise test set predictions and majority vote, as described in Sect. 2.4.

3.1 Experimental setup

All training and tests were conducted on a system with two Nvidia Tesla P100 GPUs, one 20-core Intel Xeon Scalable 6148 processor, and 384 GB of ECC 2600Mhz memory. The proposed model was trained with the Python 3.8 compiler, Tensorflow 2.3.0, and Keras 2.4.3 libraries. Framework networks were trained on a batch size of 20 using Adam optimization with a learning rate of 10^{-4} for a maximum of 500 epochs to minimize the binary cross-entropy loss function. In addition, the validation loss was monitored throughout the training. When the validation loss increased for a specified number of steps (5), early stopping was implemented, and the training was terminated. In this process, the network parameters of the model with the lowest validation loss were automatically saved for test evaluation. Similarly, benchmarking methods were trained for a maximum of 500 epochs using the (online) Adam optimizer and binary cross-entropy loss. All the experiments were performed using TUBITAK ULAKBIM, High Performance, and Grid Computing Center.

3.2 Dataset and pre-processing

We used the publicly available Covid-19 Radiography Database [24, 25]. This dataset was created from published papers, online resources, and publicly available datasets by researchers from Qatar University and the University of Dhaka, their collaborators from Pakistan and Malaysia, and with the help of medical doctors. The dataset includes chest images in four categories: Covid-19 positive case, lung opacity, normal, and non-Covid-19 viral pneumonia. Of the chest X-rays, 3616 belong to individuals with Covid-19, 6012 to individuals with lung opacity, 10192 to normal individuals, and 1345 to individuals with other viral pneumonia. The training dataset includes 2531, 4208, 7134, and 941 CXR images for Covid-19, lung opacity, normal, and other viral pneumonia categories, respectively. The validation dataset includes 542, 1528, 901, and 201 CXR images for Covid-19, lung opacity, normal, and other viral pneumonia categories, respectively. The test dataset includes 543, 903, 1530, and 203 CXR images for Covid-19, lung opacity, normal, and other viral pneumonia categories, respectively. From the training dataset, 14,814 images were used to train the model, and 3,172 images were used to validate the model. But first, all images were resized (100×100), and pixel values were normalized between 0 and 1.

3.3 Evaluation metrics

The classification performances are reported using four metrics. These are recall, precision, F1_score, and accuracy, defined as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{F1\_score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{8}$$

Here, TP denotes true positive, TN true negative, FP false positive, and FN false negative. In addition, the average weighted by the support and unweighted mean for metrics, precision, recall, and F1_score is among the results.

3.4 Experimental results

Table 1 presents the average accuracy and training time from 10 trials of the proposed MultiCOVID framework and benchmarking techniques for the classification problem, including Covid-19 pneumonia, lung opacity, normal and non-Covid-19 viral pneumonia categories. The findings show that training for the MultiCOVID framework takes significantly less time than training for the VGG16 and single SbPM. Furthermore, the suggested framework trained with ten training images from each class achieves better accuracy than other models trained with the whole training images.

The average performance metrics of the approaches derived from 10 trials are shown in Table 2. The proposed MultiCOVID framework shows better classification performance compared to other methods, especially single SbPM. Even though VGG16, a classical multi-class classification technique, performed well, it used all examples in the training phase while making this classification. In contrast, MultiCOVID completed its training using only 10 examples from each class. Furthermore, when the stability of the methods was assessed, that is, looking at the classification accuracy values from 10 trials, VGG16 was the most stable method with a standard deviation of 0.001.

Based on ten replications for each compared model, Fig. 4 depicts a composite of box plots accompanied by scattered data points, showcasing between-subjects designs and incorporating statistical information within the visualization. Statistical analyses and demonstrations were performed using the ggstatsplot package in R [26]. All the data were tested for normality before the analysis, and normality assumptions were satisfied. Therefore, the models' classification accuracies were compared using Welsch's analysis of variance (ANOVA). The Games-Howell test was used for pairwise comparisons, and Bonferroni was used as an adjustment method for p-values.

A statistically significant difference $(p < 0.001)$ exists between the proposed models (MultiCOVID_Concat and MultiCOVID_AbsDiff) and the other models under comparison (VGG16 and Single_SbPM). MultiCOVID_Concat and MultiCOVID_AbsDiff exhibited notably superior classification performance compared to the other models. Conversely, there is no statistically significant difference between the proposed models MultiCOVID_Concat and MultiCOVID_AbsDiff.

