Received: 15 October 2021 | Revised: 28 November 2021 | Accepted: 3 December 2021

DOI: 10.1002/cpe.6802

SPECIAL ISSUE PAPER

WILEY

COVID-19 detection with severity level analysis using the deep features, and wrapper-based selection of ranked features

Coşku Öksüz[1] | Oğuzhan Urhan[2] | Mehmet Kemal Güllü[3]

[1]Department of Electronics and Automation, Bozkurt Vocational School of Kastamonu University, Kastamonu, Bozkurt, Turkey

[2]Department of Electronics and Telecommunication Engineering, University of Kocaeli, Kocaeli, Turkey

[3]Department of Electrical and Electronics Engineering, University of Bakırçay, İzmir, Turkey

Correspondence
Coşku Öksüz, Department of Electronics and Automation, Bozkurt Vocational School of Kastamonu University, Kastamonu 37680, Turkey.
Email: coksuz@kastamonu.edu.tr, coskuoksuz@gmail.com

Abstract

The SARS-COV-2 virus, which causes COVID-19 disease, continues to threaten the whole world with its mutations. Many methods developed for COVID-19 detection are validated on the data sets generally including severe forms of the disease. Since the severe forms of the disease have prominent signatures on X-ray images, the performance to be achieved is high. To slow the spread of the disease, effective computer-assisted screening tools with the ability to detect the mild and the moderate forms of the disease that do not have prominent signatures are needed. In this work, various pretrained networks, namely GoogLeNet, ResNet18, SqueezeNet, ShuffleNet, EfficientNetB0, and Xception, are used as feature extractors for the COVID-19 detection with severity level analysis. The best feature extraction layer for each pre-trained network is determined to optimize the performance. After that, features obtained by the best layer are selected by following a wrapper-based feature selection strategy using the features ranked based on Laplacian scores. The experimental results achieved on two publicly available data sets including all the forms of COVID-19 disease reveal that the method generalized well on unseen data. Moreover, 66.67%, 90.32%, and 100% sensitivity are obtained in the detection of mild, moderate, and severe cases, respectively.

KEYWORDS

computer-aided diagnosis, COVID-19 detection, deep features, mild, X-ray imaging

1 | INTRODUCTION

The COVID-19 disease has left a devastating impact all over the world. As of October 13, deaths worldwide reached 4.8 M.[1] On the other hand, there are about 238 M confirmed cases.[1] The SARS-CoV-2 virus that causes the COVID-19 disease has become more contagious than ever with various mutations such as Delta.[2] In Reference 2, it is stated that a full vaccination is mandatory to suppress the SARS-CoV-2 delta variant mutation frequency. Despite this, there are still many unvaccinated people around the world. Although vaccines have been developed to combat the virus, people who have been vaccinated can also become infected and spread the disease. Currently, the RT-PCR test is the gold standard test to identify the SARS-CoV-2 virus.[3,4] Although the RT-PCR test is regarded as the gold standard for the detection of COVID-19 disease, it necessitates taking specimens from the patients via nasopharyngeal or oropharyngeal swabs.[5] Therefore, it requires health care professionals equipped with personnel protective equipment to reduce the risk of transmission. More importantly, this test, which requires intervention to patients, has low sensitivity as it produces many false-negative results.[6] There is an urgent need to develop a computer-aided diagnostic system for COVID-19 detection during the pandemic that does not require human intervention and has high sensitivity. The development of such systems is important in

order to control the spread of the disease and reduce deaths, especially in underdeveloped countries where there are not enough specialists and equipment.

Many works have been done for diagnosing the COVID-19 disease during the pandemic using medical images such as chest radiograph (X-ray) and computed tomography (CT). The methods proposed in many of the works are based on CT imaging as it gives detailed information about the lungs.[7–10] Although CT images give more detailed information about the lungs, the amount of exposed radiation doses is significantly higher than X-ray imaging.[11] In addition, X-ray is an inexpensive imaging technique available in almost every health institution. Therefore, the methods are proposed based on X-ray images in many other works.[12–15] It is aimed to develop a method based on X-ray images due to the factors mentioned above in this study. One of the issues with the proposed X-ray-based methods is the reported performance scores. The data sets used in many published studies consist of severe forms of the COVID-19 disease. Since the signatures of the severe forms of the disease can be captured more easily in X-ray images, the performance of the proposed methods is presented as high. To avoid this misleading situation, it is a requirement to use a data set containing all forms of COVID-19 disease. For this purpose, data sets including mild, moderate, and severe forms of COVID-19 disease are used in the study.

Deep learning is a recently emerging field that yields significant improvement in performance especially when the amount of training data is sufficient. The performance obtained with deep learning methods is superior to the methods adopting hand-crafted feature engineering due to better capture of the patterns characterizing the data. Deep learning models have been trained with a large number of data and learned many important combinations of low- and high-level features. For this reason, distinctive features can also be obtained by using networks as feature extractors. In this work, the lung regions segmented using the lung segmentation network (i.e., *Ensemble-LungMaskNet* model) proposed in our previous work is fed to the pretrained networks to extract only relevant features. The best feature extraction layer of each pretrained network is found by cross-validation. Then, redundant features are eliminated using a wrapper-based method that utilizes ranked features returned by a filter method. Finally, the SVM classification is done with the selected features.

The contributions of this study can be summarized as follows:

1. Our previously proposed segmentation model namely the *Ensemble-LungMaskNet*[16] is used to segment lung regions within the X-ray images. Thus, only the region of interest (ROI) is taken into account by eliminating the irrelevant regions that are not important for classification.
2. Instead of training a CNN from scratch, a pretrained network is used as a feature extractor by finding the optimal layer for feature extraction that optimizes the classification performance.
3. Redundant features have been eliminated to both optimize performance and reduce dimensionality. Accordingly, a wrapper-based feature selection strategy is followed that uses ranked features returned by a filter-based method.
4. The method is proposed based on a data set including all the forms of COVID-19 disease to obtain more realistic results.

The remaining part of the study is organized as follows. The proposed method is presented in Section 2. The experimental results are given in Section 3. Section 4 is devoted to the discussion. Finally, Section 5 is devoted to the conclusion.

2 | PROPOSED METHOD

The proposed method is shown in Figure 1. As seen from Figure 1, the method consists of consecutive stages including the segmentation stage to obtain ROI, a pretrained network to extract features from ROI, a feature selection stage, and the SVM classification stage.

2.1 | Segmentation

In the scope of the study, we make use of our previously proposed model that is, *Ensemble-LungMaskNet*[16] for lung segmentation. In this earlier work, it is demonstrated that ensembling of the pretrained encoders in different depths as a single feature extraction backbone yields superior lung segmentation performance. This explains why we preferred to use this model in this work. In the proposed framework, the regions including lungs are first returned by the *Ensemble-LungMaskNet*[16] which accepts a chest X-ray image as an input. Then, the lung mask obtained by the network is element-wise multiplied with the input image to obtain only the pixels within the lungs. Finally, the ROI is fed to the feature extraction stage.

