Received: 25 March 2020 | Revised: 8 July 2020 | Accepted: 11 July 2020

DOI: 10.1002/ima.22469

RESEARCH ARTICLE

WILEY

Classification of Coronavirus (COVID-19) from X-ray and CT images using shrunken features

Şaban Öztürk[1] | Umut Özkaya[2] | Mücahid Barstuğan[2]

[1]Electrical and Electronics Engineering, Amasya University, Amasya, Turkey

[2]Electrical and Electronics Engineering, Konya Technical University, Konya, Turkey

Correspondence
Şaban Öztürk, Electrical and Electronics Engineering, Amasya University, Amasya, Turkey.
Email: saban.ozturk@amasya.edu.tr

Abstract

Necessary screenings must be performed to control the spread of the COVID-19 in daily life and to make a preliminary diagnosis of suspicious cases. The long duration of pathological laboratory tests and the suspicious test results led the researchers to focus on different fields. Fast and accurate diagnoses are essential for effective interventions for COVID-19. The information obtained by using X-ray and Computed Tomography (CT) images is vital in making clinical diagnoses. Therefore it is aimed to develop a machine learning method for the detection of viral epidemics by analyzing X-ray and CT images. In this study, images belonging to six situations, including coronavirus images, are classified using a two-stage data enhancement approach. Since the number of images in the dataset is deficient and unbalanced, a shallow image augmentation approach was used in the first phase. It is more convenient to analyze these images with hand-crafted feature extraction methods because the dataset newly created is still insufficient to train a deep architecture. Therefore, the Synthetic minority over-sampling technique algorithm is the second data enhancement step of this study. Finally, the feature vector is reduced in size by using a stacked auto-encoder and principal component analysis methods to remove interconnected features in the feature vector. According to the obtained results, it is seen that the proposed method has leveraging performance, especially to make the diagnosis of COVID-19 in a short time and effectively. Also, it is thought to be a source of inspiration for future studies for deficient and unbalanced datasets.

KEYWORDS

classification, coronavirus, COVID-19, feature extraction, hand-crafted features, sAE

1 | INTRODUCTION

The Coronavirus (COVID-19), which appeared toward the end of 2019, caused a global epidemic problem that could spread quickly from the individual to the individual in the community. According to the World Health Organization (WHO) data, the rate of catching COVID-19 in China is between 16-21%, and the mortality rate is 2-3%.[1] The clinical symptom must be present, and also positive X-ray and CT images must be included with the positive pathological test to be able to diagnose COVID-19. Fever, cough, and shortness of breath are frequently seen as clinical symptoms.[2] Besides, positive findings should be obtained in X-ray and CT images with these symptoms. Another diagnostic method of COVID-19 is to examine the RNA sequence of the virus. However, this

method is not a very efficient technique due to a long time of diagnosis. Also, its accurate diagnosis rate is not as high as desired. Thus, diagnosis tests should be repeated in lots of time.[3] Radiological imaging techniques are essential for the detection of COVID-19. In X-rays and CT images, the COVID-19 virus shows the same features in the early and late stages. Although it shows a circular distribution within an image, it may exhibit similar characteristics with other viral epidemic lung diseases.[4] This makes it challenging to detect COVID-19 from other viral cases of pneumonia.

Machine learning techniques, which are a sub-branch of the field of artificial intelligence, are frequently used for medical applications in the concept of feature extraction and image analysis. Machine learning methods are used in the diagnosis of viral pneumonia, especially in X-ray and CT images, tumor diagnosis, and cystoscopic image analysis.[5] Viral pathogenic patterns contain several features in X-ray and CT images.[6] COVID-19 shows an irregular distribution and shading within the image.[7] When the X-ray and CT studies in the literature are examined CNN-based methods are used recently.[8,9] Studies with other feature extraction methods are generally seen in studies performed before the CNN method. These studies are highly dependent on the results produced by the methods chosen based on the researcher's experience. Also, it is not always possible to achieve the same performance for different datasets. These methods cover many techniques from edge detection algorithms to GLRLM and SFTA methods. Reference 10 used dissimilarities computed between collections of regions of interest. Then, they classified these features via a standard vector space-based classifier. Reference 11 presented a CT classification framework with three classical types of features (grayscale values, shape and texture features, and symmetric features). They used the radial basis function of the nerve network to classify image features. Reference 12 presented a comparative study using the Jeffries-Matusita (J-M) distance and the Karhunen-Loève transformation feature extraction methods. Reference 13 proposed a classification framework with an average greyscale value of images for multi-class image classification. Reference 14 suggested an automatic classification method to classify breast CT images using morphological features. When these studies with feature extraction algorithms and other similar studies in the literature are examined, it is seen that the proposed methods are mostly successful only on one dataset. Performance decreases when the same operation is done with other datasets.