Table 3 presents the classification performance metrics of SbPMs to see the effect of adopting vertical concatenation or absolute difference for merge layer in the MultiCOVID framework on the resulting pairwise predictions.

Table 1 Comparison of models for the four-class classification task

Model	Training dataset	Merge layer	Training time (min)	Accuracy
MultiCOVID_Concat	10 images per class	Concatenation	37	0.92 ± 0.007
MultiCOVID_AbsDiff	10 images per class	Absolute difference	33	0.92 ± 0.004
VGG16	Entire dataset	–	329	0.89 ± 0.001
Single_SbPM	Entire dataset	Absolute difference	321	0.70 ± 0.007

Table 2 Classification performances for the proposed framework and benchmarking techniques

	Metric	Class				Metric	Score
		Covid-19	Lung opacity	Normal	Viral pneumonia		
MultiCOVID_Concat	Precision	0.84 ± 0.029	0.91 ± 0.021	0.95 ± 0.009	0.96 ± 0.015	Macro average	0.91 ± 0.008
						Weighted average	0.92 ± 0.005
	Recall	0.96 ± 0.006	0.89 ± 0.015	0.92 ± 0.021	0.94 ± 0.012	Macro average	0.93 ± 0.004
						Weighted average	0.92 ± 0.007
	F1_score	0.89 ± 0.015	0.90 ± 0.004	0.93 ± 0.007	0.95 ± 0.005	Macro average	0.92 ± 0.006
						Weighted average	0.92 ± 0.007
	Support	543	903	1530	203	Accuracy	0.92 ± 0.007
MultiCOVID_ AbsDiff	Precision	0.91 ± 0.017	0.91 ± 0.02	0.94 ± 0.008	0.92 ± 0.033	Macro average	0.92 ± 0.01
						Weighted average	0.92 ± 0.004
	Recall	0.95 ± 0.007	0.89 ± 0.016	0.93 ± 0.016	0.97 ± 0.009	Macro average	0.93 ± 0.002
						Weighted average	0.92 ± 0.004
	F1_score	0.93 ± 0.008	0.90 ± 0.004	0.93 ± 0.005	0.94 ± 0.014	Macro average	0.93 ± 0.006
						Weighted average	0.92 ± 0.004
	Support	543	903	1530	203	Accuracy	0.92 ± 0.004
VGG16	Precision	0.88 ± 0.006	0.87 ± 0.008	0.89 ± 0.004	0.96 ± 0.005	Macro average	0.90 ± 0.003
						Weighted average	0.89 ± 0.001
	Recall	0.88 ± 0.004	0.84 ± 0.007	0.92 ± 0.007	0.90 ± 0.008	Macro average	0.88 ± 0.003
						Weighted average	0.89 ± 0.001
	F1_score	0.88 ± 0.002	0.85 ± 0.002	0.90 ± 0.001	0.93 ± 0.004	Macro average	0.89 ± 0.002
						Weighted average	0.89 ± 0.001
	Support	543	903	1530	203	Accuracy	0.89 ± 0.001
Single_SbPM	Precision	0.51 ± 0.02	0.67 ± 0.011	0.78 ± 0.006	0.63 ± 0.016	Macro average	0.65 ± 0.009
						Weighted average	0.69 ± 0.007
	Recall	0.43 ± 0.024	0.71 ± 0.008	0.76 ± 0.005	0.80 ± 0.015	Macro average	0.68 ± 0.01
						Weighted average	0.70 ± 0.007
	F1_score	0.47 ± 0.021	0.69 ± 0.009	0.77 ± 0.005	0.70 ± 0.014	Macro average	0.66 ± 0.009
						Weighted average	0.69 ± 0.007
	Support	543	903	1530	203	Accuracy	0.70 ± 0.007

The results reveal that MultiCOVID implemented with the concatenation operator performs better regarding pairwise predictions.

Figure 5 displays the confusion matrix for each approach. The confusion matrices in Fig. 5a, b demonstrate the proposed framework's remarkable performance in unhealthy CXR image classes. MultiCOVID framework with concatenation operator correctly diagnoses 521 of 543 Covid-19 patients but incorrectly diagnoses nine as lung opacity and 13 as normal. Of the 543 Covid-19 patients, the MultiCOVID framework with absolute difference operator accurately classifies 513; nevertheless, it misdiagnoses 12 as lung opacity, 17 as normal, and one as viral pneumonia. With 966 CXR images, a single SbPM has the highest misclassification score in the entire dataset and the highest misclassification score per class.