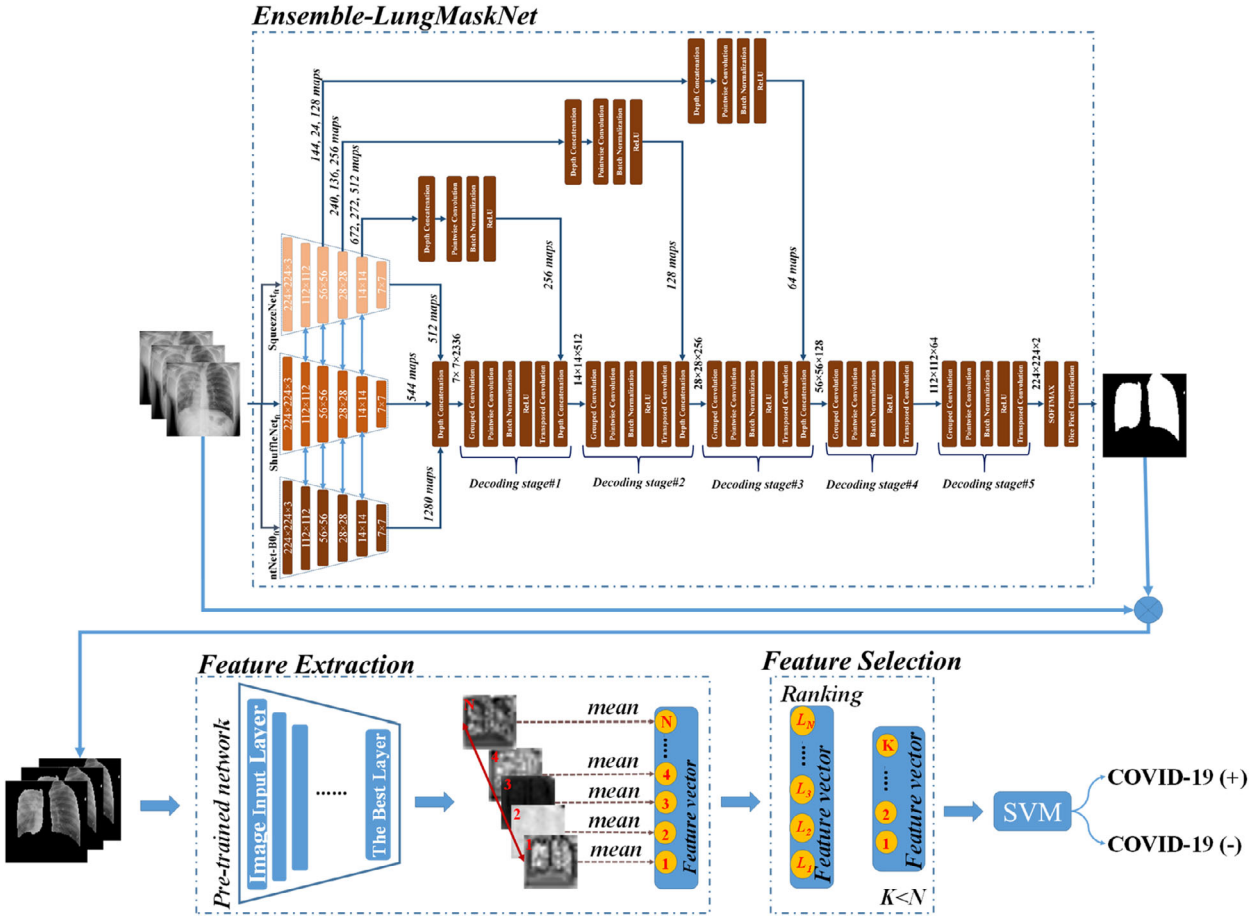


FIGURE 1 The proposed method

2.2 | Feature extraction, selection, and classification

Feature extraction aims to derive new features with reduced dimensions from the raw data while it still preserves the information within the original data. Instead of using the raw data as features, extracting features from the raw data improves the efficiency with the features that have better representative power. However, the manual feature exploring process is difficult. Feature extraction in a manual way generally depends on the predefined parameters of an algorithm. For example, the parameter of cell size of both the HOG (histogram of oriented gradients)[17] method and LBP (local binary patterns)[18] methods impact the performance significantly. On the other hand, using a pretrained network is an automated way of feature extraction since manual feature exploring procedure is avoided. Moreover, because pretrained networks are trained with over one million images, combinations of low- and high-level information have already been learned by the networks. Therefore, we have adopted to use of the pretrained network as a feature extractor. The best layer for feature extraction is determined by an experimental process. Then, the extracted features are ranked based on the Laplacian scores[19] which is a filter-based unsupervised feature selection method. Laplacian scores of each feature (L_r) are computed as given in (1), where L, D_g, and x_r are representing the Laplacian matrix, the degree matrix, and the rth feature with removed mean, respectively. The degree matrix is a diagonal square matrix where each diagonal element of the matrix is the sum of each row of a similarity matrix ($S_{i,j}$). The similarity matrix is computed by transforming the pairwise distances between the features using the kernel function. The Laplacian matrix is computed as the difference between the degree matrix, and the similarity matrix. The highest value of the Laplacian score indicates a more important feature.

$$L_r = \frac{x_r^T L x_r}{x_r^T D_g x_r}. \tag{1}$$

After that, the ranked features are selected with a wrapper-based feature selection method. Finally, the SVM model is used with the selected features because it processes high-dimensional data effectively and is relatively memory efficient.

In Figure 2, a flowchart summarizing all the stages of the proposed methodology is demonstrated.

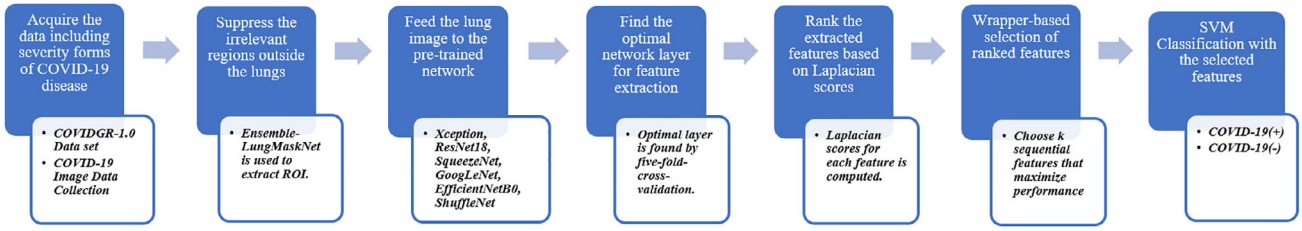


FIGURE 2 The roadmap used in the paper

3 | EXPERIMENTS AND RESULTS

3.1 | Data sets

COVIDGR-1.0 Data set[20] is used in the study. This data set is a perfectly balanced data set which consists of 852 images in total. There are 426 images labeled as COVID-19 positive, and 426 images labeled as COVID-19 negative. COVID-19 disease can have various forms, including mild, moderate, and severe. However, many data sets released during the pandemic generally contain severe forms of the COVID-19 disease. Severe forms of the disease are easily caught by models compared to other forms. As a result, the models may yield high accuracy. Since the mild forms of the disease are not readily caught, the performance may significantly decrease when the data sets contain mild forms of the disease. The *COVIDGR-1.0 Data set* contains 76 images in Normal PCR+, 100 images in mild, 171 images in moderate, and 79 images in severe forms of the disease which makes the data set challenging. This explains why we use this data set in the study.

