Applied Intelligence
https://doi.org/10.1007/s10489-021-02945-8

Decision and feature level fusion of deep features extracted from public COVID-19 data-sets

Hamza Osman Ilhan[1] · **Gorkem Serbes**[1] · **Nizamettin Aydin**[1]

Accepted: 19 October 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
The Coronavirus disease (COVID-19), which is an infectious pulmonary disorder, has affected millions of people and has been declared as a global pandemic by the WHO. Due to highly contagious nature of COVID-19 and its high possibility of causing severe conditions in the patients, the development of rapid and accurate diagnostic tools have gained importance. The real-time reverse transcription-polymerize chain reaction (RT-PCR) is used to detect the presence of Coronavirus RNA by using the mucus and saliva mixture samples taken by the nasopharyngeal swab technique. But, RT-PCR suffers from having low-sensitivity especially in the early stage. Therefore, the usage of chest radiography has been increasing in the early diagnosis of COVID-19 due to its fast imaging speed, significantly low cost and low dosage exposure of radiation. In our study, a computer-aided diagnosis system for X-ray images based on convolutional neural networks (CNNs) and ensemble learning idea, which can be used by radiologists as a supporting tool in COVID-19 detection, has been proposed. Deep feature sets extracted by using seven CNN architectures were concatenated for feature level fusion and fed to multiple classifiers in terms of decision level fusion idea with the aim of discriminating COVID-19, pneumonia and no-finding classes. In the decision level fusion idea, a majority voting scheme was applied to the resultant decisions of classifiers. The obtained accuracy values and confusion matrix based evaluation criteria were presented for three progressively created data-sets. The aspects of the proposed method that are superior to existing COVID-19 detection studies have been discussed and the fusion performance of proposed approach was validated visually by using Class Activation Mapping technique. The experimental results show that the proposed approach has attained high COVID-19 detection performance that was proven by its comparable accuracy and superior precision/recall values with the existing studies.

Keywords COVID-19 · Convolutional neural networks · Support vector machines · Feature level fusion · Decision level fusion · Ensemble learning · Class activation mapping · Transfer learning · Multistage learning

1 Introduction

The coronavirus disease 2019 (COVID-19) is a respiratory disorder, which may have varying severity respiratory symptoms from the common cold to fatal pneumonia. COVID-19 is caused by a novel coronavirus known as the severe acute respiratory syndrome coronavirus 2 (SARS-CoV2). SARS-CoV2 has very high contagious nature with a 1-14 days long incubation period. Some of the carriers may not show any symptoms while a significant amount of the patients may have minor symptoms such as dry-cough, sore throat, headache, fatigue, and sputum production. However, the virus can be fatal if the immune system of the patient is weak [72]. The conditions seen in the severe and critical patients may be the pneumonia, acute respiratory distress syndrome, pulmonary edema or multiple organ failure [15, 17]. In [19], it was stated that approximately 14% of the COVID-19 patients have experienced severe conditions such as the dyspnea, while 5% of the patients were in critical condition including respiratory failure, septic shock, or multiple organ dysfunction. Early diagnosis of the COVID-19 and the application of successful treatment is the key factor to reduce the complication and mortality in patients having underlying medical conditions such as hypertension, diabetes, cardiovascular disease and asthma [19, 26, 32, 53]. Another important factor

This article belongs to the Topical Collection: *Artificial Intelligence Applications for COVID-19, Detection, Control, Prediction, and Diagnosis*

Hamza Osman Ilhan and Gorkem Serbes contributed Equally to the work

✉ Hamza Osman Ilhan
hoilhan@yildiz.edu.tr

[1] Yildiz Technical University, Istanbul, Turkey

Published online: 30 October 2021

related with the COVID-19 is the transmission mechanism of the SARS-CoV2. The primary propagation mechanism of the SARS-CoV2 has been identified as the spread of respiratory droplets through sneezing and coughing, which have the potential to cover a distance up to 1.8 meters [13]. This highly contagious nature of the SARS-CoV2 puts any person, who has a close contact history with the patient, in a very high risk. Although, the primary source of the SARS-CoV2 transmission has been identified as the symptomatic people, asymptomatic people can also have a possibility to be a risk factor [13]. The higher risk of getting severe COVID-19 disease for the patients having existing medical conditions and being over age 60 years, and the high potential of fast propagation risk of COVID-19 results in a significant need for the fast and accurate diagnosis tools.

As the most common test technique to diagnose COVID-19, the real-time reverse transcription-polymerase chain reaction (RT-PCR) is used to detect the presence of viral RNA. In this method, a sample including a mixture of mucus and saliva is taken by using the nasopharyngeal swab technique for being assessed for virus existence. However, the RT-PCR suffers from having low-sensitivity especially in the early stage [27, 56] and it was mentioned in [93] that the chest radiography has performed very well in the early diagnosis of COVID-19. Therefore, it is believed that complementing the nucleic acid testing with chest radiography based diagnosis has promising potential in the early detection of COVID-19 [41]. Regarding the chest radiography techniques, X-rays and Computer tomography (CT) scans are the most commonly used imaging methods to diagnose the thoracic abnormalities. Although the CT scan can provide finer details of the 3D anatomy of human body, X-rays are more convenient to differentiate between viral and non-viral pneumonia due to its fast imaging speed, significantly low cost and low dosage exposing of radiation [77]. Furthermore, in [39], the most common manifestations and patterns of lung abnormality on portable chest radiography (CXR) in COVID-19 were described and it was mentioned that the CXR will likely be the most commonly utilized method for diagnosis and follow up of COVID-19 because of the infection control issues related to patient transport to CT suites, the problems experienced in CT room decontamination, and lack of CT availability in parts of the world. In [8], an experimental CXR scoring system, which was tested on hospitalized patients with COVID-19 pneumonia, was presented to quantify and monitor the severity and progression of disease. The authors found that the inter-observer agreement of the developed system was very good and the CXR based scoring is a promising tool for predicting mortality in hospitalized patients with SARS-CoV2 infection. In the light of the advantages of X-ray imaging over CT scan in the diagnosis and monitoring of COVID-19, we focus on developing a X-ray imaging based automated system which has the ability of differentiating viral pneumonia (COVID-19) from non-viral pneumonia and normal controls (No findings).

Computer-aided diagnosis (CAD) has been successfully used as a supporting tool for the diagnosis process of radiologists since 1980s [22]. The CAD systems are mostly developed as a complementary decision making approach to the diagnosis of physicians due to their advantages such as being reproducible and having the ability of detecting subtle changes that cannot be observed by the visual inspection. With respect to the usage of X-ray imaging based CAD systems in the diagnosis of thoracic diseases, the recent advances in deep learning have led to breakthrough improvements in the discrimination of viral and non-viral pneumonia. In [43], a diagnostic tool, which is based on a deep-learning framework for diagnosis of pediatric pneumonia using chest X-ray images, was proposed. In [71], the performance of customized convolutional neural networks (CNNs) to differentiate between bacterial and viral pneumonia types in pediatric CXRs was presented. Additionally, various deep learning approaches were successfully employed to diagnose pneumonia and other pathologies in [5, 40, 95]. In order to detect COVID-19 samples by using X-rays, a deep learning architecture, which employs depthwise convolutions with varying dilation rates to incorporate local and global features extracted from diversified receptive fields, was presented in [60]. In [87], various deep learning models were utilized for feature extraction and the obtained feature sets were processed using the Social Mimic optimization method. Later, the modified deep features were given to support vector machines (SVMs) with the aim of COVID-19 detection. In [69], a concatenated neural network, which is based on Xception and ResNet50V2 networks for classifying the chest X-ray images into three categories of normal, pneumonia, and COVID-19, was presented in an unbalanced data-set configuration. In [65], a patch-based CNN approach with a relatively small number of trainable parameters were given for COVID-19 diagnosis. In this method, random patches were cropped from the X-ray images and the final classification result was obtained by majority voting from inference results at multiple patch locations. In [25], a comparative individual analysis of the recent deep learning models including VGG16, VGG19, DenseNet201, InceptionResNetV2, InceptionV3, Resnet50, and MobileNetV2 was presented in the detection and classification of COVID-19. An Auxiliary Classifier Generative Adversarial Network based model was employed in [90] for generating synthetic chest X-ray CXR images to avoid overfitting and increase the generalization capability of employed CNNs. In [66], an end-to-end deep learning architecture, which was an enhanced version of the Darknet-19 model, was employed for the multi-class classification (COVID vs. No-Findings vs. Pneumonia).

Although previous studies have shed some lights on the deep learning-based diagnosis by using X-ray images and significant improvement has been obtained, none of the previous works have been able to propose a complete solution to the COVID-19 detection problem. Additionally, the COVID-19 outbreak is recent and the content of the public X-ray imaging databases is still progressing. Due to this gradual increase in the number of COVID-19 images in the public databases, a need of developing new algorithms, which have generalization capability for new COVID-19 samples, has been raised. In this study, we propose a deep features based ensemble learning model, which uses feature and decision level fusion, in order to satisfy the aforementioned needs in COVID-19 diagnosis.

The main contributions of this study are summarized as follows:

- The proposed learning model was applied to progressively created three public COVID-19 databases in order to measure its generalization capability and reduce the biasing effect that can occur in unbalanced databases.
- The individual performance of seven powerful deep learning architectures including the Mobilenet, VGG16, ResNet50, ResNet101, NasNet, InceptionV3 and Xception were presented.
- The same seven deep learning models were employed as feature extractors and the obtained individual deep features were fed to non-linear kernel SVMs with the aim of COVID-19 detection.
- The extracted deep features by using individual CNNs were concatenated to form a single feature vector (feature level fusion) which was subsequently given to classifiers.
- The decisions of the individual classifiers were combined by employing the majority voting schema (decision level fusion).
- The experimental results have demonstrated the effectiveness and robustness of the proposed ensemble approach in epidemic screening by reaching high general accuracy values accompanied with high COVID-19 F1-scores, precision and recall values.

The rest of the study is organized as follows; Section 2 introduces materials and methods. Section 3 presents the experimental results and finally, Section 4 presents the discussion and conclusion.

2 Materials and methods

In this study, an ensemble of CNNs with the aid of decision and feature level fusion idea was proposed to solve the classification problem in X-ray images for COVID-19, No-Findings and Pneumonia classes. For doing that three public X-ray datasets were employed in the experiments and the generalization capability of the proposed approach has been proven. In the ensemble of CNNs, transfer learning layout of seven deep convolutional neural network (CNN) models, which were initially pre-trained by a huge image collection repository, the ImageNet, were utilized. The employed deep networks, whose individual classification performance were also given, were the MobilenetV2, VGG16, ResNet50, ResNet101, NasNet, InceptionV3 and Xception. In addition, the same seven deep networks were also employed as deep feature extractors and the obtained deep features were fused and the resultant concatenated feature vector was fed to non-linear kernel based SVMs to increase the discrimination performance.