Figure 6 depicts the training and validation losses of the proposed framework and benchmarking methods. Training and validation for both implementations of MultiCOVID appear to be stable, and the spread of curves is quite low. Our proposed MultiCOVID_Concat converges significantly faster than benchmark approaches. Even VGG16, which has good performance metrics, requires 370 epochs. Finally, the comparative performance evaluation of MultiCOVID with ten state-of-the-art models for the four-class classification task is presented in Table 4. The experimental results indicate the superiority of our proposed framework, MultiCOVID, over the existing state-of-the-art models.

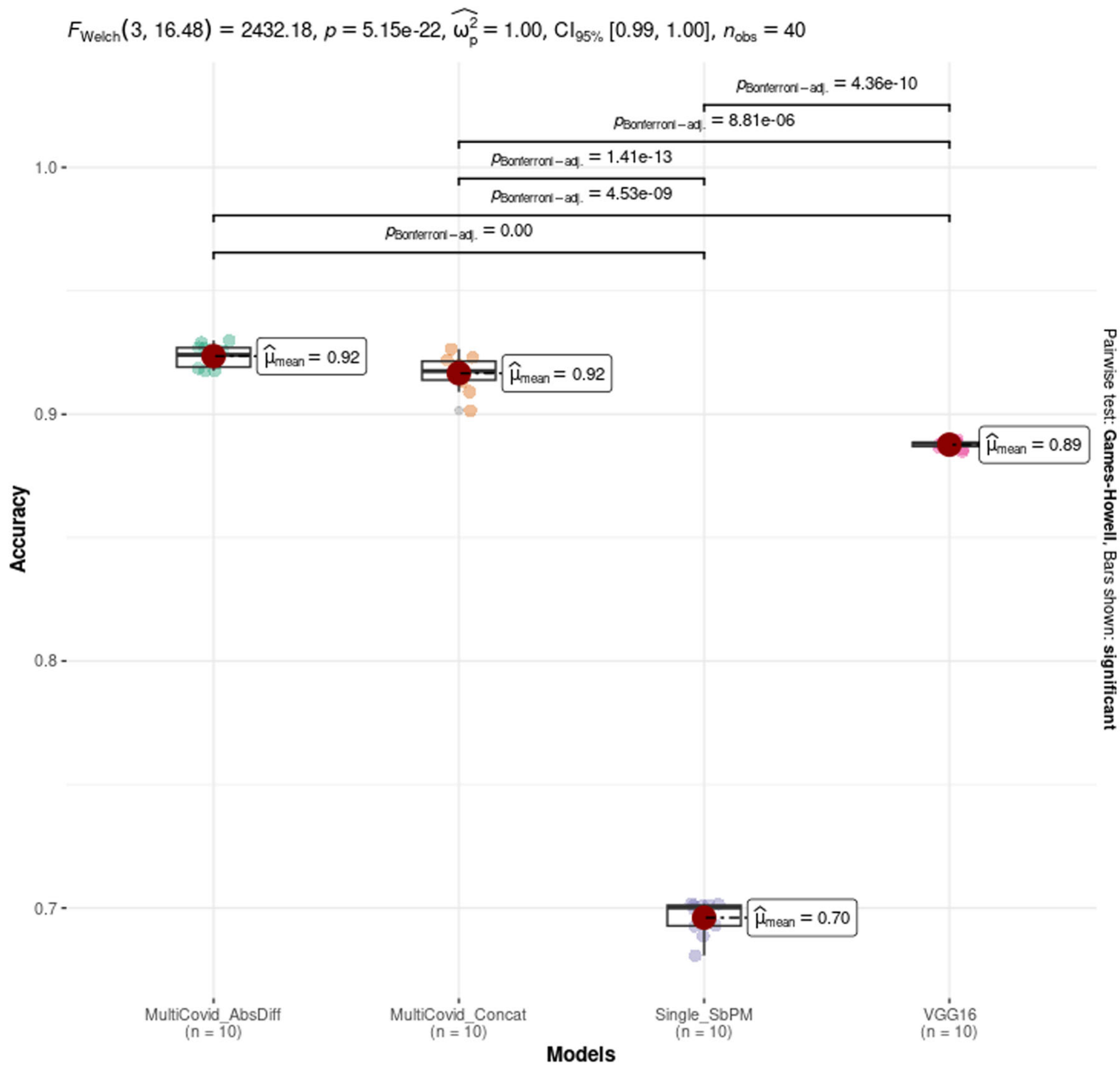


Fig. 4 Box-plot of replications of models with analysis of variance

Table 3 Evaluation results of pairwise predictions

Model	Class	Accuracy	Precision	Recall	F1_score	Roc Auc
MultiCOVID_ Concat	Covid-19	0.98	0.98	0.98	0.98	0.96
	Lung opacity	0.94	0.94	0.94	0.94	0.93
	Normal	0.93	0.93	0.93	0.93	0.93
	Viral pneumonia	0.99	0.99	0.99	0.99	0.98
MultiCOVID_ AbsDiff	Covid-19	0.96	0.96	0.96	0.96	0.94
	Lung opacity	0.90	0.90	0.90	0.90	0.88
	Normal	0.91	0.91	0.91	0.91	0.91
	Viral pneumonia	0.99	0.99	0.99	0.99	0.98

4 Discussion and conclusions

In pandemic crises, it is crucial to quickly make the right decisions regarding the disease with minimal resources. Problems with the RT-PCR test and the symptoms of Covid-19 have led to the use of medical imaging techniques that provide information in a shorter time in the diagnosis and management of the disease. Therefore, the combination of deep learning and image processing to support the assessment of Covid-19 has received considerable attention.

