Received: 15 October 2021 | Revised: 6 March 2022 | Accepted: 5 April 2022

DOI: 10.1111/coin.12526

ORIGINAL ARTICLE

Computational Intelligence WILEY

A deep learning-based approach for diagnosing COVID-19 on chest x-ray images, and a test study with clinical experts

Onur Sevli

Faculty of Engineering and Architecture, Computer Engineering Department, Burdur Mehmet Akif Ersoy University, Burdur, Turkey

Correspondence
Onur Sevli, Faculty of Engineering and Architecture, Computer Engineering Department, Burdur Mehmet Akif Ersoy University, 15030 Burdur, Turkey.
Email: onursevli@mehmetakif.edu.tr

Abstract

Pneumonia is among the common symptoms of the virus that causes COVID-19, which has turned into a worldwide pandemic. It is possible to diagnose pneumonia by examining chest radiographs. Chest x-ray (CXR) is a fast, low-cost, and practical method widely used in this field. The fact that different pathogens other than COVID-19 also cause pneumonia and the radiographic images of all are similar make it difficult to detect the source of the disease. In this study, automatic detection of COVID-19 cases over CXR images was tried to be performed using convolutional neural network (CNN), a deep learning technique. Classifications were carried out using six different architectures on the dataset consisting of 15,153 images of three different types: healthy, COVID-19, and other viral-induced pneumonia. In the classifications performed with five different state-of-art models, ResNet18, GoogLeNet, AlexNet, VGG16, and DenseNet161, and a minimal CNN architecture specific to this study, the most successful result was obtained with the ResNet18 architecture as 99.25% accuracy. Although the minimal CNN model developed for this study has a simpler structure, it was observed that it has a success to compete with more complex models. The performances of the models used in this study were compared with similar studies in the literature and it was revealed that they generally achieved higher

 © 2022 Wiley Periodicals LLC.

success. The model with the highest success was transformed into a test application, tested by 10 volunteer clinicians, and it was concluded that it provides 99.06% accuracy in practical use. This result reveals that the conducted study can play the role of a successful decision support system for experts.

KEYWORDS

chest x-ray analysis, convolutional neural network, COVID-19 diagnosis, pneumonia detection

1 | INTRODUCTION

Coronaviruses are common types of viruses discovered in the 21st century that cause respiratory diseases. Different types of coronaviruses, which cause disease in humans and animals, have caused outbreaks around the world in certain periods. After the SARS outbreak in 2002 and the MERS outbreak in 2012, COVID-19, which emerged in Wuhan, China at the end of 2019 and was caused by the SARS-CoV2 virus, caused the third major coronavirus outbreak of the 21st century. COVID-19, which continues to spread with different variants today, was declared a pandemic by the World Health Organization in March 2020. The disease, which shows a decreasing trend in certain periods, continues to be seen with the increasing number of cases and deaths because it cannot be controlled yet and is constantly mutated. In the last quarter of 2021, when this study was penned, there were approximately 220 million cases and 4.5 million COVID-19 deaths worldwide. In this period, the graph of the number of cases belonging to the top 10 countries with the highest cases worldwide is shown in Figure 1.

The family of coronaviruses is large, and some can only cause disease in humans or animals, while some can cause in both. The assumption that COVID-19 was first transmitted from an animal to a human is strong, and nowadays it can easily be transmitted from human to human

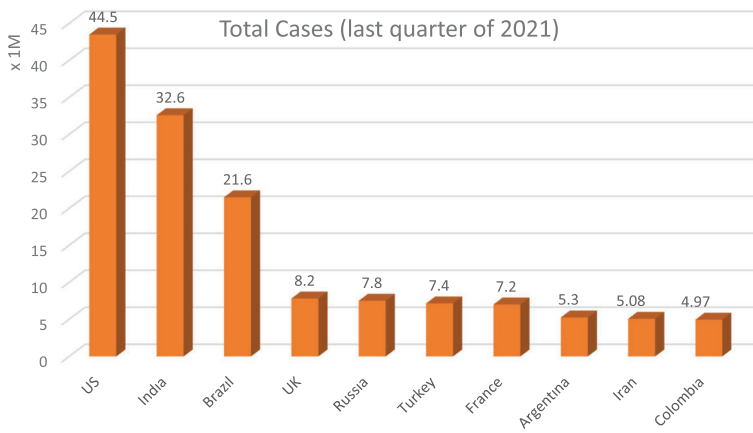


FIGURE 1 COVID-19 cases of the top 10 countries

through the respiratory tract. Common symptoms of the disease include fever, shortness of breath, cough, fatigue, loss of taste and smell. Clinical symptoms, positive pathogen tests, and radiological examination are used in the diagnosis of COVID-19.

COVID-19 has a rapidly spreading characteristic and therefore the critical step in combating the disease is the rapid detection and isolation of infected individuals. One of the gold standards used in the diagnosis of COVID-19 is RT-PCR (reverse transcription-polymerase chain reaction) test to detect SARS-CoV2 RNA in the swab. Although RT-PCR is considered a characteristic test for the diagnosis of COVID-19, it is a manual process and laborious. Despite the rapidly increasing cases of COVID-19, there have been problems in reaching adequate RT-PCR tests and obtaining results quickly. In addition, the sensitivity of the test is variable and it was reported to be relatively weak in some studies.[1] This situation necessitates resorting to alternative and more reliable methods in the diagnosis of the disease.

One of the alternative screening methods for the diagnosis of COVID-19 is radiographic examination. Pulmonary infection and pneumonia are mostly seen in patients infected with COVID-19, and lung radiography is used to diagnose this.[2] Pneumonia is inflammation of the lung tissue and can also be caused by different types of viruses, bacteria, and fungi. In addition to the SARS-CoV-2 virus that causes COVID-19 disease, different viruses such as Haemophilus influenza, rhinovirus, adenovirus also cause pneumonia, and similar chest radiographs caused by different viruses make it difficult to distinguish COVID-19 cases from the others.

Chest radiography (i.e., chest x-ray/CXR) and computed tomography (CT) are commonly used screening techniques to detect lung diseases. CXR is widely used because it is low cost, fast and practical, and exposes the patient to less radiation than CT.[3] According to data presented by the Centers for Disease Control and Prevention (CDC), before the COVID-19 pandemic in the United States, 1.3 million patients were diagnosed with pneumonia and 50,000 people died because of it, each year.[4] The number of cases and deaths in the ongoing COVID-19 pandemic is increasing day by day. The necessity of applying different treatment procedures for each disease agent requires the correct determination of the source of pneumonia. Therefore, accurate examination of CXR images is critical in diagnosing the disease and administering the right treatment.

Studies indicate that more accurate results can be obtained with the radiographic examination in the diagnosis of COVID-19.[5,6] CXR and CT images are examined by expert radiologists to look for evidence of the SARS-CoV2 virus. However, diagnosis on CXR images can be more complex than on CT, and interpretation of images requires special expertise.[7] Early diagnosis of the disease and timely implementation of the necessary treatment procedures are important in terms of reducing deaths and the rate of spread of the disease. The high demand for healthcare services in pandemic situations such as COVID-19 to combat the disease can cause bottlenecks. In addition, the fact that COVID-19 is a newly recognized virus in the world, the lack of sufficient experience in diagnosis, the debate, and uncertainties about the effectiveness of the tests applied, and the insufficient number of experts increase the need for intelligent decision support systems. Studies report that intelligent decision support systems help experts make more accurate decisions in detecting conditions that cause pneumonia, such as COVID-19.[7]

The main motivation of this study is to propose a high-performance solution to the problem of diagnosis of COVID-19 using CXR images. In this direction, five state-of-art CNN architectures, which are common in the literature, were used, and a CNN model with less complexity specific to this study was proposed. Demonstrating the performance of the proposed minimal CNN model compared to other deeper and more complex models constitutes a sub-research problem of the study. A CXR dataset consisting of 15,153 images and 3 classes (COVID-19, other viral pneumonia, and healthy) was used in this study. In the classifications performed with six different CNN

architectures, it was aimed to distinguish pneumonia, one of the common symptoms of COVID-19, from images of healthy lungs, as well as from images of pneumonia caused by other pathogens. Considering the difficulties experienced by specialists in distinguishing images of different types of pneumonia from chest radiographs, confirming that the source of the disease is COVID-19 is extremely important for the application of correct treatment procedures. The classification model with the highest accuracy was converted into an application and a pilot test study was conducted with 10 clinical experts. In this way, it was aimed to reveal the success of the model in practical use as well as the training and test successes.

2 | LITERATURE SURVEY

Recently, the number of artificial intelligence-supported solutions for different problems in the field of medicine has been increasing. Convolutional neural networks (CNNs), one of the deep learning techniques, a sub-branch of artificial intelligence, produce successful results in diagnosing cancer and other diseases through images.[8]

Numerous scientific studies have been conducted in different disciplines since the emergence of COVID-19. In addition, the number of studies on the diagnosis of pneumonia caused by COVID-19 with CNN-based systems over chest radiographs has increased especially in the last 2 years.