COVID-19 Image Data Collection[21] is used as a hold-out set in the study. In Reference 22, severity scores for each of 94 X-ray images from the *COVID-19 Image Data Collection*[21] are defined between 0 and 8 by three independent radiologists, each with at least 20 years of experience. Scoring in Reference 22 is done by the experts based on the scoring system introduced in Reference 23. According to the scoring system in Reference 23, the radiological findings of chest radiographs of COVID-19 patients are evaluated in the range of 0–1 in asymptomatic cases (or normal), 1–2 in mild cases, 3–5 in moderate cases, and 6–8 in severe cases. Based on this information, we have thresholded the consolidation scores considering the ranges determined by the experts. In Table 1, it is given that how many samples fall into each severity level of COVID-19 disease.

In Figure 3, the exemplary images from the data set used in the study are given. In the first row of Figure 3, the images are from the *COVIDGR-1.0 data set*, while the images given in the second row are from the *COVID-19 Image Data Collection*. The images given at each column, from left to right, belong to patients with PCR+, mild, moderate, and severe forms, respectively.

3.2 | Performance metrics

The performance metrics used in the study are accuracy (ACC), recall (TPR), precision (PPV), and specificity (SPC). *ACC*, *TPR*, *PPV*, and *SPC* scores are computed as given in (2), (3), (4), and (5), respectively.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \tag{2}$$

TABLE 1 The number of images found at each severity level for the 94 X-ray images from the COVID-19 image data collection

Severity level	Image count
Asymptomatic	15
Mild	30
Moderate	31
Severe	18

$$TPR = \frac{TP}{TP + FN}, \tag{3}$$

$$PPV = \frac{TP}{TP + FP}, \tag{4}$$

$$SPC = \frac{TN}{TN + FP}. \tag{5}$$

3.3 | Performance evaluation

The five-fold cross-validation (CV-5) is used for the performance evaluation of each method in the study. Accordingly, the entire data set is partitioned into five distinct subsets. Then, one of the subsets is used as the test set, while the remaining subsets are used in the training set. Eventually, the loss values obtained on each test fold are averaged to compute the CV-5 loss. The performance evaluation process is demonstrated in Figure 4.

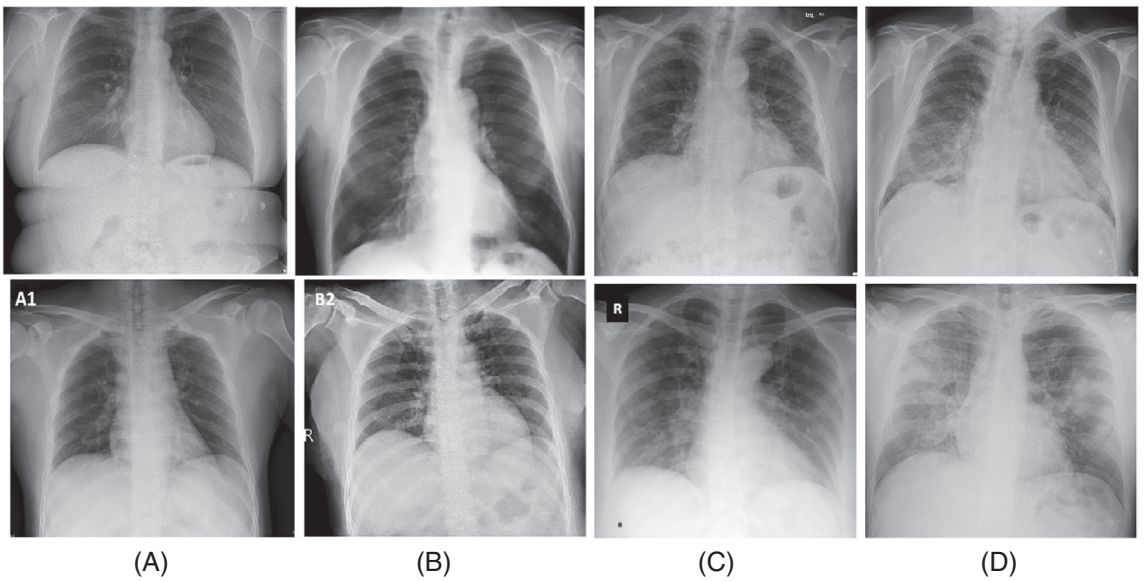


FIGURE 3 The exemplary images used in the study. The images in the first row are randomly selected from the *COVIDGR-1.0 data set* for each level of severity, while the images in the second row are randomly selected from the *COVID-19 Image Data Collection* for each level of severity. Images in columns (A), (B), (C), and (D) correspond to Normal PCR+, mild, moderate, and severe cases, respectively

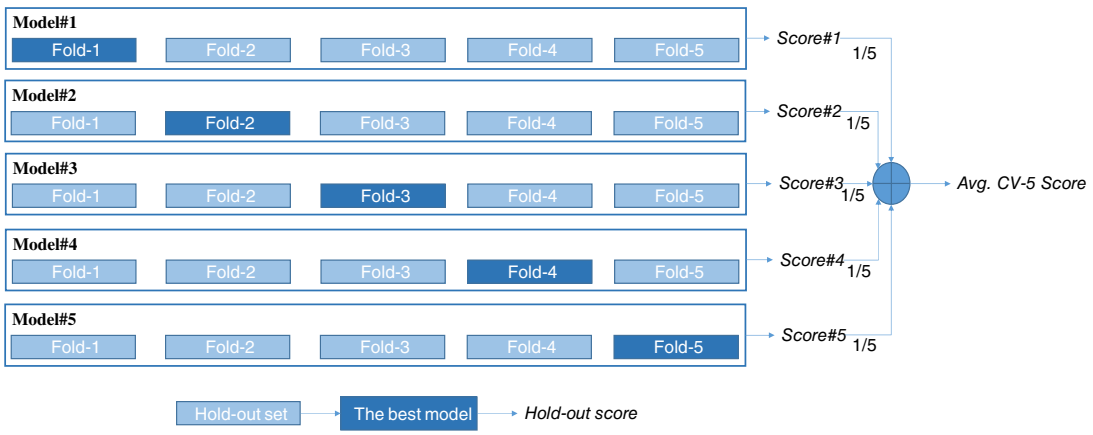


FIGURE 4 The performance evaluation process

3.4 | Classification results

3.4.1 | Classification with pretrained network features

Pretrained networks, namely ResNet18 (18-layers), SqueezeNet (18-layers), GoogLeNet (22-layers), ShuffleNet (50-layers), Xception (71-layers), and EfficientNetB0 (82-layers), are used for feature extraction in the study as they have relatively low complexity. Finding the appropriate network for feature extraction is an experimental process. Instead of directly using each model as a feature extractor, the most suitable layer of the network for feature extraction should be determined. For this purpose, the basic layers that are important in the flow of information (especially concatenation layers) for each network are taken into account. The CNN codes (feature vectors) can be directly obtained by a fully connected layer, but there is a need to convert the feature maps obtained by a convolutional layer to CNN codes. The output form of any convolutional layer for an input image is