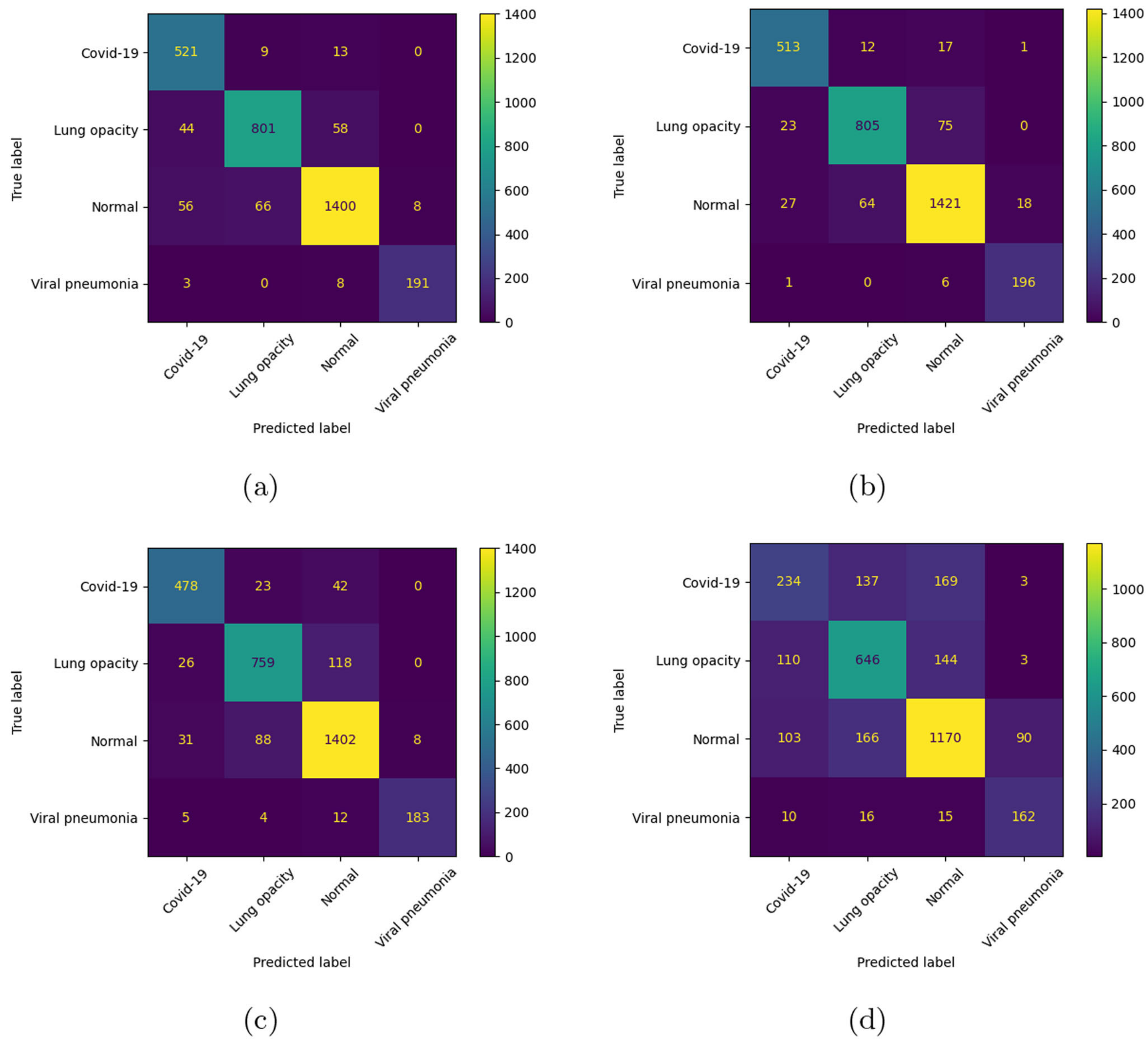


Fig. 5 Performance results of all techniques. **a** Confusion matrix of MultiCOVID_Concat. **b** Confusion matrix of MultiCOVID_AbsDiff. **c** Confusion matrix of VGG16. **d** Confusion matrix of single_SbPM

In studies on Covid-19, deep learning methods have been mainly applied to classification tasks with 2 or 3 classes, while less attention has been paid to applications for four or more classes. So far, when out of the six papers addressing the same categories and the same dataset are reviewed, it is seen that four of them applied to data augmentation and reported accuracy ranges between 92% and 96%. Brima et al. [35] suggested a framework with a ResNet-50 CNN architecture that achieved 94% accuracy using 15,241 training, 3809 validation, and 2115 test chest X-ray images. Senan et al. [36] reported 95% accuracy in their approach when they applied the data augmentation technique to the same data set and used ResNet-50 as a feature extractor. They reported 92% accuracy when they employed AlexNet. They stated the training time as 674 min 32 sec for ResNet-50 and 81 min 5 sec for AlexNet. Bashar et al. [37] achieved 95.63% classification accuracy through the VGG16 transfer learning algorithm using an enhanced augmented normalized dataset. Khan et al. [38] developed a technique based on EfficientNetB1, NasNet-Mobile, and MobileNetV2 pre-trained deep learning models. They applied an image augmentation approach to the dataset containing 21,165 images to increase the amount of data and balance the classes. With a classification accuracy of 96.13%, the regularized EfficientNetB1 model outperformed other models. Sanida et al. [39]