Jaiswal et al. carried out a study for automatic pneumonia diagnosis with a dataset consisting of 30,000 CXR images. Using ResNet101, a CNN architecture, they were able to reduce the error rate in pneumonia and normal image classification by up to 19.9%.[3] Chouhan et al., in their classification study using transfer learning on CXR images of 1346 healthy and 3883 pneumonia patients, obtained 92.86% accuracy with the AlexNet model and 94.23% accuracy with the ResNet18.[9] Chen et al. created a dataset consisting of a total of 35,355 CT images collected from 51 patients who were confirmed by laboratory tests to be COVID-19 positive and 55 control patients with other diseases. On this dataset, they achieved 95.24% accuracy in their classification study using the ResNet50 model for the detection of pneumonia caused by COVID-19.[10] In the study conducted by Liang and Zheng for the diagnosis of pediatric pneumonia, the dataset consisting of 5856 CXR images was classified with four different transfer learning methods and a developed CNN model with 49 convolution layers. The obtained accuracies were 74.2% for the VGG16 model, 81.9% for the DenseNet121, 85.3% for the InceptionV3, 87.8% for the Xception and 90.5% for the developed CNN model.[11] Wang et al. created a three-class dataset consisting of COVID-19, normal, and pneumonia cases with 13,975 CXR images collected from 13,870 patients. They achieved 92.6% test accuracy in their classification on this dataset using the CNN model they developed. They achieved 90.6% accuracy when they used the ResNet50 model.[12] Hemdan et al. performed classifications using seven different pre-trained CNN models on a dataset consisting of CXR images of 25 healthy and 25 COVID-19 patients. They achieved the highest accuracy of 90% with the VGG19 model and the lowest accuracy of 60% with the MobileNetV2.[13] Sethy and Behera achieved 98.66% accuracy using the ResNet50 model and SVM classifier in their classification study on a dataset consisting of 381 CXR images in three categories, normal, COVID-19, and other pneumonia.[14] Özkaya et al., in their study on the classification of CXR images infected with COVID-19 on a 300-sample dataset, compared the performances of three different pre-trained models and the proposed CNN model. They achieved 91.27% accuracy with the VGG16, 94.3% with the ResNet50, 91.53% with the GoogLeNet, and 95.60% with the CNN model they developed.[15] Nour et al. achieved 98.97% accuracy in their classification study on a 2905-sample CXR

dataset consisting of normal, COVID-19, and other pneumonia cases using a CNN model with five convolutional layers and SVM classifier.[16] In their study, Zhang et al. designed a ResNet network for COVID-19 detection from CXR images. They obtained 96% recall, 70.65% precision, and 95.18% AUC values in the classification they performed on the dataset consisting of 1078 samples with two classes, COVID-19 and healthy.[17] Apostolopoulos et al. performed a performance evaluation study for the detection of COVID-19 using five different models with the transfer learning method. They classified the dataset consisting of 1427 images in three different categories: COVID-19, bacterial pneumonia, and healthy. In that study, which was carried out using the VGG19, Inception, MobileNetV2, Xception, and ResNetV2 models, the highest accuracy value was obtained with the VGG19 model as 93.48%.[18] Ghoshal and Tucker achieved 92.9% accuracy in their classification study for the detection of COVID-19 using the Bayesian CNN model on a dataset consisting of 70 CXR images.[19] Uçar and Korkmaz used the pre-trained SqueezeNet model to classify the dataset containing 5310 CXR into three different classes as normal, pneumonia, and COVID-19. In that study, which they supported with the Bayesian optimization and data augmentation, they reached 98.26% accuracy.[20]

Ismael and Şengür performed classification studies on a dataset containing 180 COVID-19 and 200 healthy CXR images using the ResNet50, VGG, and the CNN model they developed and the SVM classifier. The highest accuracy was obtained as 94.7% when they used a linear kernel SVM with the ResNet50 model. They reached 91.6% accuracy with the CNN model they developed.[21] Maghdid et al. achieved 94.1% accuracy in their classification study using a pre-trained AlexNet model for the diagnosis of COVID-19 on datasets consisting of CXR images compiled from different sources.[22] Song et al. obtained 93% recall and 96% precision in their classification study using CT images of 88 COVID-19-induced, 101 bacterial-induced pneumonia, and 86 healthy individuals.[23] Wang et al. obtained 89.5% accuracy, 88% precision, and 87% recall in their study using the InceptionV3 model on 1065 CT images for the detection of COVID-19.[24] Jain et al. performed a classification for COVID-19 detection with three different pre-trained models using 6432 CXR images. In that study, they used the InceptionV3, Xception, and ResNetXt models and obtained the highest accuracy with the Xception model as 97.97%.[25] Nayak et al., in their classification study using eight different pre-trained models on a 500-sample dataset to distinguish CXR images of COVID-19 infected individuals from normal images, obtained the highest accuracy with the ResNet34 model as 98.33%. The ResNet50, VGG16, AlexNet, and SqueezeNet models provided 96.67%, the GoogLeNet 95.83%, the MobileNetV2 95% and the InceptionV3 model 92.5% accuracy.[26] Turkoglu carried out a study that enables the classification of COVID-19 and other viral pneumonia cases using CXR images with a model called COVIDetectioNet. Feature extraction was performed a the pre-trained AlexNet model on a dataset consisting of 6092 images, then the most effective features were selected with the Relief algorithm, and in the final stage, a classification was made with the SVM. The highest accuracy achieved in that study was reported as 99.18%.[27] In their study for COVID-19 detection using CXR images, Narin et al. compared the performance of various pre-trained models. Using the ResNet50 model, they reached 96.1% accuracy on a dataset of 3141 samples.[28] Khan et al. achieved 95.1% accuracy in their classification study using the 15-layer CNN model they developed on 1500 COVID-19 and 1300 normal lung CT images.[29] Shankar and Perumal proposed a new fusion model with deep learning-based feature extraction for COVID-19 diagnosis from CXR images. They obtained 94.08% accuracy, 94.85% precision, and 93.61% recall in the classification they performed using the InceptionV3 model on a total of 273 CXR images with 220 COVID-19, 27 normal, and 26 other pneumonia tags.[30] Hammoudi et al. achieved 90.54%, 93.92%, and 95.72% accuracy in their classification studies using the ResNet34, ResNet50, and DenseNet169 models on 5863 children's CXRs consisting of two classes,

normal and pneumonia.[31] Pham conducted a study comparing the performance of pre-trained deep learning models in diagnosing COVID-19. In the three-class classification performed on the dataset consisting of CXR images of 438 COVID-19, 438 normal, and 438 other viral pneumonia patients, the accuracy was 96.46% with the AlexNet, 96.20% with the GoogLeNet, and 96.25% with the SqueezeNet.[32] Aksoy and Salman obtained 98.02% accuracy in the classification study they carried out using CapsNet architecture on the dataset consisting of CXR images of 510 healthy and 509 COVID-19 patients.[33]

Ravi et al. mentioned that small-sized datasets were used in most of the deep learning-based studies carried out for the diagnosis of COVID-19 in the literature. They emphasized that although successful results were obtained in these studies, the generalizability was low due to the limited dataset. In their study, they performed classifications on a dataset consisting of 17,599 radiographic images of two types, COVID and non-COVID. They used the pre-trained EfficientNet model for feature extraction and then reached 99% accuracy in their classification with Random Forest and SVM.[34]

There are also different studies in the literature for the detection of tuberculosis,[35,36] lung cancer,[37] and various lung diseases[38] with the help of CNN on CXR and CT images. Liu et al.,[39] emphasizing that many people suffer from lung tumors and early diagnosis is important, suggested deep reinforcement learning models that have the potential to be used for detecting lung cancer. Capizzi et al.[40] proposed a model using a fuzzy system and neural network combination for the detection of lung nodules. They validated the model on CXR images with lung nodules and achieved 92.56% accuracy. Recent studies have focused more on detecting COVID-19 and other types of pneumonia.

In this study, six different classification processes were carried out on the dataset containing 15,153 CXR images, using five different state-of-art models and a minimal CNN model developed for the detection of normal, COVID-19, and other viral-induced pneumonia cases. The performances of the models were reported as accuracy, precision, recall, and F1-score metrics. A test application was developed using the model that provides the highest accuracy. This application was tested by 10 volunteer clinical experts with 12–23 years of experience in radiological image analysis. After the test, the success of the application in correctly classifying the CXR images was reported. The findings obtained confirmed that the application was successful in practice and revealed that it could be a useful decision support system for experts.