2.1 Dataset information

In our study, three databases were constructed in a progressive way to measure the classification performance and generalization ability of the proposed approach by using the combinations of three publicly available data-sets. Firstly, the data-set that has been already used in [66] was employed as the baseline reference database and it is named as DB1. DB1 consists of 125 COVID-19 images, 500 pneumonia images and 500 normal (no-finding) images. The COVID-19 images of DB1 were taken from a public data-set, which is constantly updated by researchers [21]. The remaining 1000 non-COVID X-ray images were taken from the public ChestX-ray8 dataset [92] and the DB1 was finalized with 1125 X-ray images. Secondly, at the date of this study, 353 new COVID-19 samples, which have been appended to DB1 by researchers after the publication of [66], were added to DB1 to be able to compare our study with other state-of-art findings. This new database, which contains 1478 X-ray images in total, is named as DB2. Lastly, 113 new COVID-19 samples obtained from a different domain were added to DB2 to be able to create a more balanced data-set that would be more convenient to measure performance of the proposed method. The new 113 COVID-19 samples were taken from [1] resulting in the DB3, which contains 1591 X-ray images in total. In the experiments, 5-fold cross-validation technique was applied in order to validate the results over each created dataset as in [66], [60], [91]. For each fold, the whole image set was divided into training and testing sets with the ratio of 80% and 20%, respectively. In each repetition, a new model was trained by using randomly arranged 80% of data-set, while testing was evaluated with the remaining 20% of dataset. This cross-validation approach is then repeated 5 times and, as a result, each observation (sample image) is used for testing exactly once. A short summary of the constructed data-sets with the information of training and testing sizes for each fold is given in Table 1.

Table 1 The Image Distributions over Classes in Tested Datasets

Labels	DB1 [66]			DB2 [21]			DB3 [21]+[1]		
	Train Set	Test Set	Total	Train Set	Test Set	Total	Train Set	Test Set	Total
COVID-19	100	25	125	383	95	478	473	118	591
No Findings	400	100	500	400	100	500	400	100	500
Pneumonia	400	100	500	400	100	500	400	100	500
Total			1125			1478			1591

2.2 Employed deep learning architectures

The traditional machine learning approaches, which consist of sequential sub-steps such as pre-processing, feature extraction, feature reduction/selection and classification, require domain specific expertise in order to obtain satisfactory performance in medical image analysis. The spatial and frequency domain features are the most popular approaches to obtain discriminating information from the raw images. For example, the Scale-Invariant Feature Transform (SIFT) and Maximally Stable Extreme Regions (MSER) methods are used in literature [37, 38] as the spatial domain interest point extraction techniques and the interest points based features are employed in traditional learning models subsequently. Regarding the frequency domain feature extractors like short time Fourier Transform (STFT) and wavelet transform (WT), the parameter selection procedure makes them hard to implement and dependent to user experience. On the other hand, even if the training processing times of deep learners are relatively long, they are implemented in end-to-end architectures which have no need or having minimum need for extra pre-processing steps and optimum tuning of feature extractor parameters. In contrast, traditional machine learning methods are still highly error prone and inaccurate to be used in a sensitive decision making process. Therefore, in order to benefit from the aforementioned superiorities of deep learners, seven CNN models including the MobileNetV2, VGG16, ResNet50, ResNet101, NasNet, InceptionV3 and Xception, have been applied to three public databases with the aim of three-class (COVID, No-Findings, Pneumonia) discrimination of X-ray images in the proposed study.

2.2.1 MobileNetV2

Although higher accuracy values can be achieved by using deeper and larger networks, these networks do not ensure efficiency in terms of size and speed, making them inconvenient for mobile applications. However, the fast and accurate diagnosis of COVID-19 is critical in the current pandemic condition causing the small mobile deep learning solutions more preferable. The MobileNetV2 [76], as an improvement of MobileNetV1, can be a powerful and versatile solution for mobile diagnosis of COVID-19 due to its high performance proven in various application areas including medieval writer identification [18], detecting underwater live crabs [12], real-time crowd counting [29] and remote wave gauging [11]. The main characteristic of MobileNetV2 relies on the usage of depthwise separable convolutions in which two 1D convolutions with two kernels are used instead of employing 2D convolution with a single kernel. As a result, the training phase can be carried out by using fewer parameters and less memory that results in a small and efficient model.

2.2.2 VGG16

The VGG16 [82] is a pre-trained very large CNN that was invented by VGG (Visual Geometry Group) from University of Oxford. The VGG16 was the 1st runner-up of the ILSVR (ImageNet Large Scale Visual Recognition Competition) 2014 in the classification task. The VGG16 architecture uses simple 3×3 size kernels in convolutional layers and combine them in a sequence to emulate the effect of larger receptive fields. The implemented VGG16 architecture is composed of 13 convolutional layers followed by 3 fully connected layers. Despite the simplicity of the VGG16 architecture, its memory usage and computational cost is dramatically high due to the exponentially increasing kernels.

2.2.3 ResNet50 and ResNet101

The ResNet deep learning models [31], which have introduced the "skip connections" concept, are the sub-classes of CNNs. In ResNets, some of the convolutional layers are bypassed (the concept of "skip connections") at a time and the batch normalization is applied along with non-linearities (ReLU) [67]. In ResNet architectures, the "skip connections" enables to train much deeper networks and they give the network the option to simply copy the activations from ResNet block to ResNet block, preserving

information as data goes through the layers [59]. The two versions of ResNet family, the ResNet50 and ResNet101 having 49 and 100 convolutional layers respectively, were employed in the current proposed COVID-19 diagnosis approach as a classifier and deep feature extractor.

2.2.4 NasNet

As a relatively recent network, the NASNet [100], whose CNN architecture was designed by another neural network, outperformed the previous state-of-the-art on the ILSVRC 2012 dataset. The NASNet architecture was created by use of the Neural Architecture Search (NAS) framework providing an algorithm for finding optimal neural network architectures [20]. In this algorithm, a controller recurrent neural network creates architectures aimed to perform at a specific level for a particular task, and by trial and error learns to propose better and better models [59].

2.2.5 InceptionV3

In the InceptionV3 [84], the inception modules, which are repeatedly stacked together to form a large network, are employed as an alternative to sequentially ordered convolution layers. In the inception modules, an asymmetric convolution structure is obtained by using multiple filters of various sizes resulting in more and more abundant spatial features with increased diversity. The usage of inception modules not only causes significant decrements in the number of parameters, it also increases the recognition ability of the network by using multiple scale features [99].

2.2.6 Xception

As an improved version of inception architecture, the Xception [16] algorithm uses depthwise separable convolutions which enables more efficient use of model parameters. In the Xception, the standard inception modules are replaced with the depthwise separable convolutions (enhanced inception modules) that use the depth dimension (the number of channels) as well as the spatial information. The enhanced inception modules result in stronger features including the depth information.

2.3 Transfer learning

During the analysis of medical images by using Transfer Learning, the weights of a deep-net that have been learned in the training of a CNN on a main dataset (for example ImageNet [74]) are transferred to a second CNN, which is then re-trained on labelled samples of desired medical data set using pre-learned weights. The final training phase is named as "fine tuning"; in which the certain layers of pre-trained net can be frozen (the weights of these layers stay fixed) while the remaining layers can be fine-tuned to suit the classification problem, except the last fully connected layer.

In our study, the employed CNNs were applied to COVID-19 data-sets by using the Transfer Learning strategy in the light of literature findings. In [4], it was mentioned that the performance of knowledge transfer depends on the dissimilarity between the database on which a CNN is trained and the database to which the knowledge is to be transferred. The distance between the natural image databases, that are employed for knowledge transfer, and COVID-19 data-sets is considerable. However, recent studies show that there is a potential for having benefit from knowledge transfer in medical data sets. For instance, in [6], a pre-trained CNN was employed as a feature extractor with the aim of chest pathology identification. In [89], pre-trained CNN based features have shown improved performance as they were fused with traditional handcrafted features in a nodule detection system. In addition to their feature extractor usage, the knowledge transferred CNNs can also be employed as the main classification framework with fine-tuning. For example, in [85], it was shown that the fine-tuned CNNs have performed similarly or better than the CNNs trained from scratch. In this study, pre-trained weights from [51] were transferred in either a shallow tuning or deep tuning strategy in which the weights of few layers for the former and many layers for the latter were trained. The results highlighted that medical image analysis requires deep tuning of more layers in contrast to many other computer vision tasks. In another study, it was demonstrated that fine-tuning of pre-trained networks worked better compared to nets trained from scratch in the analysis of skin lesions [62]. Additionally, in [81] knowledge transfer from natural images was applied in thoraco-abdominal lymph node detection and interstitial lung disease classification resulting in higher performance than training the CNNs from scratch. Similarly, in [14], transfer learning strategy was employed to identify the fetal abdominal standard plane and the approach revealed improved capability of the algorithm to encode the complicated appearance of the abdominal plane. In our study, due to the aforementioned superiorities of fine-tuning strategy, seven CNNs, which have already been trained by natural image database (ImageNet), were fine-tuned to extract deep features by using the X-ray samples. Later, these deep features were employed in the classification of chest X-ray images with individual and ensemble learning models.

2.4 Decision and feature level fusion

In a pattern recognition system, the ultimate goal is the design of best possible classification model for a specific problem such as the COVID-19 detection by using X-ray images. Traditionally, various classification models that have different theories and methodologies are applied to a specific pattern recognition problem, and the best model in terms of performance metrics is chosen. However, it was observed that the sets of patterns misclassified by the various classifiers would not necessarily overlap, even if one of the models has yielded the best accuracy. Hence, different classifiers may be harnessed to improve the overall performance by using their possible complementary information about the patterns to be classified, when they are used in an ensemble scheme [49]. This type of ensemble learning scheme is called decision level fusion based learning, in which the individual decisions of different models are combined to derive a consensus decision instead of relying on a single decision-making model. The hard-level combination uses the individual outputs of each classifier after they are binarized by applying a threshold to the classifier output probabilities (estimates of a posteriori probability of the class) to map them into class labels [63]. As a member of hard-level combination, the majority voting strategy simply counts the votes received from each classifier and the class that has the largest number of votes is selected as the consensus decision.

As an additional fusion strategy, the feature level fusion, in which various sets of features obtained by different feature extractors are combined, has high potential to achieve better classification performance [30, 55, 75, 88]. Feature level fusion generally consists of the concatenation of various normalized feature subsets resulting in a single feature vector forming a complete representation of different views (deep features obtained by using various CNNs). Regarding the CNNs based feature level fusion studies, even if the various CNN models are based on different configurations (or architectures), the fine-tuning of these CNN models by using the same target database (COVID-19 database in our study) consisting of concatenated feature vectors, can provide complementary information [23, 70].