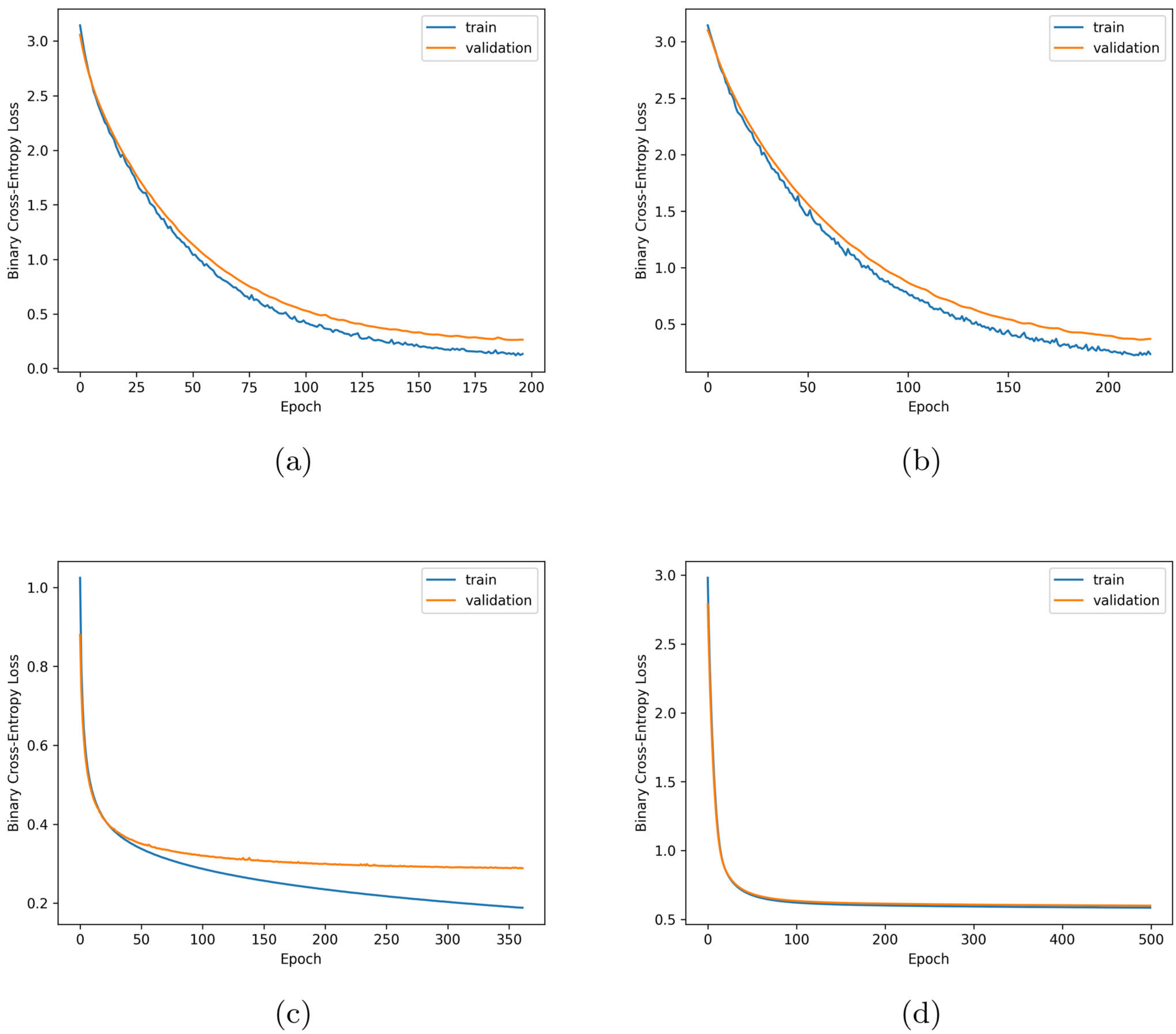


Fig. 6 Train versus validation loss curves reported for all techniques. **a** Loss curves of MultiCOVID_Concat. **b** Loss curves of MultiCOVID_AbsDiff. **c** Loss curves of VGG16. **d** Loss curves of single_SbPM

proposed a model based on the MobileNetV2 architecture. They achieved 95.8% accuracy with 14,813 images for training, 4232 images for validation, and 2120 images for testing. Using W-RESNET-50, Basu et al. [40] achieved 92.65% accuracy with a data set comprising 21,165 images and 94.57% accuracy with the second data set containing 27,790 images.

This paper proposes a framework for the multi-class classification task of X-ray images, named MultiCOVID. A pairwise model based on a Siamese neural network is addressed to each class. The framework is implemented in two ways, depending on how subnetworks' outputs' of the pairwise models are combined. One of these merging ways is done as a vertical concatenation of feature vectors, the other as the absolute difference of the corresponding elements in the extracted feature vectors of the two images of each pair. For the final classification, a majority vote is used based on the results of the pairwise models, each learning only one class.