3 | MATERIAL AND METHOD

3.1 | Dataset

The dataset used in this study was created by Chowdhury et al. by combining different public datasets of pneumonia cases from COVID-19 and other pathogens.[41] While creating this dataset, the researchers used six different public sub-datasets compiled from similar studies in the literature. These sub-datasets are Italian Society of Medical and Interventional Radiology database,[42] Novel Coronavirus Dataset,[43] COVID-19 positive CXR images from different articles,[44] COVID-19 Chest imaging database,[45] RSNA Pneumonia Detection Challenge database,[38] and CXR Pneumonia Images database.[46] The created dataset contains posterior-to-anterior and anterior-to-posterior images that radiologists frequently use in the diagnosis of lung diseases. The main purpose of the researchers in creating this dataset is to provide a resource for artificial intelligence-supported studies that will detect COVID-19 quickly and accurately and distinguish

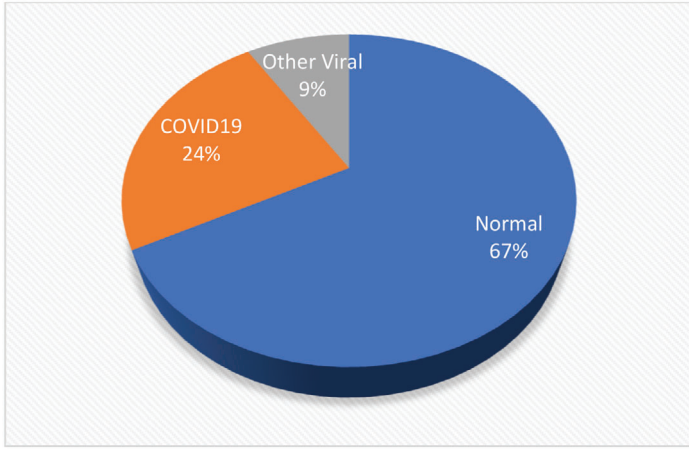


FIGURE 2 Class distributions of the dataset

it from other types of pneumonia through CXR images. For this reason, images of pneumonia caused by COVID-19 and other pathogens were added to the dataset, as well as healthy CXR images. The dataset includes a total of 15,153 CXR images, of which 10,192 are normal (healthy), 3616 are pneumonia caused by COVID-19, and 1345 are pneumonia caused by other viruses. In Figure 2 the graph shows the distribution of the classes in the dataset.

Sample images from three different classes in the dataset are given in Figure 3.

Opacifications can be seen in both COVID-19 and other viral pneumonia case images. This is not the case with healthy case images. Since the images of COVID-19 and other viral cases show very similar characteristics, its diagnosis requires special expertise.

3.2 | CNN models used for classification

CNN, which is a deep learning method, is a computer vision technique that provides automatic extraction of features from images. It is widely used in the analysis of medical images as in many fields. A typical CNN architecture includes input, convolution, activation, and pooling layers. The output of the CNN model is the features extracted from the image. The model is connected to a classification layer via the fully connected layer. The general structure of a CNN model is shown in Figure 4.

The pixels that make up the image are the inputs of the CNN model. In the convolution layer, the feature filters are stridden over the input pixels to obtain a subset of features. In the pooling layer, dimension reduction is performed. With filters called pools, the reduction process is applied to the input matrices by maximizing, minimizing, or averaging. Neural network models optionally include a dropout layer to handle overfitting. The dropout process increases the adaptability of the network to different situations by deactivating randomly selected neurons. The fully connected layer is involved in the transition to the classification layer. Here, the outputs of the CNN model are flattened and transferred to the model that will perform the classification.

Features such as the number of layers to be used in a CNN model and the number of components in each layer are tunable parameters and vary according to the process to be performed. One of the critical conditions for a neural network model to produce highly accurate results is

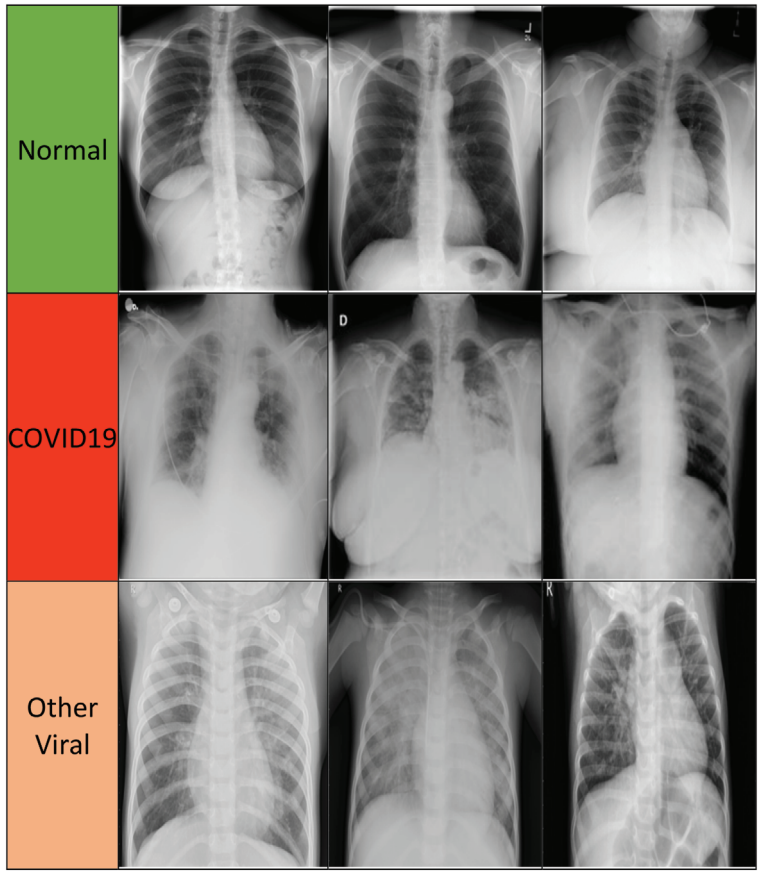


FIGURE 3 Sample images from the dataset

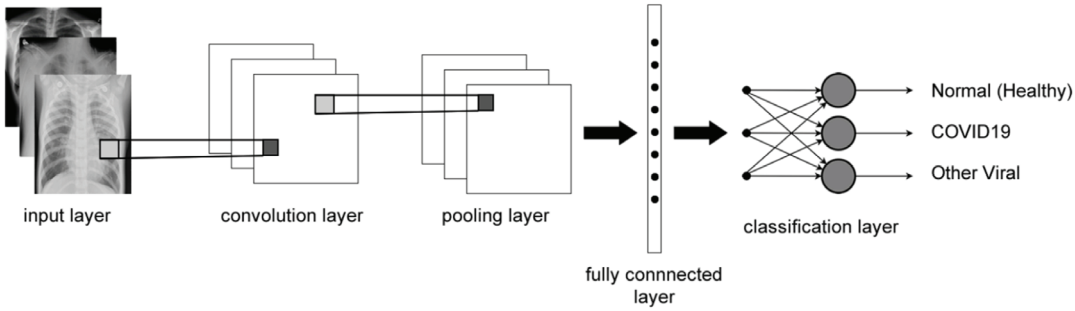


FIGURE 4 The general structure of a CNN model

to train it with a sufficient number and variety of data. State-of-art CNN models, which perform successful classifications on the ImageNet dataset consisting of approximately 14 million images, and give high accuracy results when adapted to different fields, are widely used in many studies. In this study, classifications were carried out using five of these models, which are widely used in the literature.

ResNet is a CNN model introduced by He et al. in 2015 and won first place in the ImageNet competition with a 3.5% error rate.[47] Unlike traditional sequential network models, ResNet has a structure based on microarchitecture modules. In theory, success is expected to increase as the number of layers in a model increase, but increasing the number of parameters makes training and optimization more difficult. During training, neurons with low activation become ineffective in the neural network and residues occur. ResNet is created by adding blocks that feed these residues to the next layers. There are variants of the ResNet built with a different number of weighted layers. ResNet18 used in this study is a type of ResNet model with 18 weighted layers.

Szegedy et al. created GoogLeNet by customizing the previously developed Inception architecture.[48] The model uses Inception modules that allow choosing from a large number of convolutional filter sizes in each block of the network. The Inception network stacks modules with max-pooling in certain situations to reduce the grid resolution. Consisting of 22 weighted layers, GoogLeNet was designed with efficiency and practicality in mind. It also enables successful extractions, especially on devices with limited computing resources.

AlexNet is a CNN model developed by Alex Krizhevsky that shows high success on the ImageNet dataset.[49] The model, consisting of 65,000 neurons and 60 million parameters, supports GPU acceleration to make training faster. The AlexNet model was among the top five with a 15.3% error rate in the ILSVRC-2012 image processing competition and outperformed the model that came after it by 10.9%.

VGG16 is a CNN model developed by Simonyan and Zisserman.[50] It was created by improving the AlexNet model by replacing filters with large kernel sizes with multiple 3×3 filters. The model, which includes 16 weighted layers, was among the top 5 models that provided the highest accuracy with 92.7% on the ImageNet dataset.

DenseNet, developed by Huang et al., is a CNN model in which all layers in a network are directly interconnected by weighted layers.[51] It attracted attention by providing high classification performance on huge datasets such as ImageNet and CIFAR-10. In this feedforward network, each layer takes the feature maps from all previous layers as input and forwards them to all subsequent layers. DenseNet alleviates the vanishing-gradient problem, strengthens feature propagation, and greatly reduces the number of parameters. DenseNet161 is a special form of DenseNet with 161 weighted layers.