2.5 Proposed deep features based ensemble model

In this study, seven CNN models (the MobilenetV2, VGG16, ResNet50, ResNet101, NasNet, InceptionV3, and Xception) have been used as the main structure of proposed framework. During the development of proposed method, firstly, these seven CNN models have been employed as deep feature extractors as depicted in Fig. 1. As seen in Fig. 1, the three databases were fed to the individual CNNs, which have already been pre-trained by using ImageNet [74], with the aim of network specific deep feature extraction by using a 5-fold cross-validation scheme. The optimum hyperparameters were chosen by employing a batch-size, epoch, and learning rate analysis that was based on trial and error strategy. Accordingly, the number of training epochs was chosen as 50, while a batch-size of 16 was employed. The learning rate that controls the speed of convergence was set to 0.0001, when Stochastic Gradient Descent with momentum was used as the optimization technique.

2.5.1 Learning scheme without feature level fusion

Subsequent to the deep feature extraction phase, the obtained deep features were fed to a softmax classifier satisfying the end-to-end learning scheme of classical deep learning. The classical softmax layer of CNNs, which is the generalization of logistic sigmoid function with the ability of mapping deep-features onto probability values used as outputs in discrimination problems having three or more classes, is named as "softmax classifier" in our study. The softmax classifier [45, 58, 80] is employed to

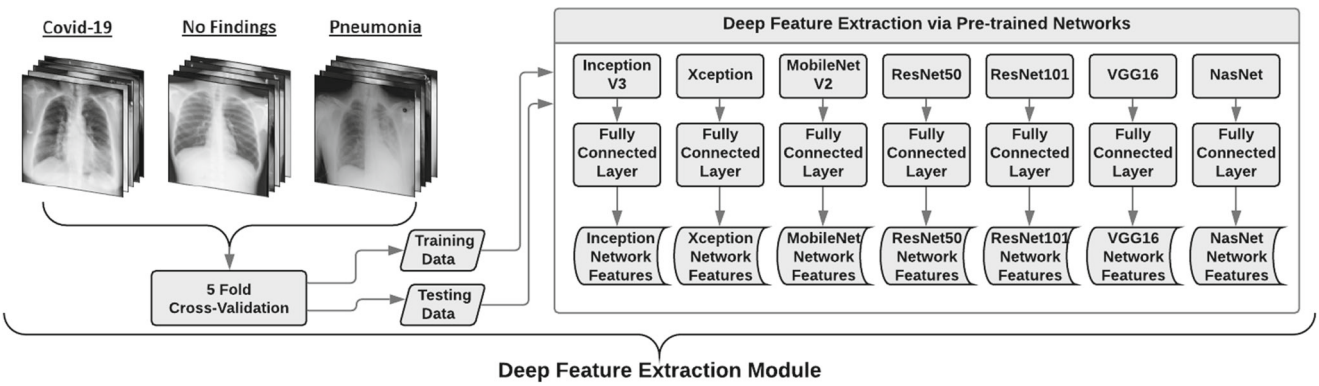


Fig. 1 The deep feature extraction module

measure the discriminating power of deep features obtained from the individual CNNs (equivalent to classical end-to-end learning with CNNs). The architecture that is followed to obtain individual CNN predictions is shown in Fig. 2 and the output of softmax classifier is named as "Individual Predictions" at the top-middle section.

In [36] and [86], it was mentioned that the CNNs, which are very good at learning invariant features, may show lower performance than the SVMs in classification. On the other hand, the SVMs are very successful at producing optimal decision surfaces from well behaved feature vectors, while having difficulties to represent the variances occurred in image features. Regarding the chest X-ray images used in our study, the areas that characterize the lung consolidation pattern may be located in various parts of the lung with changing size resulting in significant variances. Therefore, in addition to individual CNN based learning, a multistage model, in which the CNNs are employed to extract deep features that have potential to detect and recognize lung consolidation patterns, and non-linear SVMs that are trained by feeding the deep features learned by the CNNs, was presented and its performance was validated by using three databases. This multistage learning approach that uses CNNs and SVMs in a cascade connection has been successfully employed in various areas with the aim of classification performance improvement [28, 57, 83]. In this configuration, fully-connected activations of each CNN have been employed as feature extractors (given in Fig. 1) and the obtained deep feature vectors were fed to classifiers (softmax-classifier representing classical end-to-end CNN learning, SVM with RBF and Polynomial kernels) in a 5-fold validation scheme. Additionally, with the aim of performance improvement, the individual predictions obtained from classical end-to-end CNN learning (deep features that are fed to softmax classifier) and kernel based SVMs (deep features that are fed to SVMs) were fused by using the voting approach in accordance with the combinations given in Table 2. The SVM based learning configuration that uses the deep features, and the applied voting strategy was presented in Fig. 2.

2.5.2 Learning scheme with feature level fusion

Regarding the feature level fusion phase; the deep features extracted by individually employed CNNs were concatenated into a single fused feature vector directly without using any weight value. Subsequently, the fused feature vector was fed to the softmax classifier and also to the non-linear SVMs separately. After this, the individual predictions of the softmax classifier and SVMs were obtained as depicted in Fig. 3 when a concatenated feature vector was fed into. Furthermore, to benefit from the possible complementary behaviour of the learning models (softmax classifier, RBF and Polynomial SVMs), the obtained individual decisions were fused by using the majority voting. Thanks to the power of using feature and decision level fusion together, this final approach has given the best performance and was chosen as our proposed method. The detailed flowchart of proposed method including the deep feature extraction module, feature level fusion, multistage learning and decision level fusion can be seen in Fig. 3.

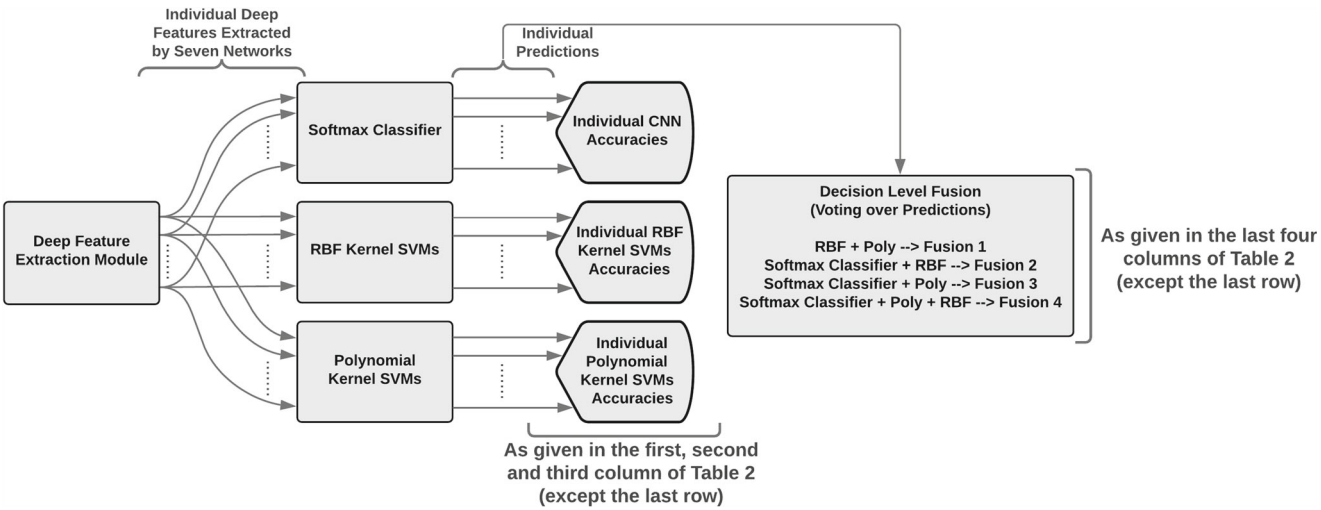


Fig. 2 The multistage learning approach and decision level fusion of individual classifiers. "Fusion 1" refers to the hard-level combination of the individual predictions obtained from RBF and Polynomial kernel based SVMs. "Fusion 2" refers to the hard-level combination of the individual predictions obtained from Softmax function and RBF kernel based SVM. "Fusion 3" refers to the hard-level combination of the individual predictions obtained from Softmax function and Polynomial kernel based SVM. "Fusion 4" refers to the hard-level combination of the individual predictions obtained from Softmax function, RBF and Polynomial kernel based SVMs

Table 2 The detailed presentation of accuracy values obtained from applied individual and ensemble learning scenarios for three data-sets (average accuracy values of 5-folds are given)

Accuracy (Standard Deviation)			Individual Classifiers			Decision Level Fusion over Predictions			
			Softmax Classifier	SVM (RBF)	SVM (Poly)	Fusion #1 RBF + Poly	Fusion #2 Softmax + RBF	Fusion #3 Softmax + Poly	Fusion #4 All
Individual Performances of Deep Neural Networks	Inception V3	DB1	87.6 (2.2)	87.6 (2.3)	87.3 (2.2)	87.5 (2)	87.3 (3.5)	87.4 (2.3)	87.7 (2.3)
		DB2	88.1 (0.9)	88.2 (0.6)	87.3 (0.8)	88 (0.6)	88.2 (0.8)	88 (0.8)	88.2 (0.6)
		DB3	87.2 (0.2)	87.8 (0.7)	87.9 (0.3)	87.8 (0.4)	87.6 (0.7)	87.6 (0.3)	87.8 (0.2)
	Xception	DB1	86.7 (3.8)	85.7 (4.2)	85.6 (3.4)	85.3 (3.8)	85.7 (3.8)	86.4 (3.5)	86.8 (3.4)
		DB2	86.9 (1.5)	85.3 (1.5)	86.6 (1.2)	85.2 (0.8)	85 (1.3)	86.5 (0.8)	86.4 (1.4)
		DB3	87.4 (0.7)	87.1 (0.2)	87.1 (0.8)	86.5 (0.4)	87.1 (0.9)	86.7 (0.4)	87.3 (0.2)
	MobileNet V2	DB1	84.2 (2.8)	86.8 (2.7)	86.0 (1.2)	86.1 (2.1)	84.7 (1.3)	84.7 (1.6)	86.1 (1.1)
		DB2	85.3 (2)	86.4 (1.6)	85.9 (1.3)	86.2 (1.3)	86 (1.6)	85.9 (1.4)	86.1 (1.5)
		DB3	86.9 (1.7)	87 (1.5)	87.1 (1.1)	86.9 (1.2)	86.8 (1.5)	86.8 (1.3)	87.1 (0.9)
	ResNet50	DB1	86.3 (2.3)	87.4 (2.5)	87.0 (2.7)	87.3 (2.5)	86.2 (2.6)	86.4 (2.5)	87.5 (2.3)
		DB2	87.5 (1.4)	87.2 (1.1)	86.7 (1.3)	87.2 (1.2)	87.1 (0.9)	87.3 (1.2)	87.5 (1.3)
		DB3	87.2 (1.8)	88.1 (1.5)	87.2 (1.4)	87.6 (1.8)	87.8 (2.1)	87.6 (1.8)	87.5 (0.9)
	ResNet101	DB1	85.5 (1.5)	85.7 (2.1)	85.5 (1.6)	85.7 (1.5)	85.4 (1.8)	85.4 (2.3)	85.8 (1.9)
		DB2	86.6 (1)	87 (0.8)	86.1 (1.3)	86.6 (1.2)	86.6 (0.5)	86.5 (1.1)	86.5 (1)
		DB3	87.3 (1.1)	87.2 (1.3)	86.9 (1.2)	86.9 (1)	87 (1.3)	87 (1.2)	87.1 (0.8)
	NasNet	DB1	85.2 (2.5)	84.8 (2.3)	84.3 (2.6)	84.4 (2.3)	84.8 (2.2)	84.4 (2.2)	84.6 (2.1)
		DB2	84.9 (1.1)	84 (1.5)	84.1 (1.7)	84.3 (1.9)	84.2 (1.5)	84.6 (1.3)	84.3 (1)
		DB3	85.8 (2.1)	85.5 (2.3)	84.1 (1.8)	84.9 (1.9)	85.9 (1.7)	85.2 (2)	85.5 (1.3)
	VGG16	DB1	85.8 (3.1)	86.3 (2.9)	85.9 (3.1)	86.3 (2.3)	86.0 (3)	85.9 (2.9)	86.1 (3.4)
		DB2	85.8 (1.9)	85.9 (0.9)	85.3 (1.1)	85.9 (1.4)	85.8 (1.6)	85.6 (1.5)	85.7 (1.3)
		DB3	86.7 (2.2)	86.7 (2.6)	86.4 (2.1)	86.7 (1.9)	86.9 (2.2)	87 (2.9)	86.5 (3.1)
Feature Level Fusion	Concatenated Vector	DB1	90.2 (2.3)	90.7 (1.7)	90.3 (2.2)	90.8 (1.6)	90.4 (2.1)	90.4 (2)	90.8 (1.7)
		DB2	90 (1.1)	89.6 (1.4)	89.8 (0.9)	89.5 (1.1)	90.1 (0.7)	90.1 (0.9)	90.5 (1)
		DB3	90.4 (1.2)	90.4 (1.6)	89.7 (1.4)	90 (1.7)	90.3 (1.4)	90.5 (1.1)	90.7 (1.8)