Besides, interest in early-stage feature extraction methods has begun to decline with the introduction of CNN and other automated feature extraction techniques.[15] Convolutional neural network architecture is a deep learning architecture that automatically extracts and classifies images from images.[16] Reference 17 proposed a hybrid model called fused perceptual hash-based CNN to reduce the classifying time of liver CT images and maintain performance. Reference 18 used a transfer learning strategy to deal with the medical image unbalance problem. Then, they compared GoogleNet, ResNet101, Xception, and MobileNetv2 performances. Reference 19 analyzed the CT scan of lung images using the assistance of optimal deep neural network and linear discriminate analysis. Reference 20 converted raw CT images to low attenuation, raw images, and high attenuation pattern rescale. Then, they resampled these three samples and classified them using CNN.

The methods in the literature have several drawbacks. When CT studies with features are examined, it is seen that the one-way features in these studies show high performance only in the dataset of interest. It also creates a computational load. On the other hand, in CNN studies, a rather sizeable labeled dataset is needed.[21,22] Unbalanced and unlabeled data pose problems for CNN training. Besides, it requires a high hardware capacity. To solve all these problems effectively, a high-performance framework is presented in this study. The classification is made from X-ray and CT images by extracted effective features. Images in the dataset consist of ARds, COVID, No finding, pneumocystis-pneumonia, Sars, and streptococcus classes. As it is known, since COVID is a very new disease, the samples are quite limited. Therefore, the dataset is not suitable for using CNN. Also, the dataset is unbalanced. This situation can cause poor performance not only when using the dataset with CNN but also the classification of features with other classifiers. Although similar problems are solved in the literature with the transfer learning approach, these studies are not suitable for the medical domain. To solve all these problems, feature vectors are created in different spatial planes using four feature extraction algorithms, and a two-stage data augmentation approach is proposed. These four feature vectors created for each image are combined into a single vector. To solve the unbalanced dataset problem, image augmentation and data over-sampling are performed to reproduce the missing number of class vectors. The reason for using these two methods is to prevent the synthetic performance effect of synthetic data generated by the Synthetic minority over-sampling technique (SMOTE) algorithm. When classified with extended features for a small number of observations, noises and various irrelevant information act as disruptors. For this reason, the size of the feature vector is reduced by using a stacked auto-encoder (sAE) and principal component analysis (PCA). The success of these two feature

reduction methods, which work according to different approaches, is compared. Thus, both storage space is saved, and response time is accelerated. Finally, data is classified with the support vector machine (SVM).

This article is organized as follows. Section 2 describes details about the dataset, the introduction of the proposed framework method, and parameters. Section 3 presents experiments and experimental results. Section 4 includes discussion, and Section 5 presents the conclusion.

2 | MATERIAL AND METHODS

2.1 | Dataset description

In the chest X-ray and CT images of patients with COVID-19,[23] there are medium-characteristic infected patterns.[24] Accurate analysis of positive and negative infected patients is essential to minimize the rate of spread of COVID-19. At the same time, the pre-recognition system should have a low level of false-positive alarms to serve more patients. Images in the dataset used to consist of ARds, COVID, No finding, pneumocystis-pneumonia, Sars, and streptococcus classes. The dataset includes 4 ARds images, 101 COVID images, 2 No finding images, 2 pneumocystis-pneumonia images, 11 Sars images, and 6 streptococcus images. Figure 1 contains some sample images from the dataset.