The main advantages and disadvantages of the models mentioned above are given in Table 1 comparatively.

4 | EXPERIMENTAL STUDY AND FINDINGS

In this study, classifications were made on a dataset consisting of three different classes and 15,153 CXR images to distinguish COVID-19-infected cases from healthy and other viral-induced pneumonia cases. For this study, five different state-of-art CNN models (ResNet18, GoogLeNet, AlexNet, VGG16, DenseNet161) and a minimal CNN model specially created for this study were used. The features extracted with each CNN model were classified by a similarly structured neural network added to the end of the models. After the classification processes, the performances of the models were evaluated and compared in terms of different metrics (accuracy, precision, recall, and F1-score). The model with the highest success was saved with its calculated parameters and then connected to a test application developed through an application programming interface (API). This application was tested by 10 different volunteer clinical experts with

TABLE 1 Advantages and disadvantages of the state-of-art models used in this study

Model	Advantages	Disadvantages
ResNet18	• Increases the depth of the network instead of widening it, so fewer additional parameters are required. • The training process is faster. • Reduces the effects of the vanishing gradient problem.	• Increased complexity. • The model requires substantial Batch Normalization. • Jump links need to be added between different layers.
GoogLeNet	• It is trained faster than VGG. • The size of the trained model is slightly smaller.	• Although it has fewer parameters, the model structure is deeper and more complex.
AlexNet	• It has the ability to quickly subsample intermediate representations. • Extracts features better than the LeNet model. • Performs well on colored images.	• It has slightly less depth and therefore takes more effort to extract image features. • Longer training is required to achieve higher accuracy.
VGG16	• The model has enhanced depth that increases its success. • More layers with smaller kernels increase non-linearity.	• Network training takes longer. • The weights in the network architecture are large. • It is more prone to the vanishing gradient problem than ResNet.
DenseNet	• Alleviates the vanishing gradient problem. • Strengthens feature propagation. • Allows the reuse of features.	• Dense connections require more memory usage. • Requires more training time.

TABLE 2 Encoded labels of each class

Image label	Normal (healthy)	COVID-19	Other viral pneumonia
Encoded label	0	1	2

professional experience ranging from 12 to 23 years. The results obtained confirmed the success of the application in practical use.

4.1 | Data preprocessing

The dataset used in this study consists of two main parts. The first part is the metadata file containing the identifying information of the x-ray images. This file contains the path to the image and the label of the disease state. The original image labels were in text format and were numerically re-encoded from 0 to 2 before classification. Label codes for each class are given in Table 2.

The dataset consists of 10,192 normal (healthy), 3616 COVID-19, and 1345 other viral-induced pneumonia images and has an imbalanced class distribution. This imbalance in the dataset may cause the classifier model to predispose to the dominant class during training and reduce its success in predicting other classes. One of the solutions to avoid this overfitting situation is data

augmentation. For this reason, the sample numbers of minority classes were increased 3 times in COVID-19 case images and 7 times in other viral pneumonia images, bringing the sample numbers closer to the majority class. Data augmentation was performed by applying operations such as rotation, flipping, shifting, reflecting, and scaling on randomly selected images from the dataset. In this study, 1/255 scaling, 20% zoom, 20° rotation, 20% horizontal and vertical shifting, and horizontal flip were applied for data augmentation. In addition to these operations, 3% noise was added to the reproduced images. During noise adding, randomly selected pixels were replaced with 1 and 0 values.

The techniques and rates used for data augmentation in this study are purely experimental and can be tuned. However, these techniques and values were found to give good results for this study. Increasing the classification success with optimizations to be made on the values may be the subject of future studies.

4.2 | Developed CNN model

The general structure of the CNN model developed for this study to classify CXR images is shown in Figure 5.

The model can be considered as basic 4 blocks and consists of a total of 16 layers. The first 3 blocks constitute the CNN architecture that provides feature extraction from the images given as input. After feature extraction, the model is connected to the neural network, which will perform the classification in the fourth block.

The 3 channel 299×299 images are the inputs of the model. In the first two layers of the first block, there are convolution operations with 32 and 64 filters each of size 3×3. After these processes, maximum pooling is applied with a 2×2 pool size. Then 25% dropout is applied. In the first layer of the second block, a convolution operation is performed with 64 filters of 3×3 size. After that, maximum pooling with a 2×2 filter and then 25% dropout is applied. In the first layer of the third block, a convolution operation is applied with 128 filters of 3×3 size. This is followed by a maximum pooling operation with a pool size of 2×2. Then 25% dropout is applied.

ReLU activation function and the same padding technique are used in all convolution layers. The feature matrix obtained as a result of the third block is flattened and transferred to the fourth block where the classification process will be performed.

In the fourth block, there is a neural network (NN) model that will perform the classification using the features extracted by the CNN model in the previous section. The first three layers of this model have 256,128 and 64 neurons, respectively. ReLU is used as the activation function in all three layers. Then 50% dropout is applied. In the last part, there is a dense layer containing

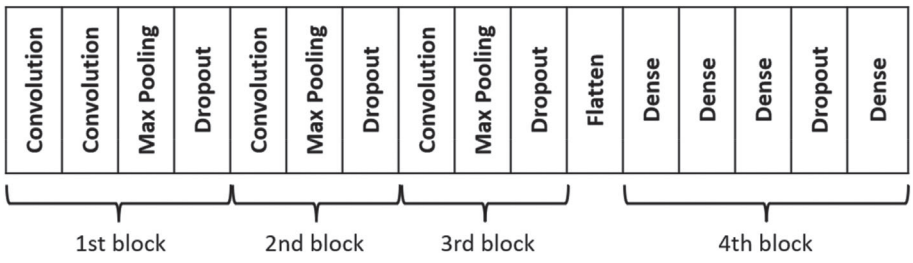


FIGURE 5 The general structure of the developed CNN model

as many neurons as the number of classes in the dataset. The softmax activation function is used for multiclass classification in this layer. This classification layer, which forms the fourth block, is also used for the other five state-of-art models in this study.

A summary of the entire model is given in Table 3. The CNN model consists of a total of 130.176 trainable parameters.

The proposed CNN model consists of 8 weighted layers, 7 hidden layers, and 1 output layer in addition to an input layer. The layer structure of the proposed model was obtained as a result of experimental studies. Adam optimizer was used in the training of the model. Grid search was used to obtain the values of the optimizer parameters used in this study. Batch size and epoch values were adjusted considering the balance between training time and model success. The hyperparameters and their values used in the training of this developed model are given in Table 4.

The main feature of the proposed model is that it is less complex than other models used. Due to its minimal structure, the number of parameters is less and the processing speed is high. It is effective in increasing the nonlinearity of the data as it contains layers with small kernels. Apart from this, the proposed model does not include an additional function, compared to other

TABLE 3 Summary of the model

	Layer type	Layer features	Output shape	Parameters
CNN model	Convolution 2D	Filters: 32, Kernel: 3 × 3, Activation: ReLU	(299, 299, 32)	896
	Convolution 2D	Filters: 64, Kernel: 3 × 3, Activation: ReLU	(299, 299, 64)	18,496
	Max pooling 2D	Pool size: 2 × 2	(149, 149, 64)	0
	Dropout	25%	(149, 149, 64)	0
	Convolution 2D	Filters: 64, Kernel: 3 × 3, Activation: ReLU	(149, 149, 64)	36,928
	Max pooling 2D	Pool size: 2 × 2	(74, 74, 64)	0
	Dropout	25%	(74, 74, 64)	0
	Convolution 2D	Filters: 128, Kernel: 3 × 3, Activation: ReLU	(74, 74, 128)	73,856
	Max pooling 2D	Pool size: 2 × 2	(37, 37, 128)	0
	Dropout	25%	(37, 37, 128)	0
	Flatten		(175,232)	0
Classifier NN	Dense	Neurons: 256, Activation: ReLU	(256)	40,141,056
	Dense	Neurons: 128, Activation: ReLU	(128)	32,896
	Dense	Neurons: 64, Activation: ReLU	(64)	8256
	Dropout	50%	(None, 64)	0
	Dense	Neurons: 3, Activation: Softmax	(None, 3)	195

TABLE 4 Hyperparameters of the model

Parameter	Value
Batch size	32
Number of epochs	200
Optimizer	Adam
Optimizer parameters	lr = 0.00001, beta1 = 0.9, verbose = 1, factor = 0.05

TABLE 5 Comparison of the proposed model with other models used in the study

Model	Number of weighted layers	Total parameters
ResNet18	18	11,511,784
GoogLeNet	22	6,797,700
AlexNet	8	62,378,344
VGG16	16	138,423,208
DenseNet161	161	28,681,000
Proposed model	8	130,176

models used in the study. The less complexity of the model provides a faster computational capability compared to deeper models. However, compared to some models with a deeper layer structure, the proposed model may need to be run longer to achieve an equivalent level of accuracy. Comparison of the proposed model with other models used in this study in terms of layer and number of parameters is given in Table 5.

The proposed model was tested for the COVID-19 diagnostic problem on the CXR images whose solution was sought in this study, and its performance in different studies is likely to vary.