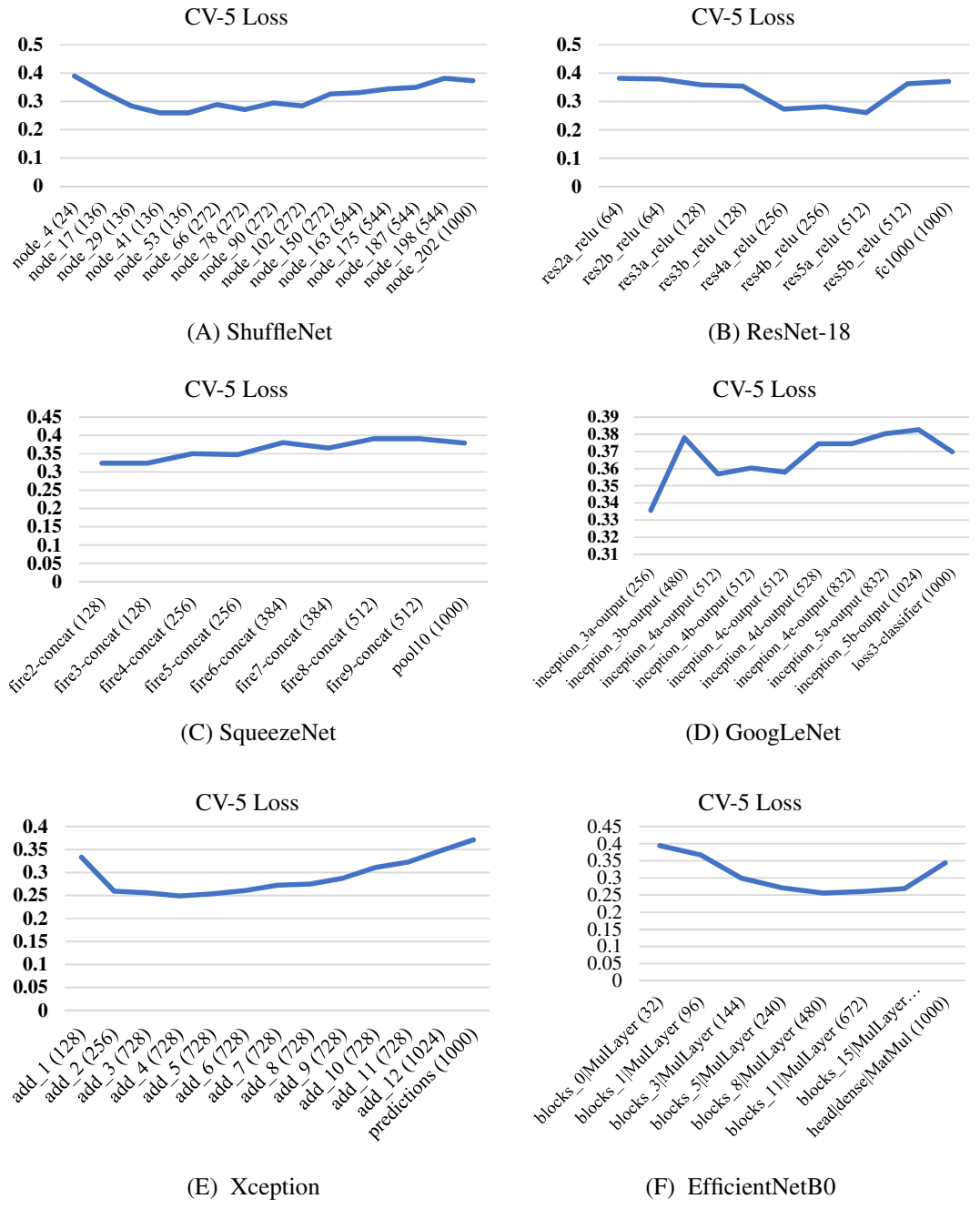


FIGURE 5 Varying values of CV-5 loss depending on the feature extraction layers of the networks. (A) ShuffleNet, (B) ResNet-18, (C) SqueezeNet, (D) GoogLeNet, (E) Xception, (F) EfficientNetB0

$W \times H \times C$ where W, H, and C are the width, the height of an image, and the number of channels in the layer, respectively. This indicates that there are C maps in size $W \times H$. As each channel corresponds to one feature, each map in size $W \times H$ is averaged to represent the map as a single value. As a result, a feature vector in size $1 \times C$ is obtained for every single image.

In Figure 5, the variation of CV-5 error according to the layers of each network used in the study is given. The values given in parentheses show the dimensionality in the vectors obtained from the relevant layer. As seen in almost all graphs given in Figure 5, CV-5 loss values decrease up to a certain layer, and then loss increases in all networks. The best feature extraction layer of the models that is, ShuffleNet, ResNet-18, SqueezeNet, GoogLeNet, Xception, and EfficientNetB0, are node_53, res5a_relu, fire3-concat, inception_3a-output, and blocks_8|MulLayer, respectively. On the other hand, the models i.e. EfficientNetB0 and Xception perform well compared to other models.

In Table 2, the average CV-5 accuracy scores are given with the scores obtained on each test fold for each model. As seen in Table 2, the performance improves as the network complexity increases. The best performance is achieved as 75.11% with the 728 features extracted from the add_4 layer of the Xception model.