2.2 | Proposed method overview

The proposed framework is shown in Figure 2. Firstly, classical image augmentation is applied to minority classes, and images are rotated, scaled, and so on. The total number of images from after image augmentation 126 increases to 260. Then, robust features are extracted from all images using four different feature extraction techniques. By combining these feature vectors, 78 features are obtained for each image. Then the feature vectors of 260 images consisting of 78 features are over-sampling with the SMOTE method. After this process, 495 feature vectors are created with 78 features. The sAE and PCA architectures are trained with these feature vectors, respectively. The purpose of the sAE and PCA algorithms in this study are to shrunken 78 features to 20 features. Finally, SVM is trained with 495 vectors containing 20 features for classification purposes.

The necessity of both image augmentation and data over-sampling (two-stage data augmentation) arises from the depth of the unbalanced structure in the dataset. In

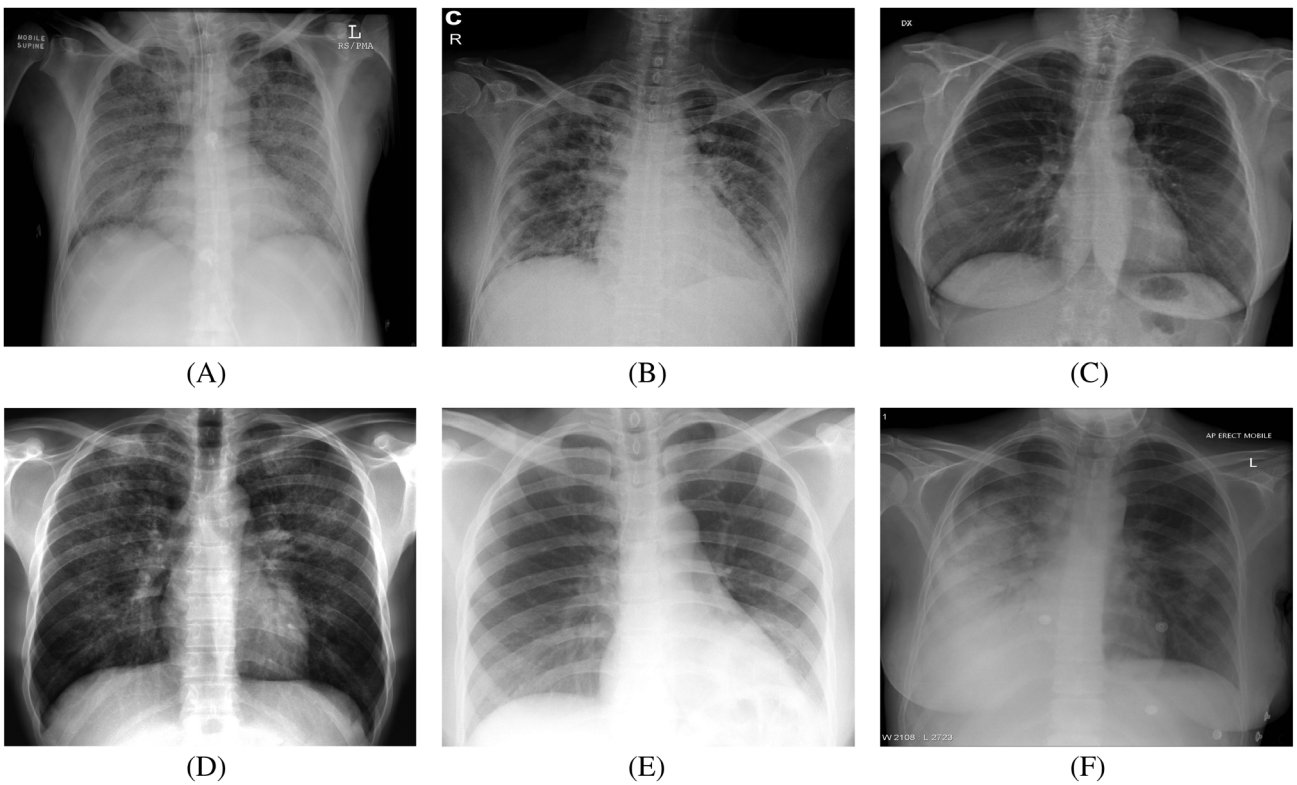

FIGURE 1 Sample images from the dataset, A, ARds, B, COVID, C, No finding, D, Pneumocystis-pneumonia, E, Sars, F, Streptococcus