4.3 | Performance measurements and comparisons

For the classification of the dataset, 5 different state-of-art models, namely ResNet18, GoogLeNet, AlexNet, VGG16, and DenseNet161, as well as the CNN model developed for this study were used. The NN model used for classification is the same in all models.

Each model was trained 200 epochs. Five-fold cross validation was applied. GPU acceleration was used during the training of the models. The performance of each training process was evaluated in terms of different metrics.

The main measure used to characterize model success is accuracy. Accuracy is calculated by dividing the number of correctly classified samples by the total number of samples. However, more detailed metrics are needed to evaluate the performance of the model, especially when working with imbalanced datasets. At this point, confusion matrices are used (Figure 6).

The confusion matrix provides an understanding of the extent to which the predictive model can distinguish the classes from each other. In the confusion matrix, samples with positive actual class and positively predicted by the model are called true positives (TP), and samples with negative actual class and negatively predicted by the model are called true negatives (TN). Samples that are predicted negatively by the model but do not have a negative actual class are called false

		Predicted class: +	Predicted class: -
Actual class	+	**TP** True Positives	**FN** False Negatives
	-	**FP** False Positives	**TN** True Negatives

FIGURE 6 Confusion matrix

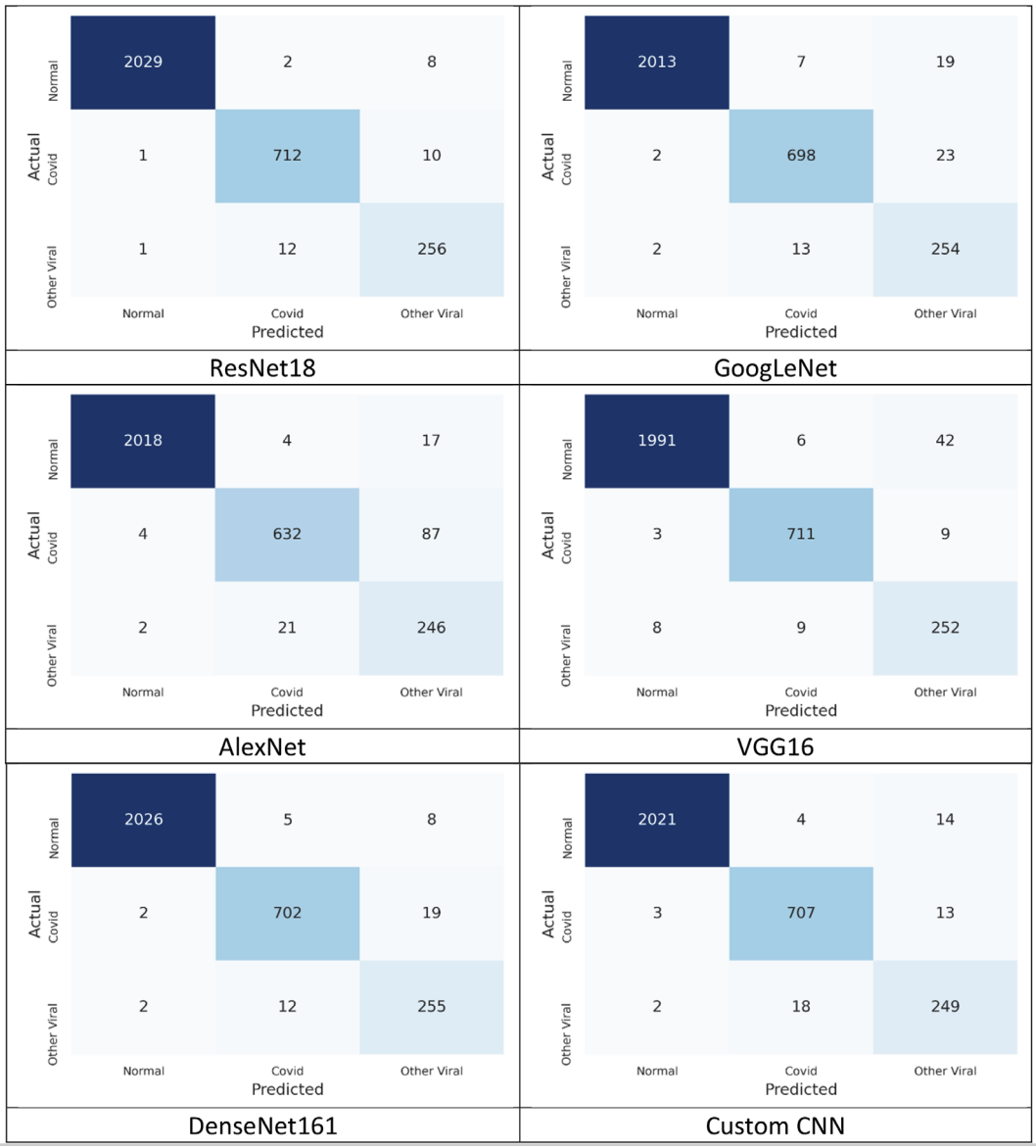
negatives (FN), and samples that are predicted positively by the model but do not have a positive actual class are called false positives (FP).

The confusion matrices obtained after training and testing of six different models used for classification performed in this study are listed in Table 6.

The confusion matrices given in Table 6 show the success of each model in distinguishing three different classes in the test dataset. The results obtained for each of these three different classes, normal (healthy), COVID-19, and other viral pneumonia, can be evaluated separately as follows: When the success of distinguishing normal cases is examined, it is seen that the ResNet18 model has the highest score by correctly detecting 2029 out of a total of 2039 normal test case samples. This model confused 23 of the normal cases with COVID-19 and 8 with other viral-induced cases. The DenseNet161 model, which ranked second in terms of success score, correctly classified 2026 normal cases, while confusing 5 cases with COVID-19 and 8 with other viral cases. The proposed model ranks third in terms of success in classifying normal cases. It correctly classified 2021 normal cases, while confusing 4 cases with COVID-19 and 14 cases with other viral cases. The fourth-ranked AlexNet model correctly classified 2018 normal cases, confused 4 cases with COVID-19 and 17 with other viral cases. The fifth-ranked GoogLeNet model classified 2013 normal cases as normal while confusing 7 of them with COVID-19 and 19 with other viral cases. The VGG16 model performed the lowest success in distinguishing normal cases. The model correctly classified 1991 normal cases, confused 6 of them with COVID-19 and 42 cases with other viral cases. When evaluated in terms of all models, it is seen that normal cases were often confused with other viral pneumonia cases. The rate of distinguishing normal cases from other cases was higher, as expected. Because while opacifications are more intense in COVID-19 and other viral case images, this is not the case in normal cases. In this respect, normal cases differ more than others. In addition, the fact that the number of normal case samples in the dataset was higher than the others had a positive effect on learning success. Another reason why normal cases are confused with other cases is that the features of images with poor quality and low clarity are likely to show similar characteristics with opacification.

It is seen that ResNet18 is the most successful model in correctly classifying COVID-19 cases. The model correctly classified 712 of a total of 723 COVID-19 test samples, confused one case with normal and 10 with other viral cases. While the VGG16 model, which ranked second, correctly detected 711 COVID-19 cases, it confused 3 of them with normal cases and 9 of them with other viral cases. The proposed model ranks third in terms of correct classification of COVID-19 cases. While classifying 707 cases correctly, it confused 3 of them with normal and 13 with other viral cases. The Densenet161 model, which ranks fourth, correctly classified 702 COVID-19 cases, while confusing 2 of them with normal and 19 with other viral cases. The fifth-ranked GoogLeNet

TABLE 6 Confusion matrices of the models



model correctly classified 698 cases, while confusing 2 of them with normal and 23 with other viral cases. The AlexNet model showed the lowest success in classifying COVID-19 cases. While classifying 632 samples correctly, it confused 4 of them with normal cases and 87 with other viral cases. When the confusion matrices are examined, it is seen that the models generally confused the images of COVID-19 cases with other viral pneumonia cases. In some other studies in the field of health, it was reported that human experts also have difficulties in distinguishing cases due to the similarity of CXR images of COVID-19 cases and other viral pneumonia cases not caused by COVID-19.[52] Because both COVID-19 and other viral pneumonia images have opacifications similar to each other. This situation makes it difficult to distinguish these two types of cases and causes them to be confused with each other more.

For other viral-induced cases, the ResNet18 model showed the highest success by correctly classifying 256 cases out of a total of 269 test images. It confused one case with normal and 12 cases with COVID-19. The DenseNet161 model, which ranked second, classified 255 cases correctly, confused 2 cases with normal and 12 cases with COVID-19. The third-ranked GoogLeNet model classified 254 images correctly, confusing 2 of them with normal and 13 with COVID-19. The fourth-ranked VGG16 model correctly classified 252 images, confusing 8 of them with normal and 9 of them with COVID-19. The proposed model ranked fifth with 249 correct classifications. The model confused 2 cases with normal and 18 cases with COVID-19. The lowest-performing AlexNet model classified 246 cases correctly while confusing 2 cases with normal and 21 cases with COVID-19. When evaluated in terms of all these results, it is seen that other viral origin cases were mostly confused with COVID-19 cases. This is because, as noted earlier, the opacifications in CXR images of these two cases show great similarities at times.