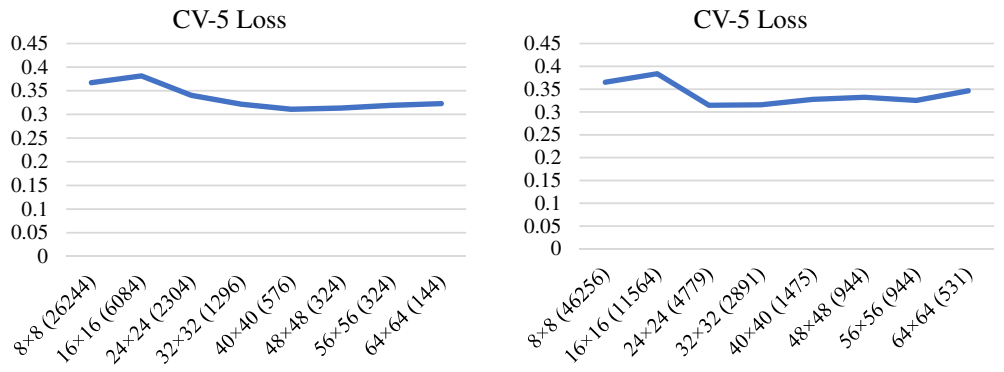
3.4.2 | Classification with hand-crafted features

The hand-crafted feature extraction methods, namely HOG,[17] LBP,[18] GLCM (gray-level co-occurrence matrix),[24] MSER (maximally stable extremal regions),[25,26] SURF (speeded up robust features),[27] Oriented Fast and Rotated BRIEF (binary robust independent elementary features)[28] and are used for comparison in the study as well. Some of these methods are highly dependent on some parameters of each algorithm as previously mentioned. For both HOG and LBP methods, cell size is one such parameter that significantly affects the classification performance. In Figure 6, varying CV-5 loss values versus cell size are given. As seen in Figure 6, increasing the cell size improves the performance up to a certain point. However, the performance decreases after a point. As seen in Figure 6A, this point is 40×40 for the HOG method where 576 features are extracted. On the other hand, it is 24×24 for the LBP method where 4779 features are extracted as seen in Figure 6B.

In Table 3, the classification performance achieved on each test fold is given in detail. Another parameter that can have an impact on the performance of the LBP is the radius (R) as well as the cell size. The effect of the R on the performance of LBP is examined between 1 and 3. The best

TABLE 2 The scores achieved on each test fold by each model trained with the feature set extracted from the optimal layer for feature extraction (the best CV-5 accuracy is marked in bold and italic)

Network	Layer	Features	Size	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Avg.
GoogLeNet	inception_4a-output	512	224×224	64.70	61.98	63.15	61.76	70.00	*64.31*
SqueezeNet	fire3-concat	128	227×227	71.17	70.76	66.08	67.64	62.35	*67.60*
ResNet-18	res5a_relu	512	224×224	72.94	73.10	76.02	74.12	73.53	*73.94*
ShuffleNet	node_53	136	224×224	74.70	70.17	78.36	74.11	72.94	*74.05*
EfficientNetB0	blocks_8|MulLayer	480	224×224	74.11	69.59	77.77	74.11	76.47	74.41
Xception	add_4	728	299×299	72.94	74.85	72.51	78.23	77.05	***75.11***



(A) HOG features-based classification (B) LBP features-based classification

FIGURE 6 Varying values of CV-5 loss depend on the cell size of the HOG and the LBP methods

TABLE 3 The accuracy scores achieved on each test fold by each feature extraction method (the best scores obtained with the parameters of each method are marked in bold and italics)

Method	Parameter settings			Features	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Avg.
HOG	Cell size: 40 × 40			576	67.65	71.35	65.50	68.82	71.18	***68.90***
LBP	Cell size: 24 × 24	R = 1		4779	66.47	73.68	66.08	67.65	68.82	*68.54*
		R = 2			72.94	71.35	65.50	68.82	66.47	***69.01***
		R = 3			69.41	70.76	64.91	68.24	64.12	*67.48*
GLCM	• Symmetrical matrix is computed using four directions • Features extracted from the GLCM: energy, correlation, homogeneity, contrast	D = 1	N = 8	4	59.41	65.50	52.63	60.00	55.29	*58.57*
		D = 2	N = 8		60.00	64.33	52.05	61.18	56.47	*58.80*
			N = 9		60.00	64.33	52.63	62.35	57.65	*59.39*
			N = 10		60.00	64.33	52.63	64.12	56.47	*59.51*
			N = 11		60.00	63.74	52.05	62.94	57.06	*59.16*
		D = 3	N = 8		59.41	65.50	52.63	64.12	57.65	***59.86***
			N = 9		60.59	64.91	52.63	63.53	56.47	*59.63*
			N = 10		60.00	64.33	52.63	62.35	56.47	*59.16*
			N = 11		61.18	64.91	52.63	62.35	55.88	*59.39*
		D = 4	N = 8		58.82	65.50	52.63	60.59	56.47	*58.80*
MSER	-			64	64.71	49.12	56.73	58.24	54.12	*56.58*
				128	64.71	50.29	50.88	67.06	54.12	***57.41***
BRIEF	Number of keypoints	7		7	45.29	47.95	48.54	56.47	52.35	*50.12*
		8		8	50.59	54.97	46.78	55.88	53.53	*52.35*
		9		9	55.29	55.56	53.80	60.00	57.65	***56.46***
		10		10	54.12	55.56	49.71	60.59	57.65	*55.52*
		11		11	54.71	51.46	49.71	58.82	57.65	*54.47*
SURF	-			64	52.35	56.14	47.95	53.53	55.29	***53.05***
				128	49.41	58.48	49.12	54.12	53.53	*52.93*

performance is obtained when the *R* is set as 2. Beyond the HOG and the LBP, one of the other well-known methods for examining the textures of the image is GLCM that measures the spatial relationship between the pixel pairs. Features such as *contrast*, *homogeneity*, *energy*, and *correlation* are extracted from GLCM in the given order and a four-dimensional feature vector is created. The important parameters for the GLCM method that may impact the performance significantly are the distance between the pixels pairs (*D*), and the number of gray levels (*N*). As seen in Table 3, while increasing *D* to 3 increases the classification performance, performance decreases after this point. The classification performance is also analyzed for $D = 2$ and 3 points where the performance increase is achieved. For this purpose, *N* is examined between 8 and 11. As seen in Table 3, classification performance increases as *N* increases up to 10 when *D* equals 2, but performance decreases as *N* increases up to 10 when *D* equals 3. Accordingly, the best performance with the GLCM is achieved when the parameters are set as $D = 3$ and $N = 8$. On the other hand, the classification performance obtained by methods such as MSER, BRIEF, and SURF lags behind even GLCM.