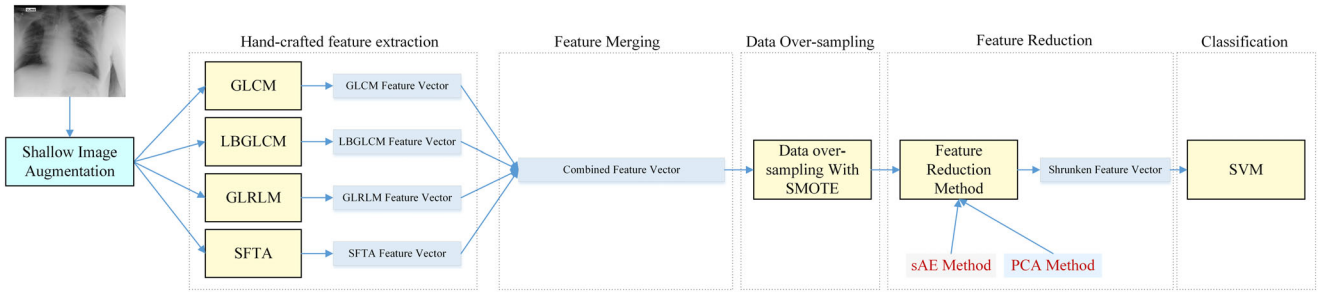

FIGURE 2 Overview of the proposed method [Color figure can be viewed at wileyonlinelibrary.com]

case only image augmentation is applied, there are only two images in many classes, and almost the same images will be produced. In this case, in-class overfitting occurs. When using only synthetic data over-sampling method, synthetic performance data is obtained, and performance may remain low in real applications.

2.2.1 | The feature extraction techniques

Gray level co-occurrence matrix (GLCM), local binary gray level co-occurrence matrix (LBGLCM), gray level run length matrix (GLRLM), and segmentation-based fractal texture analysis (SFTA) features have been extracted to classify pandemic diseases.

Gray level co-occurrence matrix: Features square matrix is defined as *G(i,j)*. Four different directions are used to divide the *G* matrix into normalized typical formation matrices. These directions are defined as vertical, horizontal, left, and right cross directions. These are computed for each of these adjacent directions. These texture features are defined as angular secondary moment, contrast, correlation, the sum of squares, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference entropy, difference variance, information measures of correlation 1, information measures of correlation 2, autocorrelation, dissimilarity, cluster shade, cluster prominence, maximum probability, and the inverse difference.[25]

Local binary gray level co-occurrence matrix: The LBGLCM feature extraction method is a hybrid technique used together with local binary pattern (LBP) and GLCM. LBP technique is applied to the gray-level image. Then, GLCM features are extracted from the obtained LBP texture image. GLCM method takes into account neighboring pixels at the feature extraction stage. It does not perform any operation on other local patterns in the image. Textural and spatial information in the image is obtained together with the LBGLCM method. Simultaneous acquisition of this information increases the availability of the LBGLCM algorithm in image processing applications.[26]

- Initialize V_{SFTA}, T, T_A, T_B

for 1: $T_A \cup T_B$

- I<=multithresholding(I)
- Δ<=findborders(I)
- V_{SFTA}(i)=Boxcounting(I)
- V_{SFTA}(i+1)=Mean(I)
- V_{SFTA}(i+2)=Pixelcounting(I)

end

FIGURE 3 Pseudocode of SFTA [Color figure can be viewed at wileyonlinelibrary.com]