When all confusion matrices are examined, it is seen that ResNet18 is the most successful model in distinguishing classes from each other. However, all classification models, including this one, seem to confuse COVID-19 cases more with other viral cases. In addition, it is seen that normal cases and other viral cases are confused, albeit to a lesser extent. By visualizing the features of the images to be classified extracted by the CNN models, making it easier to the similarities and differences between the classes. t-SNE (t-Distributed Stochastic Neighbor Embedding) is one of the methods that enables the visualization of data through dimension reduction by creating representations of high-dimensional data in a lower-dimensional space.[53] t-SNE, a non-parametric, unsupervised machine learning method, uses a stepwise iterative approach to find a lower-dimensional representation of the original data while preserving knowledge about the local neighborhood of the data. The representation created by passing the features obtained from the CNN model, which showed the highest success of this study, to t-SNE is given in Figure 7. The parameters used to create the t-SNE representation and their values are: n_components = 2, perplexity = 10, early_exaggeration = 12, learning_rate = 200, n_iter = 5000, n_iter_without_progress = 300, min_grad_norm = 0.0000001, metric = 'euclidean', init = 'random', verbose = 0, method = 'barnes_hut', angle = 0.5, n_jobs = −1.

As can be seen in Figure 7, the image features of COVID-19 and other viral cases overlap more frequently in certain regions. This explains why classifiers are more likely to confuse cases of COVID-19 and other viral-induced. In addition, it is seen that the features of normal and other viral case images overlap. It can be said that this overlap is slightly less than the other. It also seems that COVID-19 and normal image features intersect in a much smaller region. These overlaps seen in the t-SNE feature representations reveal that the features of different case images are confused

FIGURE 7 t-SNE feature representation

TABLE 7 Performance metrics

Metric	Formula	Explanation
Accuracy	(TP + TN)/(TP + TN + FP + FN)	Refers to the overall success
Precision	TP/(TP + FP)	How accurate the positive predictions are
Recall	TP/(TP + FN)	Coverage of true positive samples
F1-score	2 * (Precision * Recall)/(Precision + Recall)	The harmonic mean of precision and recall

TABLE 8 Performance scores of models

Model	Accuracy	Precision	Recall	F1-score
ResNet18	0.9925	0.9713	0.9772	0.9742
DenseNet161	0.9894	0.9595	0.9708	0.9650
Custom CNN	0.9881	0.9565	0.9649	0.9606
GoogLeNet	0.9855	0.9428	0.9656	0.9535
VGG16	0.9831	0.9352	0.9656	0.9493
AlexNet	0.9703	0.8873	0.9261	0.9014

at certain regions and therefore misclassifications occur. In order to reduce misclassifications, it is concluded that models and training processes should be improved in future studies, and the number and variety of samples should be increased, especially in more frequently confused classes.

Different metrics obtained from the confusion matrices were used to examine the model performances in more detail. These metrics, their formulas, and explanations are given in Table 7.

The accuracy, precision, recall, and F1-score values obtained after the training and testing processes of the models are given in Table 8.

When the results in Table 8 are examined, it is seen that the accuracy of all models is over 97% and their success is high. The highest accuracy of 99.25% was obtained with the model that implements the ResNet18 architecture. The proposed CNN model has the third highest score with 98.81% accuracy. The AlexNet provided the lowest accuracy of 97.03%. The accuracy metric characterizes the overall success of the models. In this sense, the overall success order of the models can be made as follows: ResNet18 > DenseNet161 > Custom CNN > GoogLeNet > VGG16 > AlexNet.

The precision value is the hit rate on samples that the model classifies as positive. In terms of this value, the ResNet18 model has the highest score of 97.13%. The success of the ResNet18 model in distinguishing classes was discussed in detail when interpreting the confusion matrices. The precision value expressed here is the result of these correct detections. In terms of precision, the DenseNet161 ranks second with 95.95%. The proposed CNN model has the third highest score with 95.65%. The other three models are GoogLeNet with 94.28%, VGG16 with 93.52%, and AlexNet with 88.73%, respectively. In terms of precision score, model rankings are similar to accuracy score rankings.

The recall value shows how many of the actual positive values are correctly determined. In terms of recall values, the highest score was obtained with the ResNet18 model as 97.72%. The DenseNet161 follows this model with 97.08%. The GoogLeNet is third with 96.56% and the VGG16

is fourth with 96.56%. The proposed CNN model ranks fifth with 96.49%. The AlexNet model has the lowest score with 92.61%.

The balance of precision and recall values, which deals with the hit of positive values from two different aspects, the model's predictions, and the actual values, is expressed with the F1-score. In this sense, the F1-score can be thought of as a combined summary for precision and recall values. In terms of F1-score, the highest success belongs to the ResNet18 model with 97.42%. Other models are DenseNet161 with 96.50%, the recommended model with 96.06%, GoogLeNet with 95.35%, VGG16 with 94.93% and AlexNet with 90.14%, respectively.

When evaluated in terms of all the measurements obtained, it is seen that the most successful model is ResNet18. The model with the lowest success is AlexNet. Although the CNN model developed for this study has a minimal architecture compared to other models used, it is seen that it is more successful than the other three state-of-art models (GoogLeNet, VGG16, AlexNet) according to the performance measurements obtained. In addition, although fewer parameters were used for this model, the 98.81% accuracy rate obtained is very close to the two models (ResNet18 and DenseNet161) that provide higher accuracy. When the performance of the developed model is evaluated according to the simplicity of its architecture, it can be said that this model is highly successful.

The aim is to achieve as high accuracy as possible with the classifier models, but the fast response of the model is also expected. The minimal structure of a model allows it to show faster reflexes by using fewer parameters. The model proposed in this study has a more minimal structure compared to the others and shown high success. Therefore, the proposed model offers an ideal balance in terms of classification success and performance compared to other models.

In addition, it should be noted that the proposed model shown high success for the problem situation in this study, but has not yet been tested for different situations. It is expected that the proposed model will show a similar success to this study in studies such as radiographic image classification, tumor, and cancer detection on medical images. However, if the number of classes in the dataset and the classification complexity increase, the suggested minimal structure may be limited.

The comparison of the performance measurements obtained in this study with similar studies in the literature is given in Table 9. These studies performed operations on similar datasets consisting of different numbers of radiographic images. The data size used in this study is 15,153.

When the studies in Table 9 are examined, it is seen that the 99.25% accuracy obtained with the ResNet18 model in this study is higher than all the other studies in the table. The closest 99.18% accuracy rate to this study was obtained by Turkoglu et al.,[27] using less than half the number of images of this study and for a binary classification problem. Although the difference between the success rates of these two studies is small, the fewer images and fewer classes reduce the generalizability. In addition, since the number of classes is less, the accuracy is more likely to be higher since it will reduce the complexity of the problem. In addition to the higher accuracy obtained in this study, it can be said that its generalizability is higher due to the use of more images and classes.

The study conducted by Ravi et al.[34] has similar accuracy as this study but they used a two-class (COVID and non-COVID) dataset in their study. There are opacifications in COVID-19 radiographs, but this is not the case in normal case radiographs. Therefore, the distinction between COVID and non-COVID cases can be made relatively easily. However, in the dataset used in this study, there were images of pneumonia cases that were not caused by COVID-19 and caused by other pathogens as a third class. The similarity between COVID-19 and other viral pneumonia radiographs makes it difficult to distinguish between these two classes. Therefore, this study has

TABLE 9 Comparison of the results obtained in this study with the literature

Reference	Number of samples	Method	The highest Acc (%)
Chouhan et al.[9]	5229	AlexNet	92.86
		ResNet18	94.23
Chen et al.[10]	35,355	ResNet50	95.24
Liang and Zheng[11]	5856	Custom CNN	90.5
Wang et al.[12]	13,975	Custom CNN	92.6
Hemdan et al.[13]	50	VGG19	90
Sethy and Behera[14]	381	ResNet50 + SVM	98.66
Özkaya et al.[15]	300	Custom CNN	95.60
Nour et al.[16]	2905	Custom CNN + SVM	98.97
Apostolopoulos[18]	1427	Vgg19	93.48
Ghoshal and Tucker[19]	70	Bayesian CNN	92.9
Uçar and Korkmaz[20]	5310	SqueezeNet + Bayesian optimization	98.26
Ismael and Şengür[21]	380	ResNet50 + Linear kernel SVM	94.7
		Developed CNN	91.6
Wang et al.[24]	1065	InceptionV3	89.5
Jain et al.[25]	6432	Xception	97.97
Nayak et al.[26]	500	ResNet34	98.33
Turkoglu[27]	6092	Alexnet + Relief + SVM	99.18
Narin et al.[28]	3141	ResNet50	96.1
Khan et al.[29]	2800	Developed CNN	95.1
Shankar and Perumal[30]	273	InceptionV3	94.08
Hammoudi et al.[31]	5863	DenseNet169	95.72
Pham[32]	1314	AlexNet	96.46
Aksoy and Salman[33]	1019	CapsNet	98.02
Ravi et al.[34]	17,599	EfficientNet	99
This study	15,153	ResNet18	**99.25**
		DenseNet161	98.94
		Custom CNN	98.81
		GoogLeNet	98.55
		VGG16	98.31
		AlexNet	97.03

a more complex structure since it has more image classes and difficulty in distinguishing between COVID-19 and other viral pneumonia cases. Although the 99.25% accuracy value obtained in this study is close to the other study in terms of quantity, it is the result of the analysis performed on a more detailed and inclusive dataset.