3.4.3 | Classification after feature selection

In this section, redundant features are eliminated following a wrapper-based method. First, the features extracted by each method are ranked based on the Laplacian scores (the importance scores). Then, it is revealed that how many ranked features are required to optimize the performance with the SVM classification. Accordingly, the SVM model is trained for the first ranked feature, the first two ranked features, the first three ranked features, and so on. The obtained results are given in Figure 7. In Figure 7, while the red curves demonstrate the training loss, the blue curves demonstrate the CV-5 loss. All diagnostic curves seen in Figure 7 show indicators of overfitting after a certain amount of ranked features are used. As a

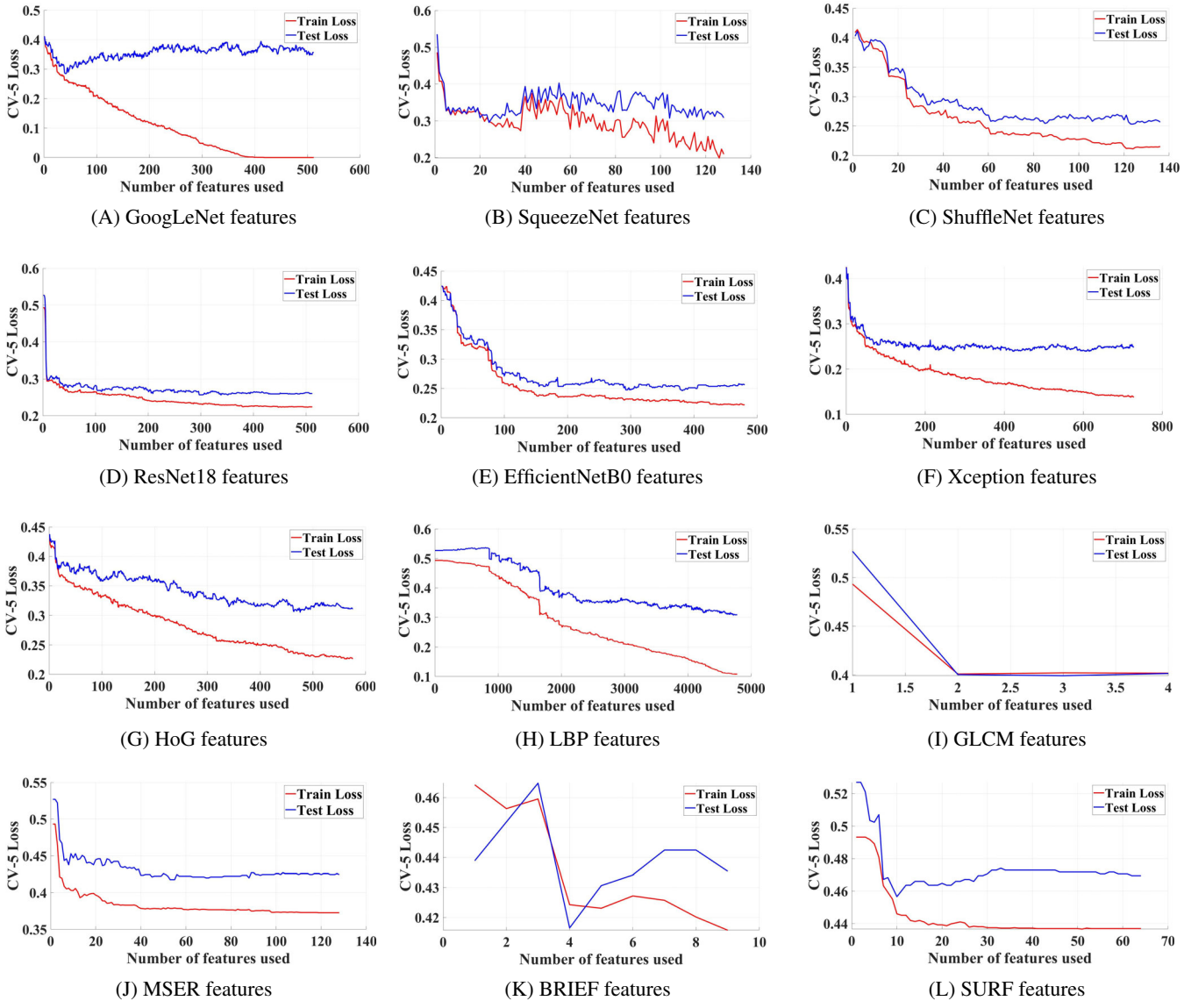


FIGURE 7 The effect of wrapper-based feature selection based on the ranked features on classification performance

result of the feature selection process, the first 10, 53, 4, 3, 476, 4718, 40, 24, 123, 197, 381, and 605 features are selected from the SURF, MSER, BRIEF, GLCM, HOG, LBP, GoogLeNet, SqueezeNet, ShuffleNet, ResNet-18, EfficientNetB0, and Xception feature sets, respectively.

In Table 4, the performance achieved with each method is given after feature selection. As seen in Table 4, feature selection leads to an improvement for all the methods. The feature selection increases the classification performance by 1.29%, 0.82%, 1.88%, 0.24%, 0.58%, 0.36%, 7.17%, 2.82%, 0.6%, 0.36%, 0.94%, and 0.95% with the features of SURF, MSER, BRIEF, GLCM, HoG, LBP, GoogLeNet, SqueezeNet, ShuffleNet, ResNet-18, EfficientNet-B0, and Xception, respectively. Accordingly, the average increase in the classification performance is 1.50%. The best performance is achieved with the Xception features as 76.06%.

In Table 5, the classification results before and after feature selection for Xception network features are given comparatively over the confusion matrix. As seen in Table 5, wrapper-based selection of ranked features increases true positives and true negatives and reduces the total misclassification cost to 204 from 212.

4 | DISCUSSION

Many hand-crafted feature extraction methods were developed in the literature to identify textural patterns within the images. In this work, we have considered well-known methods that is, GLCM, HOG, LBP, MSER, SURF, and BRIEF to catch the patterns related to the COVID-19

TABLE 4 The accuracy scores obtained on each test fold after feature selection (the best CV-5 accuracy is marked in bold and italic)

Feature set		Dimensionality	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Avg.
Hand-crafted	SURF features	10	51.18	59.06	49.12	54.12	58.24	*54.34*
	MSER features	53	64.71	50.29	54.97	67.65	53.53	*58.23*
	BRIEF features	4	63.53	52.63	54.39	60.59	60.59	*58.34*
	GLCM features	3	60.59	66.08	52.63	63.53	57.65	*60.10*
	HOG features	476	66.47	73.10	66.67	70.59	70.59	*69.48*
	LBP features	4718	72.94	72.51	65.50	68.82	67.06	*69.37*
Deep features	GoogLeNet	40	69.41	70.76	69.01	74.12	74.12	*71.48*
	SqueezeNet	24	68.82	72.51	70.76	71.76	68.24	*70.42*
	ShuffleNet	123	75.88	71.35	78.95	74.71	72.35	*74.65*
	ResNet-18	297	74.12	73.68	74.85	75.88	72.94	*74.30*
	EfficientNetB0	381	75.29	73.20	78.36	72.94	77.06	*75.35*
	Xception	605	72.94	75.44	75.44	78.24	78.24	***76.06***

TABLE 5 The confusion matrix represents the classification results before and after feature selection