The standard deviations of accuracy values obtained from 5-folds are also presented in parentheses to show the robustness of applied approaches

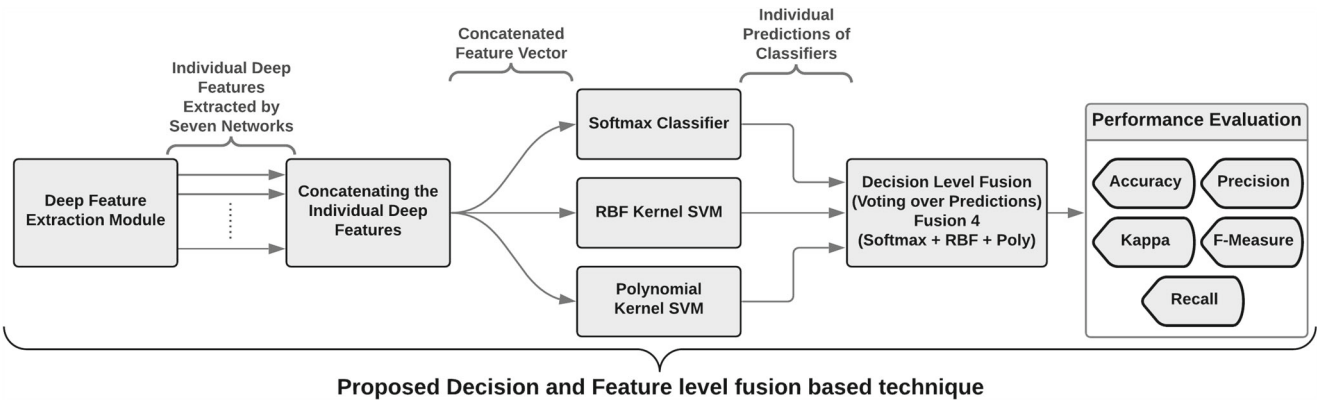


Fig. 3 The Flowchart of the proposed method employing feature and decision level fusion

3 Experimental results

3.1 Performance of individual classifiers

The individual performance of the employed seven CNN models plus the results of concatenated feature vector can be seen in the first column of Table 2 in terms of the accuracy metric (each presented accuracy value was calculated as the mean of 5-folds and the standard deviation of these 5-folds were also represented for clarification). The highest accuracy values were obtained as 87.6% and 88.1% for the DB1 and DB2 respectively by using InceptionV3, while the best classification performance was achieved as 87.4% for DB3 by employing the Xception net. In contrast, when the poorest individual performances are investigated, it is seen that the MobileNetV2 had the worst accuracy value as 84.2% for DB1, while the NasNet has ended up with the accuracy values as 84.9% and 85.8% for DB2 and DB3 respectively. The second and third columns of Table 2 show the accuracy values obtained by multistage learning scheme, which uses non-linear SVM kernels, for the individual deep feature sets and also for the concatenated feature vector as given in bottom row group. As seen in column 2, the highest accuracy value was obtained by using the RBF kernel as 87.6% for the DB1 with no increment compared to softmax classifier. On the other hand, the RBF kernel based SVM learning, which were fed by InceptionV3 deep features, has slightly increased best accuracy value to 88.2% for DB2, while the ResNet50 has reached to 88.1% for DB3 by using RBF kernel based multistage approach. In addition, the columns 4, 5, 6 and 7 indicate the accuracy values obtained by using the decision level fusion strategy composed of the combinations of softmax classifier, radial basis function (RBF) and polynomial kernel based SVMs as highlighted in the Table 2.

3.2 Performance obtained by feature and decision level fusion

Regarding the effect of feature level fusion, the bottom row group (named as "Feature Level Fusion") of Table 2 and the Fig. 4, in which the error values obtained from the three COVID-19 databases for the individual softmax classifier based learning models plus concatenated feature vector can be investigated. As seen in Fig. 4, the error values, which were obtained from the deep feature vector formed by using feature level fusion, are significantly lower than individual softmax classifier performance by reaching 9.8%, 10% and 9.6% errors for the DB1, DB2 and DB3 respectively. As understood from Table 2 and the Fig. 4, not a specific individual deep feature set (extracted by using a specific CNN) has outperformed the others for all three databases. This situation indicates that there is a significant need for ensemble learning which may pave the way for the complementary information achievement. It should also be noted that the error value for DB1 was even further reduced by 0.5% when RBF kernel based multistage learning algorithm was applied.

The contribution of decision level fusion can be investigated by using the right-side of Table 2 and the Fig. 5. In Fig. 5, the conventional classification performance of the softmax classifier (as it is used in traditional CNN based learning) was chosen as the reference baseline performance for seven CNN based deep feature extraction schemes. For comparison, the increments or decrements seen in the accuracy values obtained by the multistage SVM based learning and the decision level fusion were represented for each deep feature set plus the concatenated feature vector (obtained by the feature level fusion). When the Table 2 is investigated, it is seen that the highest accuracy values within the entire test set combinations were obtained as

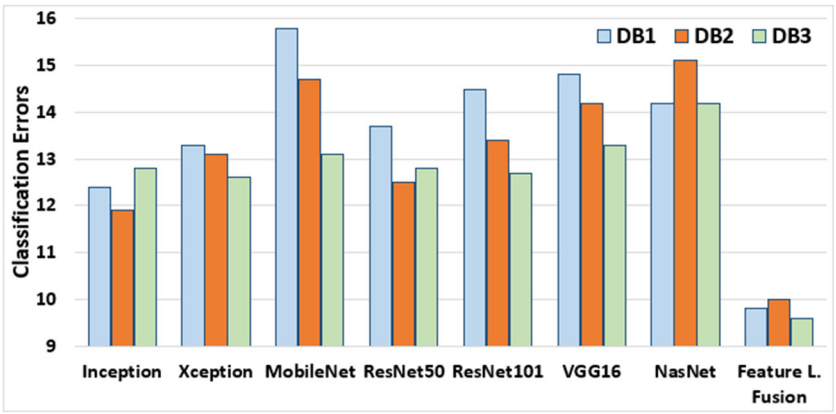


Fig. 4 Classification errors of the individual learning and the feature level fusion schemes when the softmax classifier is employed

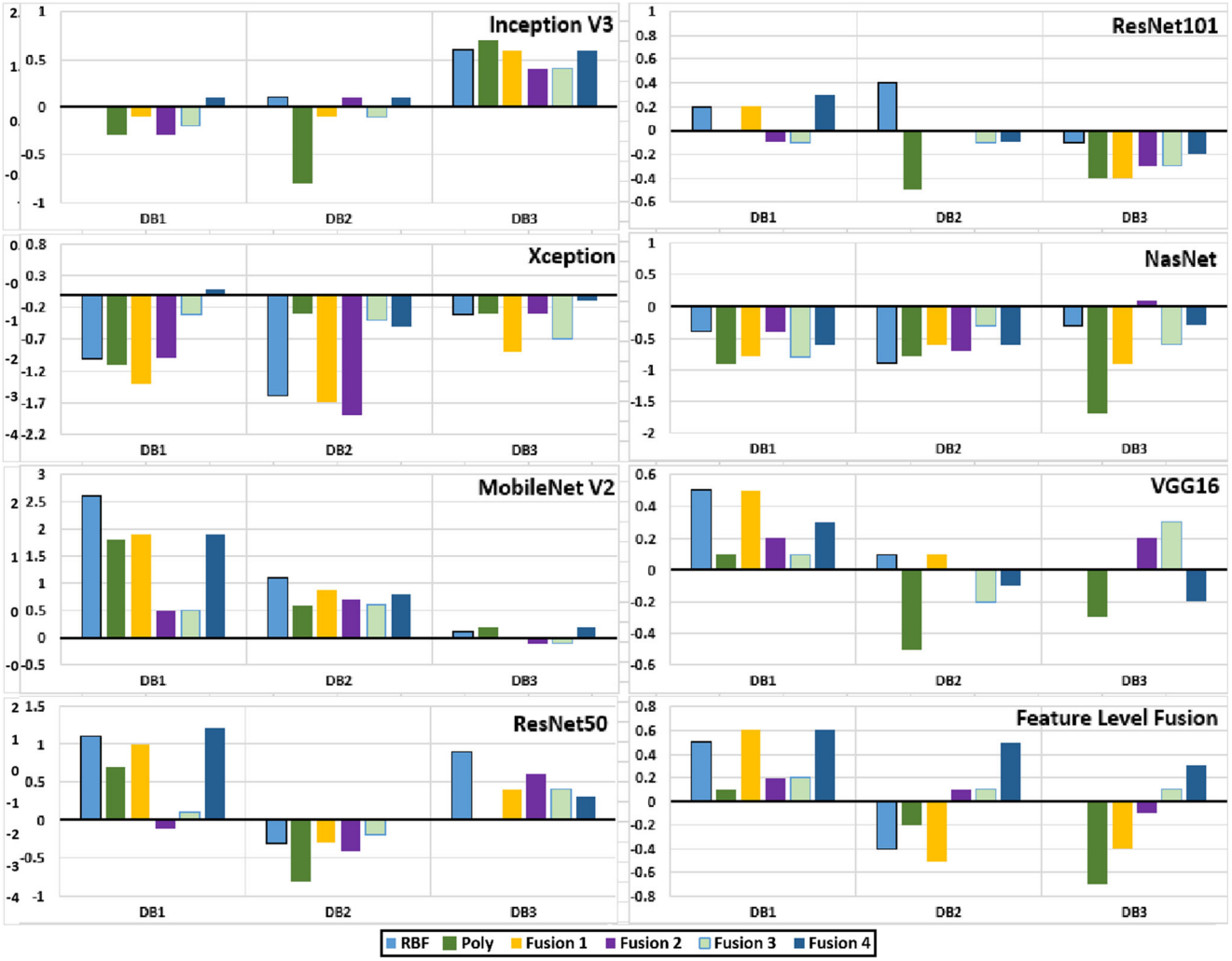


Fig. 5 The accuracy variations compared to CNNs when multistage learning and/or majority voting is applied are presented