The third highest accuracy rate of 98.97% in Table 9 was obtained by Nour et al.[16] for a three-class classification problem, similar to this study, and on a dataset consisting of 2905 images. Although the accuracy rate is close to this study, the number of images used is about one-fifth, and therefore its generalizability is lower. Similarly, the study conducted by Sethy and Behera[14] is a three-class classification problem and the obtained accuracy is 98.66%. However, the use of fewer images also reduces the generalizability.

The closest to this study in terms of data size is the classification made by Wang et al.[12] which used 13,975 images. The accuracy they obtained in the three-class classification with the CNN model they created was 92.6%, which is 6.65% lower than this study. With a higher number of 35,355 images from this study, Chen et al.[10] reached an accuracy of 95.24% in a binary classification problem (COVID-19 or healthy). Although the number of samples was more than twice the number of this study and the number of classes was less, 4% lower accuracy was obtained compared to this study. Table 9 shows that the ResNet18 model used in this study provides 5% higher accuracy in the binary classification study performed by Chouhan[9] on 5229 images and using the same architecture.

In addition to the ResNet18 model used in this study, which provides higher accuracy than all of the studies in Table 9, it is seen that the other five models used in this study show higher success than the other studies in general. While the average of success of the other studies in the Table 9 is 95.0025%, the average of the models used in this study is 98.4816%. From this point of view, it is seen that the AlexNet model, which has the lowest success in the study, provides a 2% higher accuracy than the average of other studies. The CNN model specific to this study showed higher success than the AlexNet, VGG16, and GoogLeNet models used in this study, as well as providing higher accuracy than most of the similar CNN models developed specifically for the study in the literature. The purpose of including this custom CNN model in this study was to be seen that a success that can compete with similar studies can be achieved even though it has a minimal architecture and fewer parameters compared to other models. The parameters used in the construction of this model are experimental and open to improvement by optimizing.

4.4 | Test application and expert evaluation

All the trained parameters of the classification model with ResNet18 architecture, which showed the highest success among six different CNN models used in this study, were saved to serve an end-user application. The saved model was linked to an easy-to-use test application via an application programming interface (API). This developed application can run smoothly on different platforms and screen resolutions.

Through this developed application, experts can select (select image) and analyze (analyze) test images on their own devices. After the selected image was classified by the recorded classifier model, the result of the analysis is displayed at the bottom of the page (Figure 8).

Test studies of this application were carried out with 10 volunteer clinical experts with professional experience ranging from 12 to 23 years in radiological analysis. All images used in the testing process were anonymous and different from the images used in training the models. The use of different data samples in this testing process allows revealing the success of the model

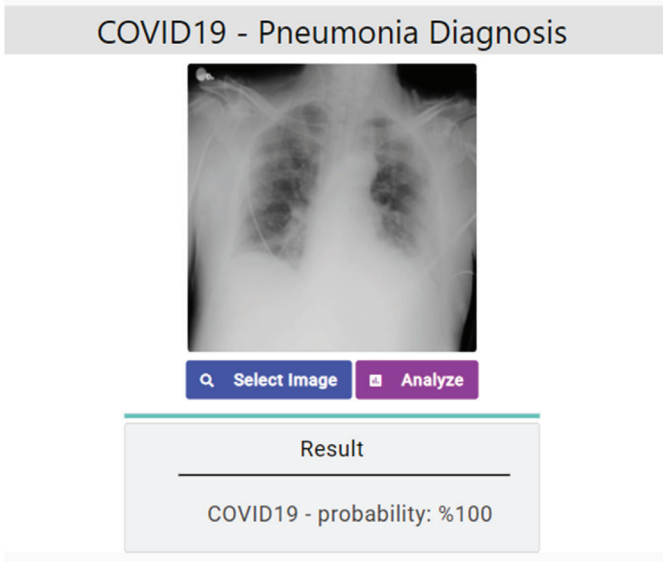


FIGURE 8 Test application

against different input situations. Working with different datasets makes the success of the model more generalizable. In this process, the experts tested the model with anonymous CXR images from their own archives or different sources. At this stage, the confidentiality of personal data was respected and no unauthorized or unethical data was used. This test study is a validation of the success of the classification model in practice.

Each independent expert used 150 different images from three different classes (COVID-19, normal, and other viral) in the test run. The test application analyzed the given image and reported the probability of the class it belongs to, and then this result was checked by the relevant expert.

The experts reported the results of the tests they performed in summary form. Each report includes information about how many images were correctly classified and how many were misclassified, and which classes were confused with each other. The results obtained in 10 different test processes using a total of 1500 different images are given in Table 10.

As a result of the test processes using 1500 images, it was confirmed by the experts that a total of 1486 images were correctly classified by the application. When the results were examined, it was seen that the pneumonia cases caused by COVID-19 were more confused with the cases of other viral-induced pneumonia. Based on their own experience, the experts stated that it was difficult for human experts to distinguish images that the classifier model confused with each other. There were also results in which other viral-induced pneumonia cases were classified as normal. According to the feedback from the experts, it was stated that viral cases mixed with normal cases mostly consisted of early stage and unclear images. In summary, in cases of outliers, the distinctiveness of the model in classification decreases. In order to enable the model to distinguish conflict situations more easily, it is necessary to increase the number of situations that are difficult to distinguish in the dataset used in the training of the model. When the independent results of each test are evaluated, the success of the overall test application is 99.06%. It is seen that the success of this test in practice coincides with the measurements obtained in the training of the model used.

TABLE 10 Test results

Expert	Number of correct predictions	Number of wrong predictions	Success rate (%)	Confusions
1	149	1	99.33	COVID-19—other viral (1)
2	150	0	100	-
3	148	2	98.66	COVID-19—other viral (2)
4	146	4	97.33	COVID-19—other viral (3) other viral—COVID-19 (1)
5	150	0	100	-
6	148	2	98.66	COVID-19—other viral (1) other viral—normal (1)
7	149	1	99.33	Other viral—normal (1)
8	147	3	98	COVID-19—other viral (3)
9	149	1	99.33	Other viral—COVID-19 (1)
10	150	0	100	-
	1486	14	99.06	

While the studies in the literature emphasize that using CXR images in the diagnosis of COVID-19 is practical and effective, they report that making an accurate diagnosis through radiographic images requires expertise, but there is still difficulty in reaching a sufficient number of experienced experts.[7] In addition, there are studies reporting that artificial intelligence-based decision support systems provide higher accuracy compared to human experts.[54] Instead of comparing the success of the system developed in this study with experts, it is emphasized that the system which was tested in practical use by experts can play the role of a successful decision support system.

5 | CONCLUSION

The timely and accurate diagnosis of COVID-19, which continues to spread rapidly nowadays, is of great importance to control the pandemic. Difficulties in reaching RT-PCR tests, which are accepted as a standard in the diagnosis of COVID-19, and hesitations about test sensitivity make it necessary to apply more reliable techniques. Analysis of CXR images is an important diagnostic in detecting pneumonia caused by COVID-19. However, the fact that there are also pneumonia cases that are not caused by COVID-19 and it is difficult to distinguish these cases from each other, as well as the fact that the number of experienced experts in this field is not yet sufficient, has accelerated the studies on different decision support systems. In this study, classifications were performed on the dataset consisting of 15,153 CXR images of COVID-19, healthy, and other viral pneumonia cases, using six different CNN architectures. The highest success rate of 99.25% was obtained with the classification model using the ResNet18 architecture. In addition, the minimal structured CNN model developed specifically for this study showed a success very close to the highest value, with an accuracy of 98.81%. The model with the highest success was transformed into a test application and tested by 10 volunteer clinical experts. In the tests performed, the

success of the system in practice was found to be 99.06%. This result shows that this study can play the role of a successful decision support system for experts.

It is possible to increase the success of diagnosis by increasing the number and variety of data used in the training of models. It is also possible to increase the success with parameter optimizations of the models. In future works, it is aimed to obtain more generalizable results by disseminating the application tests.

ACKNOWLEDGMENT

I would like to thank the researchers who shared the data set used in this study and the experts who contributed to the application testing process.

CONFLICT OF INTEREST

The author has no conflicts of interest to declare that are relevant to the content of this article.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/tawsifurrahman/covid19-radiography-database, reference number 39.