		Predictions	
N = 852		**COVID-19(+)**	**COVID-19(−)**
Actuals	COVID-19(+)	331/335	95/91
	COVID-19(−)	117/113	309/313

disease. The different textural patterns are caught by each hand-crafted based method with their different assumptions to encode the spatial information. As seen in Table 3, the maximum performance with the GLCM is achieved when the parameters that are, distance between the pixels, and the number of gray levels, are set to three, and eight, respectively. Asides from the GLCM, the other methods namely the HOG and the LBP are histogram-based methods which are aiming to divide the input image into nonoverlapping local regions to encode local spatial information. Decreasing the cell size parameter leads to extracting more detailed features as there will be more local regions. However, it increases the dimensionality at the same time which causes to redundant features come into play. On the other hand, increasing the cell size results in extracting the global features and reduces the dimensionality as the number of local regions is diminished. As seen in Figure 6A, increasing the cell size parameter up to 40×40 for the HOG method reduces the loss value which is again raised after this point. Similarly, as seen in Figure 6B, increasing the cell size parameter up to 24×24 for the LBP method reduces the loss value which is again raised after this point. Accordingly, extracting more or less detail reduces the performance which indicates that there is always a trade-off. These are also demonstrating the parameter dependence of the hand-crafted-based methods. Another thing to mention is that only the spatial relationship between the pixel pairs is taken into account with the GLCM method. Therefore, the achieved performance by the GLCM is poor compared to the LBP, and the HOG. As seen in Table 6, this results in over 9% better performance of the HOG and LBP methods than GLCM. Performance with MSER, BRIEF, and SURF, other hand-based feature extraction methods where intensity-based changes are taken into account, is slightly better than random guessing (especially the SURF method). This suggests that these feature sets are not important enough for the classification task at hand.

Beyond the hand-crafted based feature extraction methods, using a pretrained network as a feature extractor offers an automatic way to feature extraction without requiring any parameter setting. Moreover, spatial information can be better encoded by pretrained networks as many hidden patterns have been already learned. Thus, as seen in Table 6, the classification performance using the features extracted by all the pretrained networks is superior compared to hand-crafted features-based classification performance. As can be seen in Table 6, better performance is achieved with deeper networks such as ShuffleNet, EfficientNetB0, and Xception. This shows that the features obtained from these networks are more distinctive. The best performance is achieved with 76.06% accuracy when Xception network features are used. On the other hand, moderate, and severe forms of the COVID-19 disease are easily captured by each method as shown in Table 6. However, mild forms of the disease are

TABLE 6 Comparing each method with the literature using *COVIDGR-1.0 data set (the best scores are marked in bold and italic)*

Method	TPR_{PCR+}%	TPR_{Mild}%	$TPR_{moderate}$%	TPR_{severe}%	$TPR_{COVID19+}$%	$PPV_{COVID19+}$%	$SPC_{COVID19-}$%	Avg ACC%
Tabik et al.[29]	-	46.00	85.38	97.22	72.59	78.67	79.76	76.18
Lin et al.[30]	***60.00***	***73.68***	***91.43***	***100.0***	***86.05***	***83.10***	***91.30***	***93.02***
MSER features	49.20	56.91	65.23	61.78	57.30	58.40	59.20	58.23
BRIEF features	54.07	53.10	55.84	51.22	63.10	57.60	53.50	58.34
GLCM features	43.53	52.30	66.25	68.08	61.30	59.90	58.90	60.10
HOG features	47.41	64.57	77.39	89.03	68.10	70.00	70.90	69.48
LBP features	45.74	59.78	73.14	92.59	70.20	69.10	68.50	69.37
SqueezeNet features	42.30	62.66	82.27	92.10	68.80	71.10	72.10	70.42
GoogLeNet features	52.23	57.53	78.29	92.02	71.60	71.40	71.40	71.48
ResNet-18 features	35.47	58.16	81.33	95.03	78.90	72.30	69.70	74.30
ShuffleNet features	41.68	61.39	81.85	98.67	76.50	73.80	72.80	74.65
EfficientNetB0 features	31.17	60.14	82.97	96.50	80.00	73.20	70.70	75.35
Xception features	36.92	60.51	85.95	96.36	78.60	74.80	73.50	76.06

TABLE 7 Evaluation of the proposed method on the hold-out set

Method	TPR_{PCR+} 15-images	TPR_{mild} 30-images	$TPR_{moderate}$ 31-images	TPR_{severe} 18-images	$TPR_{covid19(+)}$ 94-images
Xception features (before feature selection)	33.33%	63.33%	93.55%	100%	75.53%
Xception features (after feature selection)	40.00%	66.67%	90.32%	100%	76.60%

not easily detected. The detection of mild cases is extremely important to prevent the spread of the disease. The COVID-SDNet method proposed by Tabik et al.[29] reaches 46% sensitivity in detecting mild cases. Our proposed method with Xception network features yields 60.51% sensitivity which is quite better than random guessing. The cases labeled as PCR+ by the experts indicate that there are no visual signs of the disease on X-ray images. These are also known as asymptomatic. In Table 6, the results achieved by each method confirm this by excepting the work of Lin et al.[30] Lin et al.[30] proposed an adaptive attention-based framework for COVID-19 detection which consists of 311.04 M parameters. The data set used is combined with a pneumonia data set. Combining different data sets is dangerous, as patient demographics may differ. Another issue is that the reported performance scores on the *COVIDGR-1.0 data set* are not based on cross-validation which may lead to over-optimistic results. All these can cause bias in learning. In Reference 30, reported 60% of sensitivity obtained in Normal PCR+ cases that have no signature on X-ray images reveals this bias.

In Table 7, the proposed method with Xception features is evaluated on the hold-out set. In Table 7, the results achieved with Xception features before feature selection are also given to compare. As seen in Table 7, COVID-19 detection sensitivity is increased by 1.07% after feature selection on the hold-out set. At the same time, the detection sensitivity of PCR+ and mild cases increased by 6.67% and 3.34%, respectively. Accordingly, the method generalizes well on unseen data.

5 | CONCLUSION

In this work, a ROI-based classification scheme is proposed for the detection of COVID-19 disease from X-ray images. The proposed method consists of a segmentation network to suppress the regions outside the lungs, a pretrained network as the feature extractor, a wrapper-based feature selection stage using the rankings returned by a filter-based unsupervised method, and the SVM classification. Many works regarding the COVID-19 disease published during the pandemic reported the accuracies of over 90%. The most likely reason for this is that the vast majority of cases in the data sets used in these studies are easily identifiable severe forms of the disease. In this study, our proposed method is cross-validated on a challenging data set containing all the forms of COVID-19 disease. The experimental results on a hold-out set demonstrate that classification with the

selected Xception network features yields good generalization ability. Moreover, the mild cases which are crucial for controlling the spread of the SARS-CoV-2 virus can be identified with the proposed method. One of the weaknesses of our proposed method is that it has a modular form in which some parts (i.e., feature extraction and selection) are individually optimized and then combined. This requires manually feeding the output of one module to the input of another module. By designing an end-to-end trainable deep learning model that will combine the segmentation, feature extraction, and classification tasks, the task at hand can be optimized all at once. Even so, the proposed method in this work offers a simple solution against the many complex methods developed in the literature for COVID-19 detection. Therefore, it may be an assistive tool to triage patients. In future work, we aim to develop a multitask framework that will combine segmentation and classification tasks.