MultiCOVID, which includes four class-specific SbPMs using vertical concatenation for subnetwork encodings, achieved approximately 92% classification performance. Furthermore, it is observed that the proposed models perform much better compared to single SbBM, which uses the same underlying structure (SbPM). This is because single SbPM relies on a single SbPM model to solve the multi-class classification problem. It attempts to learn a single model to identify similar pairs for all classes. While this may seem like a good idea initially, it becomes a more complex problem as the number of classes increases. On

Table 4 Comparison with state-of-the-art models

Model	Training dataset	Training time (min)	Accuracy
MultiCOVID_Concat	10 images per class	37	0.92 ± 0.007
MultiCOVID_AbsDiff	10 images per class	33	0.92 ± 0.004
ConvNeXtBase [27]	Entire dataset	45	0.64 ± 0.008
ConvNeXtXLarge [27]	Entire dataset	209	0.78 ± 0.004
DenseNet201 [28]	Entire dataset	55	0.88 ± 0.002
EfficientNetV2L [29]	Entire dataset	34	0.60 ± 0.017
EfficientNetV2S [29]	Entire dataset	43	0.63 ± 0.006
InceptionV3 [30]	Entire dataset	158	0.76 ± 0.004
MobileNetV2 [31]	Entire dataset	45	0.88 ± 0.005
NASNetLarge [32]	Entire dataset	31	0.84 ± 0.002
ResNet152V2 [33]	Entire dataset	55	0.86 ± 0.004
Xception [34]	Entire dataset	35	0.84 ± 0.003