90.8%, 90.5% and 90.7% for the DB1, DB2 and DB3 respectively, when the fourth decision level fusion approach, including the majority voting of hard labels obtained by softmax classifier, RBF and polynomial SVMs, was employed. As illustrated in Fig. 5, almost for all multistage SVM based learning and decision level fusion cases applied to MobileNetV2 based deep features, up to 2.5% increase in accuracy rate was achieved. On the contrary, approximately all the accuracy values obtained by decision level fusion, when they were applied to Xception and NasNet based deep features, were lower than the baseline softmax classifier performance. In accordance with the remaining VGG16, ResNet50, ResNet101 and InceptionV3 based deep features, neither the positive nor the negative effect of multistage learning and decision level fusion was clearly seen. For instance, up to 1% increase in the accuracy values was seen for the InceptionV3 and ResNet50 based scenarios in DB3, while slight improvements have been achieved by using VGG16 based scenarios for DB1. However, it should be noted that the proposed Fusion 4 approach has provided accuracy increments up to 0.6% in all data-sets for the feature-level fusion case as shown in the bottom-right side of Fig. 5.

3.3 Performance comparison by using confusion matrix based metrics

As alternative objective evaluation criteria, the confusion matrix based metrics were calculated to be able to show the performance of proposed approach. For doing this, the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values were obtained for each database. The confusion matrices obtained by the Fusion #4 strategy applied to 3 databases were given as Fig. 6 for further

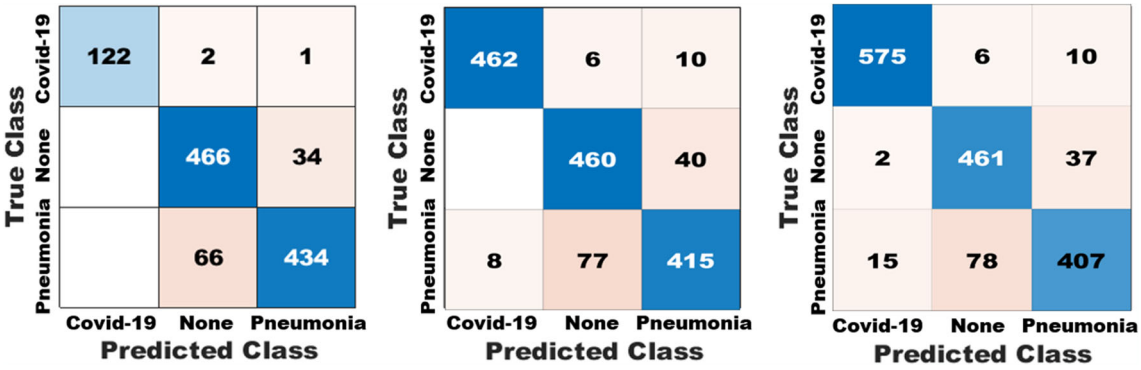


Fig. 6 Confusion Matrices obtained from Fusion #4 strategy (Left DB1, Middle DB2, Right DB3)

understanding. After the confusion matrices were obtained, 5 objective evaluation metrics were calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

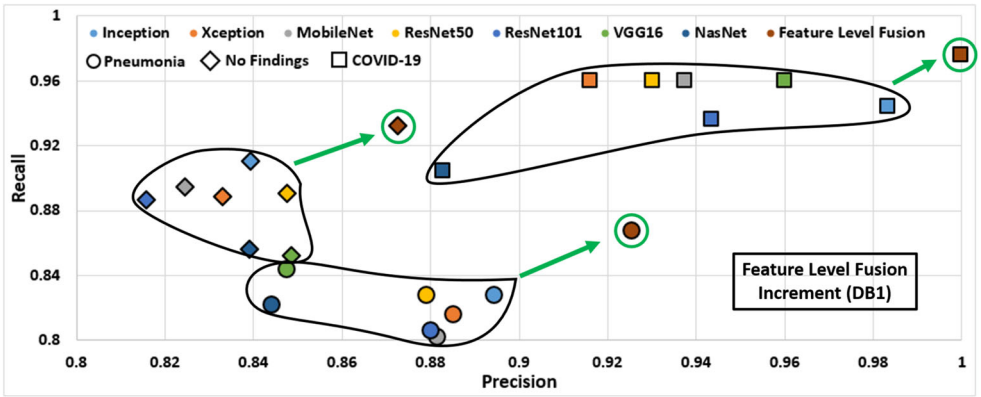
$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$F1\,score = 2\,\frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

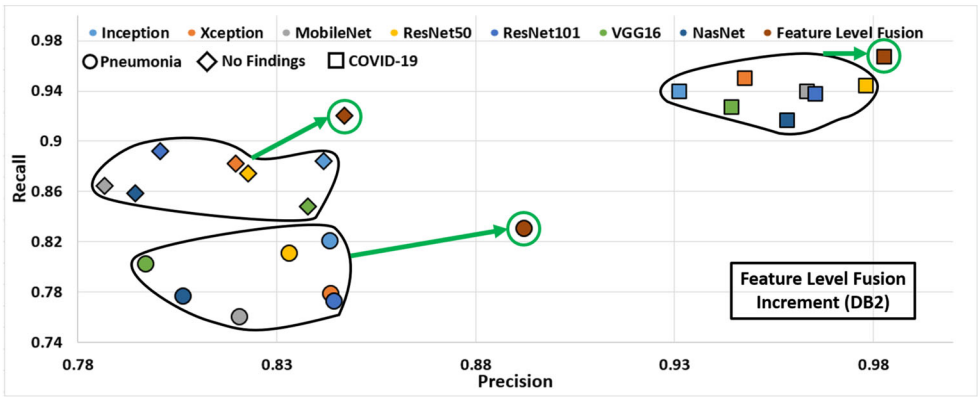
Among these, the accuracy indicates the ratio of the correctly classified samples over entire datasets. The precision emphasizes how precise the learning model is out of those predicted positive samples, how much of the predicted positives are actual positive. The precision is an important parameter to determine when the costs of FP predictions is high. Moreover, the recall measures how much of the actual positive samples are captured by the model by labeling it as positive (TP). The recall is an essential parameter when there is a high cost associated with FN samples. The behaviour of precision vs recall of the COVID-19, pneumonia and no-finding classes obtained by using majority voted decisions, described as Fusion #4, of individual deep feature sets (obtained by a specific CNN) plus the concatenated feature vector (obtained by the feature level fusion) is given in Fig. 7. It is seen that the precision and recall values obtained by the concatenated feature vector were higher than the individual deep feature sets in almost all cases. As presented in Table 2, the highest accuracy values were obtained when the Fusion #4 strategy was applied for all three databases. The obtained precision and recall values for Fusion #4 strategy is also depicted in Table 3 to go in deeper investigation. Addition to precision and recall, the specificity metric is given in the Table 3 to signify the proportion of negatives that are correctly identified by the proposed approach. As seen in this table, in almost all classes and databases, the highest precision, recall and F1-scores were obtained for the COVID-19 class which has the highest priority in our classification problem. In addition, an important evaluation metric named as Kappa, which is a statistical measure of inter-annotator agreement for categorical items by comparing an observed accuracy with an expected accuracy [61], was given for all databases in Table 3. As mentioned in [52], the Kappa values greater than 0.80 are called almost perfect classification. Hence, the obtained Kappa values (0.845, 0.857 and 0.86 for DB1, DB2 and DB3 respectively) shows the success of proposed approach following Fusion #4 strategy in COVID-19 diagnosis problem. As a final point to remark, all the F1-measure values indicating how precise the classifier is (what percentage of the samples assigned to a certain class is classified correctly), as well as how robust it is (what percentage of the samples belonging to a certain class is classified correctly), were quite high for the COVID-19 class, showing the success of proposed Fusion #4 strategy.

4 Discussion and conclusion

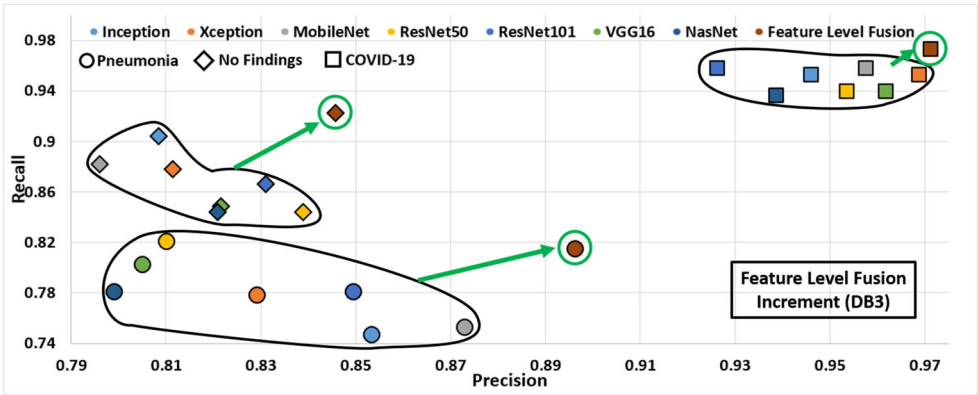
Although the RT-PCR is the most common technique to diagnose COVID-19, chest radiography based approaches have been extensively used as complementary diagnosis tools due to the low-sensitivity drawback of RT-PCR especially seen in the early stage of COVID-19. The X-ray scanning has been preferred as the primary radiography based imaging approach in COVID-19 detection due to its fast imaging speed, low cost and low dosage exposing of radiation compared to CT. However, the interpretation success of X-ray images strongly depends on the radiologist's experience and visual inspection of the X-



(a)



(b)



(c)

Fig. 7 Obtained precision and recall values of Fusion #4 strategy for each individual CNN and concatenated feature vector

Table 3 The detailed presentation of evaluation metrics for the Fusion #4 strategy