DATA AVAILABILITY STATEMENT

Data available on request from the authors.

ORCID

Coşku Öksüz https://orcid.org/0000-0001-7116-2734

REFERENCES

1. WHO coronavirus (COVID-19) dashboard; 2021. Accessed October 13, 2021. https://covid19.who.int
2. Yeh TY, Contreras GP. 2021. Full vaccination against COVID-19 suppresses SARS-CoV-2 delta variant and spike gene mutation frequencies and generates purifying selection pressure. medRxiv doi: 10.1101/2021.08.08.21261768
3. Li C, Zhao C, Bao J, Tang B, Wang Y, Gu B. Laboratory diagnosis of coronavirus disease-2019 (COVID-19). *Clin Chim Acta*. 2020;510:35-46. doi:10.1016/j.cca.2020.06.045
4. Fan L, Liu S. CT and COVID-19: Chinese experience and recommendations concerning detection, staging and follow-up. *Eur Radiol*. 2020;30(9):5214-5216. doi:10.1007/s00330-020-06898-3
5. Caulley L, Corsten M, Eapen L, et al. Salivary detection of COVID-19. *Ann Intern Med*. 2021;174(1):131-133. doi:10.7326/M20-4738
6. Xue H, Jin Z. The appropriate position of radiology in COVID-19 diagnosis and treatment—current status and opinion from China. *Chin J Acad Radiol*. 2020;3(1):1-3. doi:10.1007/s42058-020-00030-6
7. Alshazly H, Linse C, Barth E, Martinetz T. Explainable COVID-19 detection using chest CT scans and deep learning. *Sensors*. 2021;21(2):455. doi:10.3390/s21020455–476.
8. Zheng C, Deng X Fu Q. 2020. Deep learning-based detection for COVID-19 from chest CT using weak label. medRxiv doi: 10.1101/2020.03.12.20027185
9. Kumar R, Khan AA, Kumar J, et al. Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging. *IEEE Sens J*. 2021;21(14):16301-16314. doi:10.1109/JSEN.2021.3076767
10. Zhang HT, Zhang JS, Zhang HH, et al. Automated detection and quantification of COVID-19 pneumonia: CT imaging analysis by a deep learning-based software. *Eur J Nucl Med Mol Imaging*. 2020;47(11):2525-2532. doi:10.1007/s00259-020-04953-1
11. Mettler FA, Huda W, Yoshizumi TT, Mahesh M. Effective doses in radiology and diagnostic nuclear medicine: a catalog. *Radiology*. 2008;248(1):254-263. doi:10.1148/radiol.2481071451
12. 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
13. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra AU. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput Biol Med*. 2020;121:103792. doi:10.1016/j.compbiomed.2020.103792
14. Mangal A, Kalia S & Rajgopal H et al. CovidAID: COVID-19 detection using chest X-ray. arXiv:200409803 [Cs, Eess]; April 21, 2020. Accessed November 25, 2020. http://arxiv.org/abs/2004.09803
15. Öksüz C, Urhan O & Güllü MK Ensemble-CVDNet: a deep learning based end-to-end classification framework for COVID-19 detection using ensembles of networks. arXiv:201209132 [Eess]. Published online December 9, 2020. Accessed December 20, 2020. http://arxiv.org/abs/2012.09132.
16. Oksuz C, Urhan O & Gullu MK Ensemble-LungMaskNet: automated lung segmentation using ensembled deep encoders. Proceedings of the 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA); 2021:1-8; IEEE. doi:10.1109/INISTA52262.2021.9548367
17. Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); Vol 1; 2005:886-893. doi:10.1109/CVPR.2005.177
18. Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell*. 2002;24(7):971-987. doi:10.1109/TPAMI.2002.1017623
19. He X, Cai D, Niyogi P. Laplacian score for feature selection. In: Weiss Y, Schölkopf B, Platt J, eds. *Advances in Neural Information Processing Systems*. Vol 18. MIT Press; 2006. Accessed October 3, 2021.
20. Ari-Dasci/OD-Covidgr. ARI-DaSCI; 2020. Accessed April 14, 2021. https://github.com/ari-dasci/OD-covidgr
21. Cohen JP, Morrison P, Dao L, Roth K, Duong TQ, Ghassemi M. COVID-19 image data collection: prospective predictions are the future. arXiv:200611988 [Cs, Eess, q-Bio]; December 14, 2020. Accessed March 26, 2021. http://arxiv.org/abs/2006.11988
22. Cohen JP, Dao L, Roth K, et al. Predicting COVID-19 pneumonia severity on chest X-ray with deep learning. *Cureus*. 2020;28:e9448. doi:10.7759/cureus.9448
23. Wong HYF, Lam HYS, Fong AHT, et al. Frequency and distribution of chest radiographic findings in patients positive for COVID-19. *Radiology*. 2020;296(2):E72-E78. doi:10.1148/radiol.2020201160
24. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern*. 1973;SMC-3(6):610-621. doi:10.1109/TSMC.1973.4309314
25. Nistér D, Stewénius H. Linear time maximally stable extremal regions. In: Forsyth D, Torr P, Zisserman A, eds. *Computer Vision – ECCV 2008*. Lecture Notes in Computer Science. Vol 2008. Springer; 2008:183-196. doi:10.1007/978-3-540-88688-4&uscore;14

26. Matas J, Chum O, Urban M, Pajdla T. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vis Comput*. 2004;22(10):761-767. doi:10.1016/j.imavis.2004.02.006
27. Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features (SURF). *Comput Vis Image Understand*. 2008;110(3):346-359. doi:10.1016/j.cviu.2007.09.014
28. Rublee E, Rabaud V, Konolige K & Bradski G ORB: an efficient alternative to SIFT or SURF. Proceedings of the 2011 International Conference on Computer Vision; 2011:2564–2571. doi:10.1109/ICCV.2011.6126544
29. Tabik S, Gomez-Rios A, Martin-Rodriguez JL, et al. COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on chest X-ray images. *IEEE J Biomed Health Inform*. 2020;24(12):3595-3605. doi:10.1109/JBHI.2020.3037127
30. Lin Z, He Z, Xie S, et al. AANet: adaptive attention network for COVID-19 detection from chest X-ray images. *IEEE Trans Neural Netw Learn Syst*. 2021;32(11):4781-4792. doi:10.1109/TNNLS.2021.3114747

How to cite this article: Öksüz C, Urhan O, Güllü MK. COVID-19 detection with severity level analysis using the deep features, and wrapper-based selection of ranked features. *Concurrency Computat Pract Exper*. 2021;e6802. doi: 10.1002/cpe.6802