the other hand, the proposed models learn a separate SbPM model for each class, thereby capturing class-specific features better. This demonstrates the effectiveness of the proposed approach for multi-class classification problems, but there are still aspects that can be further improved. Another critical point is that the proposed models achieve this performance with only 10 examples from each class and, therefore, in a much shorter time. In contrast, VGG16 and single SbPM have not attained this performance even though they use all of the training data and much more training time. The results demonstrate the potential of the proposed framework, which reduces a large amount of data dependency and training time. These benefits become essential in choosing an approach that can quickly adapt to the problem, especially in the early stages of new health emergencies, when data are lacking and quick decision-making is required. However, it can be said as a limitation that CT or other image datasets are not considered in the study.

Future studies may consider integrating different pre-trained models and other hyperparameter tuning methods into the proposed framework. Although many approaches have already been developed to recognize Covid-19 disease from image data, the need to create assistive systems for more sensitive assessments with different datasets remains. In this context, the proposed framework can be used in follow-up radiographs to classify changes related to the course or severity of the disease. In this way, the right treatment decisions will further increase the benefit of imaging technologies. Efforts for automatic and accurate detection of Covid-19 based on medical images will undoubtedly lead to significant advances in using artificial intelligence technology to fight against other diseases or pandemics.

Acknowledgements The numerical calculations reported in this paper were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

Author contributions All authors contributed equally to the study conception, design, material preparation, analyses, and the writing of the paper. All authors read and approved the final manuscript.

Funding No funding was received for conducting this study.

Data availability Not applicable.

Materials availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval and consent to participate Not applicable.

References

1. Ucar F, Korkmaz D (2020) Covidiagnosis-net: deep bayes-squeezenet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. Med Hypotheses 140:109761
2. Benmalek E, Elmhamdi J, Jilbab A (2021) Comparing CT scan and chest X-ray imaging for COVID-19 diagnosis. Biomed Eng Adv 1:100003
3. Bhardwaj P, Kaur A (2021) A novel and efficient deep learning approach for COVID-19 detection using X-ray imaging modality. Int J Imaging Syst Technol 31(4):1775–1791
4. Kong W, Agarwal P.P (2020) Chest imaging appearance of COVID-19 infection. Radiol: Cardiothorac Imaging 2(1):200028
5. Verma D, Bose C, Tufchi N, Pant K, Tripathi V, Thapliyal A (2020) An efficient framework for identification of tuberculosis and pneumonia in chest X-ray images using neural network. Proced Comput Sci 171:217–224
6. Yoon SH, Lee KH, Kim JY, Lee YK, Ko H, Kim KH, Park CM, Kim Y-H (2020) Chest radiographic and CT findings of the 2019 novel coronavirus disease (COVID-19): analysis of nine patients treated in Korea. Korean J Radiol 21(4):494–500
7. Zhao W, Zhong Z, Xie X, Yu Q, Liu J (2020) Relation between chest CT findings and clinical conditions of coronavirus disease (COVID-19) pneumonia: a multicenter study. Am J Roentgenol 214(5):1072–1077

8. Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A (2020) Coronavirus disease 2019 (COVID-19): a systematic review of imaging findings in 919 patients. Am J Roentgenol 215(1):87–93
9. Chamorro EM, Tascón AD, Sanz LI, Vélez SO, Nacenta SB (2021) Radiologic diagnosis of patients with COVID-19. Radiología (English Edition) 63(1):56–73
10. Sawyer D, Fiaidhi J, Mohammed S (2021) Few shot learning of COVID-19 classification based on sequential and pretrained models: a thick data approach. In: 2021 IEEE 45th annual computers, software, and applications conference (COMPSAC), pp 1832–1836. IEEE
11. Jadon S (2021) COVID-19 detection from scarce chest X-ray image data using few-shot deep learning approach. In: Medical imaging 2021: imaging informatics for healthcare, research, and applications, vol 11601, pp 161–170. SPIE
12. Li MD, Arun NT, Gidwani M, Chang K, Deng F, Little BP, Mendoza DP, Lang M, Lee S.I, O'Shea A, Parakh A, Singh P, Kalpathy-Cramer J (2020) Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. Radiol: Artif Intell 2(4):200079
13. Shalu H, Harikrishnan P, Das A, Mandal M, Sali HM, Kadiwala J (2020) A data-efficient deep learning based smartphone application for detection of pulmonary diseases using chest X-rays. arXiv preprint arXiv:2008.08912
14. Jiang Y, Chen H, Ko H, Han DK (2021) Few-shot learning for CT scan based COVID-19 diagnosis. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1045–1049. IEEE
15. Li Z, Zhao S, Chen Y, Luo F, Kang Z, Cai S, Zhao W, Liu J, Zhao D, Li Y (2021) A deep-learning-based framework for severity assessment of COVID-19 with CT images. Expert Syst Appl 185:115616
16. Shorfuzzaman M, Hossain MS (2021) Metacovid: a siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. Pattern Recogn 113:107700
17. Abugabah A, Mehmood A, Al Zubi AA, Sanzogni L (2022) Smart COVID-3d-scnn: a novel method to classify X-ray images of COVID-19. Comput Syst Sci Eng 41(3):997–1008
18. Al Rahhal MM, Bazi Y, Jomaa RM, AlShibli A, Alajlan N, Mekhalfi ML, Melgani F (2022) COVID-19 detection in CT/X-ray imagery using vision transformers. J Personal Med 12(2):310
19. Nneji GU, Cai J, Monday HN, Hossin MA, Nahar S, Mgbejime GT, Deng J (2022) Fine-tuned siamese network with modified enhanced super-resolution gan plus based on low-quality chest X-ray images for COVID-19 identification. Diagnostics 12(3):717
20. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis (IJCV) 115(3):211–252
21. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
22. Iqbal H (2018) Harisiqbal88/plotneuralnet v1.0.0. Available at: https://doi.org/10.5281/zenodo.2526396
23. Hsu C-W, Lin C-J (2002) A comparison of methods for multiclass support vector machines. IEEE Trans Neural Netw 13(2):415–425
24. Chowdhury ME, Rahman T, Khandakar A, Mazhar R, Kadir MA. Mahbub ZB, Islam KR, Khan MS. Iqbal A, Al Emadi N, Reaz MBI, Islam MT (2020) Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 8, 132665–132676
25. Rahman T, Khandakar A, Qiblawey Y, Tahir A, Kiranyaz S, Kashem SBA, Islam MT, Al Maadeed S, Zughaier SM, Khan MS, Chowdhury ME (2021) Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Comput Biol Med 132:104319
26. Patil I (2021) Visualizations with statistical details: the ggstatsplot approach. J Open Source Softw 6(61):3167
27. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S (2022) A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11976–11986
28. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
29. Tan M, Le QV (2021) Efficientnetv2: smaller models and faster training. In: International conference on machine learning, pp 10096–10106. PMLR
30. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
31. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
32. Zoph B, Vasudevan V, Shlens J, Le QV (2018) Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8697–8710
33. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. In: Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pp 630–645. Springer
34. Chollet, F (2017) Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
35. Brima Y, Atemkeng M, Tankio Djiokap S, Ebiele J, Tchakounté F (2021) Transfer learning for the detection and diagnosis of types of pneumonia including pneumonia induced by COVID-19 from chest X-ray images. Diagnostics 11(8):1480
36. Senan EM, Alzahrani A, Alzahrani MY, Alsharif N, Aldhyani TH (2021) Automated diagnosis of chest X-ray for early detection of COVID-19 disease. Comput Math Methods Med 2021:1–10
37. Bashar A, Latif G, Ben Brahim G, Mohammad N, Alghazo J (2021) COVID-19 pneumonia detection using optimized deep learning techniques. Diagnostics 11(11):1972
38. Khan E, Rehman MZU, Ahmed F, Alfouzan FA, Alzahrani NM, Ahmad J (2022) Chest X-ray classification for the detection of COVID-19 using deep learning techniques. Sensors 22(3):1211
39. Sanida T, Sideris A, Tsiktsiris D, Dasygenis M (2022) Lightweight neural network for COVID-19 detection from chest X-ray images implemented on an embedded system. Technologies 10(2):37
40. Basu A, Das S, Mullick SS, Das S (2023) Do pre-processing and class imbalance matter to the deep image classifiers for COVID-19 detection? An explainable analysis. IEEE Trans Artif Intell 4(2):229–241