ORCID

Onur Sevli https://orcid.org/0000-0002-8933-8395

REFERENCES

1. Fang Y, Zhang H, Xie J, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology*. 2020;296(2):E115-E117. doi:10.1148/radiol.2020200432
2. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet*. 2020;395(10223): 507-513.
3. Jaiswal AK, Tiwari P, Kumar S, Gupta D, Khanna A, Rodrigues JJPC. Identifying pneumonia in chest X-rays: a deep learning approach. *Measurement*. 2019;145:511-518. doi:10.1016/j.measurement.2019.05.076
4. CDC. Pneumonia. *Centers for Disease Control and Prevention*; October 28, 2020. Accessed August 11, 2021. https://www.cdc.gov/dotw/pneumonia
5. Guan H. Clinical and thin-section CT features of patients with the COVID-19 in Wuhan. *Radiol Pract*. 2020;35:125-130.
6. Ng M-Y, Lee EYP, Yang J, et al. Imaging profile of the COVID-19 infection: radiologic findings and literature review. *Radiol Cardiothorac Imaging*. 2020;2(1):e200034.
7. Antin B, Kravitz J, Martayan E. Detecting pneumonia in chest X-rays with supervised learning. Semantic-scholar Org; 2017.
8. Kallianos K, Mongan J, Antani S, et al. How far have we come? Artificial intelligence for chest radiograph interpretation. *Clin Radiol*. 2019;74(5):338-345. doi:10.1016/j.crad.2018.12.015
9. Chouhan V, Singh SK, Khamparia A, et al. A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Appl Sci*. 2020;10(2):559. doi:10.3390/app10020559
10. Chen J, Wu L, Zhang J, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. *Sci Rep*. 2020;10(1):1-11.
11. Liang G, Zheng L. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Comput Methods Programs Biomed*. 2020;187:104964. doi:10.1016/j.cmpb.2019.06.023
12. Wang L, Lin ZQ, Wong A. COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci Rep*. 2020;10:1-12.
13. Hemdan EED, Shouman MA, Karar ME. Covidx-net: a framework of deep learning classifiers to diagnose COVID-19 in X-ray images. ArXiv Prepr. ArXiv200311055; 2020.
14. Sethy PK, Behera SK. Detection of coronavirus disease (Covid-19) based on deep features; 2020.

15. Özkaya U, Öztürk Ş, Barstugan M. Coronavirus (COVID-19) classification using deep features fusion and ranking technique. In: Hassanien A-E, Dey N, Elghamrawy S, eds. *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach*. Springer; 2020:281-295.
16. Nour M, Cömert Z, Polat K. A novel medical diagnosis model for COVID-19 infection detection based on deep features and Bayesian optimization. *Appl Soft Comput*. 2020;97:106580. doi:10.1016/j.asoc.2020.106580
17. Zhang J, Xie Y, Li Y, Shen C, Xia Y. COVID-19 screening on chest X-ray images using deep learning based anomaly detection. ArXiv Prepr. ArXiv200312338; Vol. 27, 2020.
18. Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Phys Eng Sci Med*. 2020;43(2):635-640.
19. Ghoshal B, Tucker A. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. ArXiv Prepr. ArXiv200310769; 2020.
20. Ucar F, Korkmaz D. COVIDiagnosis-net: deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. *Med Hypotheses*. 2020;140:109761.
21. Ismael AM, Şengür A. Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert Syst Appl*. 2021;164:114054. doi:10.1016/j.eswa.2020.114054
22. Maghdid HS, Asaad AT, Ghafoor KZ, Sadiq AS, Mirjalili S, Khan MK. Diagnosing COVID-19 pneumonia from X-ray and CT images using deep learning and transfer learning algorithms. In: Agaian SS, Asari VK, DelMarco SP, Jassim SA, eds. *Multimodal Image Exploitation and Learning*. SPIE; Vol 11734; 2021:117340E.
23. Song Y, Zheng S, Li L, et al. Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. *IEEE/ACM Trans Comput Biol Bioinform*. 2021;18:2775-2780. doi:10.1109/TCBB.2021.3065361
24. Wang S, Kang B, Ma J, et al. A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). *Eur Radiol*. 2021;31:1-9.
25. Jain R, Gupta M, Taneja S, Hemanth DJ. Deep learning based detection and analysis of COVID-19 on chest X-ray images. *Appl Intell*. 2021;51(3):1690-1700.
26. Nayak SR, Nayak DR, Sinha U, Arora V, Pachori RB. Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: a comprehensive study. *Biomed Signal Process Control*. 2021;64:102365.
27. Turkoglu M. COVIDetectioNet: COVID-19 diagnosis system based on X-ray images using features selected from pre-learned deep features ensemble. *Appl Intell*. 2021;51(3):1213-1226.
28. Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Anal Appl*. 2021;24:1-14.
29. Khan MA, Kadry S, Zhang YD, et al. Prediction of COVID-19 - pneumonia based on selected deep features and one class kernel extreme learning machine. *Comput Electr Eng*. 2021;90:106960. doi:10.1016/j.compeleceng.2020.106960
30. Shankar K, Perumal E. A novel hand-crafted with deep learning features based fusion model for COVID-19 diagnosis and classification using chest X-ray images. *Complex Intell Syst*. 2021;7(3):1277-1293.
31. Hammoudi K, Benhabiles H, Melkemi M, et al. Deep learning on chest X-ray images to detect and evaluate pneumonia cases at the era of COVID-19. *J Med Syst*. 2021;45(7):1-10.
32. Pham TD. Classification of COVID-19 chest X-rays with deep learning: new models or fine tuning? *Health Inf Sci Syst*. 2021;9(1):1-11.
33. Aksoy B, Salman OKM. Detection of COVID-19 disease in chest X-ray images with capsul networks: application with cloud computing. *J Exp Theor Artif Intell*. 2021;33(3):527-541. doi:10.1080/0952813X.2021.1908431
34. Ravi V, Narasimhan H, Chakraborty C, Pham TD. Deep learning-based meta-classifier approach for COVID-19 classification using CT scan and chest X-ray images. *Multimed Syst*. 2021;27(3):1-15. doi:10.1007/s00530-021-00826-1
35. Souza JC, Diniz JOB, Ferreira JL, da Silva GLF, Silva AC, de Paiva AC. An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks. *Comput Methods Programs Biomed*. 2019;177:285-296.
36. Liu C Cao Y, Alcantara M, Liu B, Brunette M, Peinado J, Curioso W. TX-CNN: detecting tuberculosis in chest X-ray images using convolutional neural network. Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP); 2017:2314-2318. doi: 10.1109/ICIP.2017.8296695

37. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*. 2019;290(1):218-228. doi:10.1148/radiol.2018180237
38. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017:2097-2106.
39. Liu Z, Yao C, Yu H, Wu T. Deep reinforcement learning with its application for lung cancer detection in medical internet of things. *Future Gener Comput Syst*. 2019;97:1-9. doi:10.1016/j.future.2019.02.068
40. Capizzi G, Sciuto GL, Napoli C, Połap D, Woźniak M. Small lung nodules detection based on fuzzy-logic and probabilistic neural network with bioinspired reinforcement learning. *IEEE Trans Fuzzy Syst*. 2020;28(6):1178-1189. doi:10.1109/TFUZZ.2019.2952831
41. Chowdhury MEH, Rahman T, Khandakar A, et al. Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access*. 2020;8:132665-132676. doi:10.1109/ACCESS.2020.3010287
42. COVID-19 DATABASE – SIRM; Accessed May 03, 2021. https://sirm.org/category/covid-19/
43. Cohen JP. COVID-chestxray database; 2019. Accessed May 03, 2021. [Online]. https://github.com/ieee8023/covid-chestxray-dataset
44. Rahman T, Khandakar A, Qiblawey Y, et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Comput Biol Med*. 2021;132:104319. doi:10.1016/j.compbiomed.2021.104319
45. C. Imaging. This is a thread of COVID-19 CXR (all SARS-CoV-2 PCR+) from my hospital (Spain). I hope it could help; 2020.
46. Chest X-ray images (Pneumonia). Accessed May 03, 2021. https://kaggle.com/paultimothymooney/chest-xray-pneumonia
47. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition; 2015.
48. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015:1-9.
49. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60(6):84-90. doi:10.1145/3065386
50. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ArXiv Prepr. ArXiv14091556, 2014.
51. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks; 2018.
52. Yin Z, Kang Z, Yang D, Ding S, Luo H, Xiao E. A comparison of clinical and chest CT findings in patients with influenza a (H1N1) virus infection and coronavirus disease (COVID-19). *Am J Roentgenol*. 2020;215(5):1065-1071.
53. Ravi V, Narasimhan H, Pham TD. EfficientNet-based convolutional neural networks for tuberculosis classification. In: Pham TD, Yan H, Ashraf MW, Sjöberg F, eds. *Advances in Artificial Intelligence, Computation, and Data Science*. Springer; 2021:227-244.
54. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med*. 2018;15(11):e1002686.

How to cite this article: Sevli O. A deep learning-based approach for diagnosing COVID-19 on chest x-ray images, and a test study with clinical experts. *Computational Intelligence*. 2022;1-25. doi: 10.1111/coin.12526