	DB1 ([66])			DB2 ([21])			DB3 ([21] + [1])		
	COVID-19	No Findings	Pneumonia	COVID-19	No Findings	Pneumonia	COVID-19	No Findings	Pneumonia
Precision	100	87.26	92.53	98.29	84.71	89.24	97.13	84.58	89.65
Recall	97.6	93.2	86.8	96.65	92	83	97.29	92.2	81.4
Specificity	100	89.1	94.4	99.1	91.4	94.86	98.1	92.1	95.7
F-Score	0.98	0.9	0.9	0.97	0.88	0.86	0.97	0.88	0.85
Accuracy		90.8			90.5			90.7	
Kappa		0.845			0.857			0.86	

ray images belonging to several patients takes significant time and effort. In order to increase the objectivity of the X-ray imaging interpretation and decrease the required time and effort, CAD systems have been used as supporting decision mechanisms in the detection of COVID-19 cases. In this respect, several studies employing deep networks as the decision tool were published lately as depicted in Table 4. Some of the previous studies have treated the COVID-19 diagnosis as a binary classification problem. For instance, five pre-trained CNN based models, which were using binary classification in their last layer, were employed in [64] for the COVID-19 X-ray image detection. In [33], deep learning models were introduced to confirm only positive or negative COVID-19 cases as an another binary classification approach. Additionally, an approach based on building two models, the first one aimed to detect whether a chest X-ray is related to a healthy subject or to a generic pulmonary disease patient, was studied in [10]. In the second phase of this study, the X-ray image was given to an another model that aims to detect whether the pulmonary disease is COVID-19. However, we have aimed to design a COVID-19 detection framework that is built on a three class learning model in the proposed study. Therefore, the studies which were using binary classification were not added to our comparison table. Regarding the number of employed COVID-19 X-ray samples, although sufficient number of X-ray samples to train a learning model exist in [60], [65], [91], [2], [68] and [96], the ratio of COVID-19 samples is very low compared to the distribution of the remaining classes. However, most of the learning models tend to work on balanced class distributions or equal misclassification costs, and the performance of these learning methods can be significantly compromised when imbalanced data sets, like the employed COVID-19 vs non COVID-19 distributions seen in [2, 60, 65, 68, 91, 96], are used. Therefore, in our study, the employed data bases were progressively created, starting from the usage of samples given in [66] as DB1, till minimum imbalance between employed classes was achieved in DB3. As an addition to the scores obtained in [66], the CNN model studied in [48] was also tested on the same DB1 data-set, and the results obtained from that study were also compared with our proposed approach's performance. As seen in Table 4, our method has outperformed [66], [60], [48] and [96] in terms of accuracy, precision and recall metrics, while our algorithm provides competitive performance compared to [65], [91] and [2] in terms of accuracy. It should be noted that our approach applied to DB1 is having the same number of image samples and same cross-validation strategy compared to [66], while a similar 5-fold cross-validation with different number of X-ray samples was carried out in [60] and [96]. Furthermore, the precision values obtained by using our method were significantly higher than [66], [60], [65], [68] and [96], while higher performance was achieved in terms of recall compared to [91] and [2].

A similar approach to our study, which uses deep features obtained from various CNNs and a SVM based classification strategy, was given in [79]. However, in [79] the deep features are fed to SVMS in an individual manner, while the fused deep features have been fed to non-linear SVMs in our proposed study. Additionally, a voting based decision level fusion strategy is also tested on the X-ray data-sets in our proposed approach. As a contribution of these feature+decision level fusion, it is seen that higher precision and recall values compared to [79] have been obtained when the proposed approach was applied to DB1, which has similar number of COVID-19 image samples as in [79].

As seen in Table 2 and the Fig. 4, none of the individual learning models has been significantly outperformed the others. However, accuracy improvements up to 2.5% were achieved when feature level fusion has been applied to obtained deep features. When the multistage learning and decision level fusion approaches are investigated, it is seen that the accuracy rises up to 2% and 0.5% have been obtained for the deep features extracted by using MobileNetV2 and VGG16 respectively. The supportive effect of SVM usage and majority voting for these two CNNs can be related to their sizes, which are the cause of possible underfitting and overfitting. As mentioned in [97], small networks such as the MobileNetV2 usually suffer from underfitting, while very large models such as the VGG16 may have trouble with overfitting [35]. However, a learner such as SVM, which is good at producing optimal decision surfaces even there is noise on the data, can have positive effect on the classification accuracy similar to our case. On the contrary, same multistage learning and majority voting strategy did not work well, resulting accuracy reductions for the deep features obtained by Xception and NasNet. When the architecture of NasNet is investigated, it is seen that the NasNet was constructed by a neural architecture search based optimization carried out by using reinforcement learning. As a result of this process, the well-designed scalable and convolutional cells are defined in the optimum way, resulting in an architecture that is prone to produce robust features as in our case. In a similar way, in the Xception, the usage of depthwise separable convolutions paves the way of efficiently usage of model parameters producing stronger features. Hereby, the cascade connection of the SVMs to the last FC layer of Xception and NasNet plus the usage of majority voting has no supportive effect in classification. So, the network based discrimination is more than enough for these two CNNs.

Another important fact that needs to be discussed about our proposed system, in which the Fusion #4 strategy was applied, is the obtained high precision and recall values.

Table 4 Performance comparison of related works on COVID-19 detection problem with the proposed method

Paper	Architecture	Total Number of Images	Class Names and number of Images	Accuracy (%)	Precision (%) (COVID-19)	Recall (%) (COVID-19)
Ozturk et al. 2020 [66]	DarkCovidNet	1125	COVID-19 (125) No Findings (500) Pneumonia (500)	87.02	80.7	97.87
		625	COVID-19 (125) No Findings (500)	98.08	97.97	90.65
Mahmud et al. 2020 [60]	CovXNet	5856	COVID-19 (305) No Findings (1583) Viral Pneumonia (1493) Bacterial Pneumonia (2780)	90.2	90.8	89.9
Oh et al. 2020 [65]	A patch-based ResNet18	15043	COVID-19 (180) Pneumonia (6012) Normal (8851)	91.9	76.9	100
Quan et al. 2021 [68]	A Deep Learning Framework based on DenseNet and CapsNet	9432	COVID-19 (781) Pneumonia (5734) Normal (2917)	90.66	87.27	96
Wang et al. 2020 [91]	COVID-Net	13975	COVID-19 (358) Pneumonia (5551) Normal (8066)	93.3	98.91	91
Ioannis and Mpeslana 2020 [2]	VGG-19	1428	COVID-19 (224) Normal (504) Pneumonia (700)	93.48	98.75	92.85

Table 4 (continued)

Paper	Architecture	Total Number of Images	Class Names and number of Images	Accuracy (%)	Precision (%) (COVID-19)	Recall (%) (COVID-19)
Sethy and Behera 2020 [79]	ResNet-50 + SVM	381	COVID-19 (127) Pneumonia (127) Non-Covid (127)	95.33	93.47	95.33
Narin et al. 2021 [64]	ResNet-50	100	COVID-19 (50) Non-Covid (50)	98	100	96
Hemdan et al. 2020 [33]	COVIDX-Net	50	COVID-19 (25) Non-Covid (25)	90	83	100
Kahn et al. 2020 [48]	A CNN model based on Xception Architecture	1125	COVID-19 (125) No Findings (500) Pneumonia (500)	90.21	97	89
Zhang et al. 2020 [96]	Confidence-aware Anomaly Detection (CAAD)	43583	COVID-19 (106) Normal(107) Viral Pneumonia (5977) Healthy + Non-Viral (37393)	72.77	71.7	73.83
Proposed Study	Deep Feature and Decision Level Fusion	1125	COVID-19 (125) No Findings (500) Pneumonia (500)	90.84	100	97.6
		1478	COVID-19 (478) No Findings (500) Pneumonia (500)	90.5	98.29	96.65
		1591	COVID-19 (591) No Findings (500) Pneumonia (500)	90.7	97.13	97.29

The precision value is directly related with the number of FP samples and low precision in COVID-19 means high number of healthy subjects that are misdiagnosed as COVID-19. An early quarantine measure applied to COVID-19 patients is employed as the fundamental disease control strategy across the countries [73]. Apart from the physical damages, the quarantine may cause dramatic psychological effects on the mental health. In previous studies, it was reported that the psychological impact of quarantine can vary from immediate effects such as irritability, fear of spreading infection to family members, confusion, anger, loneliness, anxiety, frustration, denial, insomnia, despair, depression, to extremes of consequences including suicide [7, 9, 24]. Therefore, the FP samples frequently seen in a COVID-19 detection system may cause significant undesired psychological and social consequences. However, as seen in Table 3, the proposed system has precision values, belonging to COVID-19 class, as 100%, 98.29% and 97.13% for the DB1, DB2 and DB3 respectively showing its almost perfect FP sample reduction performance. The recall metric, which is directly connected to FN samples, is also essential in COVID-19 detection because of the high cost associated with FN samples. Misdiagnosing a COVID-19 patient may cause dramatic consequences due to the very easy and fast transmission mechanism of the SARS-CoV2. The subject misdiagnosed as normal can spread the disease to his/her close environment in a very short time resulting in new patients who are ready to spread the disease further. However, thanks to our proposed approach, high recall values, reaching up to 97.6%, 96.65% and 97.29% in DB1, DB2 and DB3 respectively, were obtained by using the Fusion #4 strategy.

Although the deep learning approaches have enabled unprecedented breakthroughs in medical image analysis, the interpretable modules are sacrificed for uninterpretable ones that achieve higher performance through greater abstraction (more layers) and tighter integration (end-to-end training) in CNNs [78]. However, in [98], the Class Activation Mapping (CAM) technique, which is a way of producing visual explanations of the predictions of deep learning models [3], was proposed to make the CNNs more transparent and explainable. By using the CAM technique, useful knowledge about the employed prediction regions in the COVID-19 detection problem can be investigated. For example, the failure regions can be visually identified for the wrongly classified samples and necessary modifications in the learning models can be applied towards the most fruitful research directions. Besides, for a deep model, which is very strong in diagnosis, the CAM technique can visually identify the lung consolidation patterns as a supportive diagnostic tool for doctors. In Figure 8, two CAMs obtained from COVID-19 samples are given with the aim of visual validation of employed CNNs. In the CAMs, the red color highlights the lung regions where the employed CNN model focuses on (activating around those patterns) most during the discrimination. In Figure 8, upper row, the CAMs obtained with six CNNs, excluding VGG16 due to inability of representing its CAM by using employed approach, for an 83 year old male having mitral insufficiency, pulmonary hypertension and atrial fibrillation with COVID-19 infection, can be seen. In this patient, Ground-glass opacification (GGO) and consolidation in the right upper lobe and left lower lobe is seen as the indicators of COVID-19. The InceptionV3 and ResNet50 have correctly localized the right upper lobe pattern, while

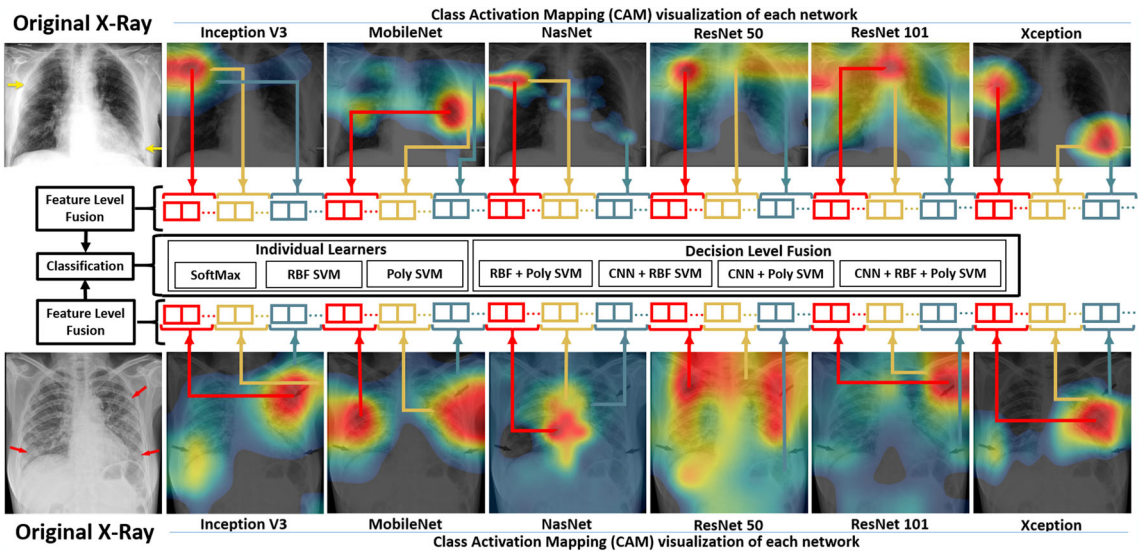


Fig. 8 CAM visualizations of two patients obtained by six CNNs (top and bottom rows) and the flow-chart of employed feature level fusion (middle row)

missing the left lower lobe. However, the Xception has successfully detected both two pathological regions with high spatial resolution. In the bottom row, the CAMs of a 53-year-old female, whose X-ray contains multifocal patchy opacities in both lungs, was depicted. This case is a good example to see the effect of feature level fusion of different CNNs because of the existing three separate opacity patterns. While the MobileNet has strong focus on right side single pattern, the InceptionV3, ResNet50 and Xception has low activation on right side. However, the ResNet101 and InceptionV3 have highly focused on left side upper pattern, while the Xception and MobileNet has significant activity near the left side lower pattern. When the complementary effect of these CAMs is considered, it is obvious that the fusion of features obtained by these CNNs would have higher discriminating power. In the middle part of Fig. 8, a flowchart explaining how the features obtained from various CNNs are concatenated is given for further understanding. Additionally, in Fig. 9, X-ray images belonging to the same patient with bilateral GGO are shown. The image in upper row is taken on the second day of diagnosis while the bottom row X-ray image is taken on the fourth day. As it can be seen, the active regions belonging to a specific CNN are consistent and not dramatically changing towards second and fourth day images.

In future research, we aim to focus on following research paths related with COVID-19 for further improvement; i) a different version of the feature level fusion, in which the features obtained from the various layers of the same CNN are concatenated, can be employed instead of the fusion of features obtained from the last FC layer of different type CNNs. By doing that diverse features, which contain more semantic information in the top layers and more low-level information in bottom layers, can be combined to provide more discriminative information. ii) the concatenation of feature sub-sets obtained from various deep-nets may cause two possible drawbacks for the subsequent learning models. First drawback would be the complexity increase [46], while the second can show itself as the difficulty in pattern identification due to curse of dimensionality, which is referred as having more features than the number of observations. A feature selection approach, such as the ReliefF [50], can applied to the concatenated feature set to reduce the learning algorithm complexity and prevent a possible overfitting scenario [47]. iii) since the outbreak is recent, the number of COVID-19 X-ray images, which can be used in CAD system design studies, is very limited. Even though there exists studies that uses Generative Adversarial Networks (GANs) [90] and attention guided augmentations [54] for increasing the number of training samples, the performance can be improved by using Progressive Growing GAN [42] for augmentation. Besides, the quality of artificial COVID-19 samples can be improved by integrating more labeled data into the learning process by using GANs. iv) the Canonical Correlation Analysis (CCA) [34, 44], which aims at measuring linear relationships between two sets of variables by using the within-set and between-set sample covariance matrices, can be employed as a feature fusion approach instead of simple concatenation of deep features. By utilizing the multi-view features (the deep features extracted from different CNNs and/or from the different layers of the same CNN), more discriminating features having maximized correlation between various sets can be attained with the hope of performance increase in COVID-19 detection. v) the hyperparameters, which are adjusted prior to the learning process and affect how the learning algorithm fits the model to data, can be tuned by using automatic tuning algorithms such as the Bayesian optimization [94]. In this way, the optimum hyperparameters for the COVID-19 detection problem can be tuned for both CNNs and SVMs to obtain higher performance.

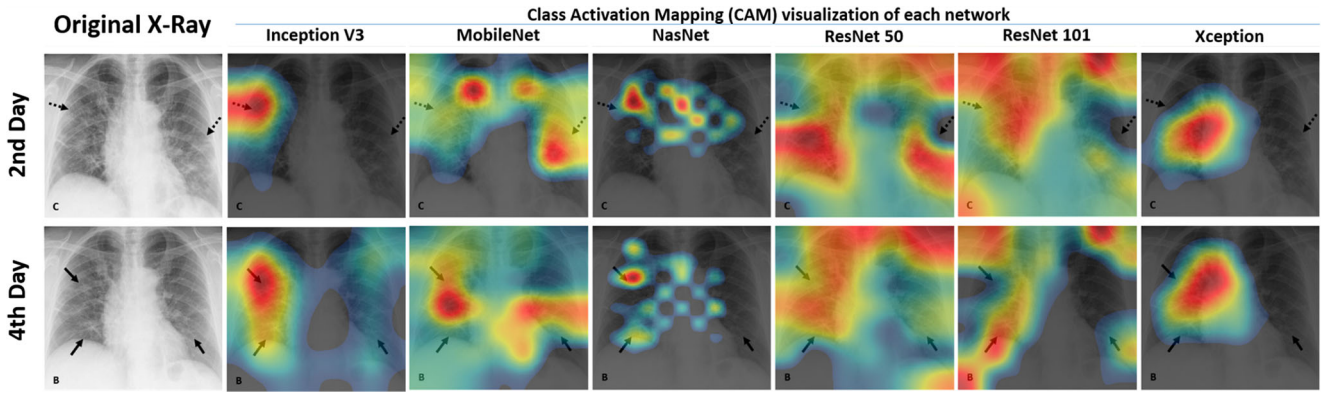


Fig. 9 CAM visualizations of the same patient on the second and fourth day of diagnosis

Funding This research received no external funding.

Declarations

Conflict of Interests The authors declare no conflict of interest.

References

1. Agchung: Actualmed COVID-19 Chest X-ray Dataset Initiative (2020). https://github.com/agchung/Actualmed-COVID-chestxray-dataset
2. Apostolopoulos ID, Mpesiana TA (2020) Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine, p 1
3. Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, Hoebel K, Gupta S, Patel J, Gidwani M et al (2020) Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. arXiv preprint arXiv:2008.02766
4. Azizpour H, Razavian AS, Sullivan J, Maki A, Carlsson S (2015) Factors of transferability for a generic convnet representation. IEEE transactions on pattern analysis and machine intelligence 38(9):1790–1802
5. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A (2019) Comparison of deep learning approaches for multi-label chest x-ray classification. Scientific reports 9(1):1–10
6. Bar Y, Diamant I, Wolf L, Greenspan H (2015) Deep learning with non-medical training used for chest pathology identification. In: Medical imaging 2015: computer-aided diagnosis, vol. 9414, p. 94140v. International society for optics and photonics
7. Barbisch D, Koenig KL, Shih FY (2015) Is there a case for quarantine? perspectives from sars to ebola. Disaster medicine and public health preparedness 9(5):547–553
8. Borghesi A, Maroldi R (2020) Covid-19 outbreak in italy: experimental chest x-ray scoring system for quantifying and monitoring disease progression. La radiologia medica, p 1
9. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, Rubin GJ (2020) The psychological impact of quarantine and how to reduce it: rapid review of the evidence. The lancet 395(10227):912–920
10. Brunese L, Mercaldo F, Reginelli A, Santone A (2020) Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays. Computer Methods and Programs in Biomedicine p 105608
11. Buscombe D, Carini RJ, Harrison SR, Chickadel CC, Warrick JA (2020) Optical wave gauging using deep neural networks. Coast Eng 155:103593
12. Cao S, Zhao D, Liu X, Sun Y (2020) Real-time robust detector for underwater live crabs based on deep learning. Comput Electron Agric 172:105339
13. Chamola V, Hassija V, Gupta V, Guizani M (2020) A comprehensive review of the covid-19 pandemic and the role of iot, drones, ai, blockchain, and 5g in managing its impact. IEEE Access 8:90225–90265
14. Chen H, Ni D, Qin J, Li S, Yang X, Wang T, Heng PA (2015) Standard plane localization in fetal ultrasound via domain transferred deep neural networks. IEEE journal of biomedical and health informatics 19(5):1627–1636
15. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y et al (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. The Lancet 395(10223):507–513
16. 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
17. Chung M, Bernheim A, Mei X, Zhang N, Huang M, Zeng X, Cui J, Xu W, Yang Y, Fayad ZA et al (2020) Ct imaging features of 2019 novel coronavirus (2019-ncov). Radiology 295(1): 202–207
18. Cilia ND, De Stefano C, Fontanella F, Marrocco C, Molinara M, Di Freca AS (2020) An end-to-end deep learning system for medieval writer identification. Pattern Recogn Lett 129:137–143
19. Clerkin KJ, Fried JA, Raikhelkar J, Sayer G, Griffin JM, Masoumi A, Jain SS, Burkhoff D, Kumaraiah D, Rabbani L et al (2020) Covid-19 and cardiovascular disease. Circulation 141(20):1648–1655
20. Cogan T, Cogan M, Tamil L (2019) Mapgi: Accurate identification of anatomical landmarks and diseased tissue in gastrointestinal tract using deep learning. Computers in biology and medicine 111:103351
21. Cohen JP, Morrison P, Dao L, Roth K, Duong TQ, Ghassemi M (2020) Covid-19 image data collection:, Prospective predictions are the future. arXiv preprint arXiv:2006.11988
22. Doi K (2007) Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics 31(4-5):198–211
23. Du P, Li E, Xia J, Samat A, Bai X (2018) Feature and model level fusion of pretrained cnn for remote sensing scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12(8):2600–2611
24. Dubey S, Biswas P, Ghosh R, Chatterjee S, Dubey MJ, Chatterjee S, Lahiri D, Lavie CJ (2020) Psychosocial impact of covid-19. Diabetes & Metabolic syndrome: Clinical Research & Reviews
25. Elasnaoui K, Chawki Y (2020) Using x-ray images and deep learning for automated detection of coronavirus disease. Journal of Biomolecular Structure and Dynamics (just-accepted), pp 1–22
26. Fang L, Karakiulakis G, Roth M (2020) Are patients with hypertension and diabetes mellitus at increased risk for covid-19 infection? The Lancet. Respir Med 8(4):e21
27. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, Ji W (2020) Sensitivity of chest ct for covid-19: comparison to rt-pcr. Radiology p 200432
28. Fuhad K, Tuba JF, Sarker M, Ali R, Momen S, Mohammed N, Rahman T (2020) Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application. Diagnostics 10(5):329
29. Gao C, Wang P, Gao Y (2019) Mobilecount: an efficient encoder-decoder framework for real-time crowd counting. In: Chinese conference on pattern recognition and computer vision (PRCV), pp. 582–595. Springer
30. Gunatilaka AH, Baertlein BA (2001) Feature-level and decision-level fusion of noncoincidently sampled sensors for land mine detection. IEEE transactions on pattern analysis and machine intelligence 23(6):577–589
31. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778
32. Hegde S (2020) Does asthma make covid-19 worse?
33. Hemdan EED, Shouman MA, Karar ME (2020) Covidx-net:, A framework of deep learning classifiers to diagnose covid-19 in x-ray images. arXiv preprint arXiv:2003.11055
34. Hotelling H (1936) Relations between two sets of variates. Biometrika 28(3/4):321–377. http://www.jstor.org/stable/2333955

35. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets:, Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861
36. Huang F, LeCun Y (2006) Large-scale learning with svm and convolutional netw for generic object recognition. In: 2006 IEEE Computer society conference on computer vision and pattern recognition, vol. 10
37. Ilhan HO, Serbes G, Aydin N (2020) Automated sperm morphology analysis approach using a directional masking technique. Computers in Biology and Medicine p 103845
38. Ilhan HO, Sigirci IO, Serbes G, Aydin N (2020) A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods. Medical & Biological Engineering & Computing, pp 1–22
39. Jacobi A, Chung M, Bernheim A, Eber C (2020) Portable chest x-ray in coronavirus disease-19 (covid-19): A pictorial review Clinical Imaging
40. Jaiswal AK, Tiwari P, Kumar S, Gupta D, Khanna A, Rodrigues JJ (2019) Identifying pneumonia in chest x-rays: a deep learning approach. Measurement 145:511–518
41. Kanne JP, Little BP, Chung JH, Elicker BM, Ketai LH (2020) Essentials for radiologists on covid-19: an update—radiology scientific expert panel
42. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196
43. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F et al (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172(5):1122–1131
44. Kettenring JR (1971) Canonical analysis of several sets of variables. Biometrika 58(3):433–451
45. Khamparia A, Saini G, Pandey B, Tiwari S, Gupta D, Khanna A (2019) Kdsae: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network. Multimedia Tools and Applications, pp 1–16
46. Khan A, Chefranov A, Demirel H (2021) Image scene geometry recognition using low-level features fusion at multi-layer deep cnn. Neurocomputing 440:111–126
47. Khan A, Eker A, Chefranov A, Demirel H (2021) White blood cell type identification using multi-layer convolutional features with an extreme-learning machine. Biomedical Signal Processing and Control 69:102932
48. Khan AI, Shah JL, Bhat MM (2020) Coronet: a deep neural network for detection and diagnosis of covid-19 from chest x-ray images. Comput Methods Prog Biomed 196: 105581
49. Kittler J, Hatef M, Duin RP, Matas J (1998) On combining classifiers. IEEE transactions on pattern analysis and machine intelligence 20(3):226–239
50. Kononenko I (1994) Estimating attributes: Analysis and extensions of relief. In: European conference on machine learning, pp. 171–182. Springer
51. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105
52. Landis JR, Koch GG (1977) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics, pp 363–374
53. Li B, Yang J, Zhao F, Zhi L, Wang X, Liu L, Bi Z, Zhao Y (2020) Prevalence and impact of cardiovascular metabolic diseases on covid-19 in china. Clin Res Cardiol 109(5):531–538
54. Li J, Wang Y, Wang S, Wang J, Liu J, Jin Q, Sun L (2021) Multiscale attention guided network for covid-19 diagnosis using chest x-ray images. IEEE Journal of Biomedical and Health Informatics 25(5):1336–1346
55. Li S, Kwok JY, Tsang IH, Wang Y (2004) Fusing images with different focuses using support vector machines. IEEE Transactions on neural networks 15(6):1555–1561
56. Li Y, Xia L (2020) Coronavirus disease 2019 (covid-19): role of chest ct in diagnosis and management. Am J Roentgenol 214(6):1280–1286
57. Liang WJ, Zhang H, Zhang GF, Cao HX (2019) Rice blast disease recognition using a deep convolutional neural network. Scientific reports 9(1):1–10
58. Liao B, Xu J, Lv J, Zhou S (2015) An image retrieval method for binary images based on dbn and softmax classifier. IETE Tech Rev 32(4):294–303
59. Lundervold AS, Lundervold A (2019) An overview of deep learning in medical imaging focusing on mri. Zeitschrift fü,r Medizinische Physik 29(2):102–127
60. Mahmud T, Rahman MA, Fattah SA (2020) Covxnet: A multi-dilation convolutional neural network for automatic covid-19 and other pneumonia detection from chest x-ray images with transferable multi-receptive feature optimization. Computers in Biology and Medicine p 103869
61. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochemia medica:, Biochemia medica 22(3):276–282
62. Menegola A, Fornaciali M, Pires R, Avila S, Valle E (2016) Towards automated melanoma screening:, Exploring transfer learning schemes. arXiv preprint arXiv:1609.01228
63. Mohandes M, Deriche M, Aliyu SO (2018) Classifiers combination techniques: a comprehensive review. IEEE Access 6:19626–19639
64. Narin A, Kaya C, Pamuk Z (2021) Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. Pattern Anal Applic, pp 1–14
65. Oh Y, Park S, Ye JC (2020) Deep learning covid-19 features on cxr using limited training data sets IEEE Transactions on Medical Imaging
66. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR (2020) Automated detection of covid-19 cases using deep neural networks with x-ray images. Computers in Biology and Medicine p 103792
67. Pacheco AG, Krohling RA (2020) The impact of patient clinical information on automated skin cancer detection. Computers in biology and medicine 116:103545
68. Quan H, Xu X, Zheng T, Li Z, Zhao M, Cui X (2021) Densecapsnet: Detection of covid-19 from x-ray images using a capsule neural network. Computers in biology and medicine 133:104399
69. Rahimzadeh M, Attar A (2020) A modified deep convolutional neural network for detecting covid-19 and pneumonia from chest x-ray images based on the concatenation of xception and resnet50v2. Informatics in Medicine Unlocked p 100360
70. Raja K, Venkatesh S, Christoph Busch R et al (2017) Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 10–18
71. Rajaraman S, Candemir S, Kim I, Thoma G, Antani S (2018) Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. Appl Sci 8(10):1715
72. Razai MS, Doerholt K, Ladhani S, Oakeshott P (2020) Coronavirus disease 2019 (covid-19): a guide for uk gps. BMJ **368**

73. Rubin GJ, Wessely S (2020) The psychological effects of quarantining a city. Bmj, **368**
74. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. International journal of computer vision 115(3):211–252
75. Sakar CO, Serbes G, Gunduz A, Tunc HC, Nizam H, Sakar BE, Tutuncu M, Aydin T, Isenkul ME, Apaydin H (2019) A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform. Appl Soft Comput 74:255–263
76. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520
77. Self WH, Courtney DM, McNaughton CD, Wunderink RG, Kline JA (2013) High discordance of chest x-ray and computed tomography for detection of pulmonary opacities in ed patients: implications for diagnosing pneumonia. The American journal of emergency medicine 31(2):401–405
78. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp. 618–626
79. Sethy PK, Behera SK (2020) Detection of coronavirus disease (covid-19) based on deep features. Preprints **2020030300**, 2020
80. Sharif MI, Li JP, Khan MA, Saleem MA (2020) Active deep neural network features selection for segmentation and recognition of brain tumors using mri images. Pattern Recogn Lett 129:181–189
81. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM (2016) Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35(5):1285–1298
82. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
83. Su R, Liu T, Sun C, Jin Q, Jennane R, Wei L (2020) Fusing convolutional neural network features with hand-crafted features for osteoporosis diagnoses. Neurocomputing 385:300–309
84. 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
85. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J (2016) Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35(5):1299–1312
86. Tang Y (2013) Deep learning using support vector machines. CoRR, abs/1306.0239 **2**
87. Toğaçar M, Ergen B, Cömert Z (2020) Covid-19 detection using deep learning models to exploit social mimic optimization and structured chest x-ray images using fuzzy color and stacking approaches. Computers in Biology and Medicine p 103805
88. Ulukaya S, Serbes G, Kahya YP (2017) Overcomplete discrete wavelet transform based respiratory sound discrimination with feature and decision level fusion. Biomedical Signal Processing and Control 38:322–336
89. Van Ginneken B, Setio AA, Jacobs C, Ciompi F (2015) Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In: 2015 IEEE 12Th international symposium on biomedical imaging (ISBI), pp. 286–289. IEEE
90. Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR (2020) Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. IEEE Access 8:91916–91923
91. Wang L, Lin ZQ, Wong A (2020) Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific Reports 10(1):1–12
92. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2017) Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2097–2106
93. Wang Y, Dong C, Hu Y, Li C, Ren Q, Zhang X, Shi H, Zhou M (2020) Temporal changes of ct findings in 90 patients with covid-19 pneumonia: a longitudinal study. Radiology p 200843
94. Wu J, Chen XY, Zhang H, Xiong LD, Lei H, Deng SH (2019) Hyperparameter optimization for machine learning models based on bayesian optimization. Journal of Electronic Science and Technology 17(1):26–40
95. Yadav SS, Jadhav SM (2019) Deep convolutional neural network based medical image classification for disease diagnosis. Journal of Big Data 6(1):113
96. Zhang J, Xie Y, Pang G, Liao Z, Verjans J, Li W, Sun Z, He J, Li Y, Shen C et al (2020) Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. IEEE transactions on medical imaging 40(3):879–890
97. Zhang X, Zhou X, Lin M, Sun J (2018) Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6848–6856
98. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929
99. Zhuang X, Zhang T (2019) Detection of sick broilers by digital image processing and deep learning. Biosyst Eng 179:106–116
100. 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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Hamza Osman Ilhan is assistant professor in YTU. His research interests are in the areas of image and signal processing, machine learning and pattern recognition with applications to biomedical engineering.



Gorkem Serbes has experience in biomedical engineering, digital signal processing and pattern recognition for more than 10 years, and he has authored over 40 peer-reviewed manuscripts in these fields.



Nizamettin Aydin is now with Computer Engineering Department and serves as the head of the department at YTU. He has experience in biomedical engineering, digital signal processing and pattern recognition.