Gray level run length matrix: GLRLM uses higher-order statistical methods to extract the spatial features of gray level pixels. The obtained feature matrix is two-dimensional. Each value in the matrix shows the total formation value of the gray level. GLRLM features are seven in total. These high statistical features are the short-run emphasis, long-run emphasis, gray-level non-uniformity, run-length non-uniformity, run percentage, low gray-level run emphasis, and high gray-level run emphasis.[27]

Segmentation-based fractal texture analysis: In texture analysis, low computing time and efficient feature extraction are critical. SFTA technique is a method that can be evaluated in this concept. In the SFTA method, the image is converted into a binary structure by multiple thresholding technique. Thresholding values of t_1, t_2, t_3,.. t_n are performed. Interclass and in-class variance values are used to determine the threshold sets. To minimize the in-class variance value, the optimum threshold number is applied to the image regions.

Figure 3 shows the feature extraction stages of the Pseudocode SFTA algorithm. V_{SFTA} represents the obtained feature vector. Initially, multiple threshold values (T), all pairs of contiguous thresholds (T_A), and threshold values (T_B) corresponding to maximum gray levels are determined. Then, segmented images pixels,

borders, and V_{SFTA} are updated for all threshold values in a loop. The asymptotic complexity of the obtained V_{SFTA} vector is $O(N \cdot |T|)$. While N expresses the number of pixels, $|T|$ shows the number of different thresholds resulting from the multi-level Otsu algorithm.[28]

2.2.2 | Over-sampling with SMOTE

SMOTE is useful for overcoming unbalanced class distribution problems for classification tasks.[29] The SMOTE algorithm increases the number of samples related to the minority class by producing synthetic samples. To oversamples minority class data is used at a specific rate. This ratio can be selected differently for each class. Let $X = [X_1, X_2, ..., X_n]$ be the feature vectors of each class. The total number of classes is represented by n. If the minority class is represented by X_{minor}, synthetic points are created by determining data points about X_{minor}. K neighborhood value is used to determine data points. According to K-nearest neighbors, Equation 1 is used to calculate a new sample.

$$X^i_{new} = X^i_{minor} + \sigma\left(X^t_{minor} - X^i_{minor}\right), t = 1, 2, ..., K \quad (1)$$

where X^i_{new} represents the new synthetic sample. σ represents a random number uniformly distributed within the range [0,1]. In this study, the number of samples of the majority class is determined to create an almost equal number of samples for minority classes. Accordingly, the "ratio" is calculated, and the samples in the minority classes are reproduced. As the upper ratio limit, eight times the number of samples in the minority class is selected. This criterion applies only to examples in the "no finding" class.

2.2.3 | Stacked auto-encoder

An AE architecture consists of two layers, encoder, and decoder, whose main purpose is to re-interpret the relationship between input and output.[30] In the proposed study, AE architecture expresses the high-dimensional feature vector with fewer parameters. AE architecture, trained in an unsupervised style, creates a relationship between input and output. In AE architecture, more than one AE is added in series. Let $X = \{x_n\}^N$ be an N feature vector. New shrunken feature vectors to be obtained at the sAE output $F{:}X \rightarrow [0,1]^{Nxk}$, is a k-bit vector ($b_n \varepsilon [0,1]^k$) for each feature vector x_n. The output of the AE layer is calculated as in Equation 2.

$$h_l = f(w_l . x_n + b_l) \quad (2)$$

where h_l is the output, w_l represents weights, x_n is nth feature vector, f is an activation function, and b_l represents bias parameter. $f(x) = 1/(1 + e^{-x})$ is used as activation function. The proposed stacked AE architecture is designed to produce 20 features from 78 features. Accordingly, 78 features are reduced to 40 in the first AE structure. These 40 features are applied as inputs to the second AE structure input. At the output of the last AE structure, 20 features are obtained. The sAE part of our framework is presented in Figure 4.

2.2.4 | Principle component analysis

The main purpose of Principle Component Analysis (PCA) is to extract the most significant features from the

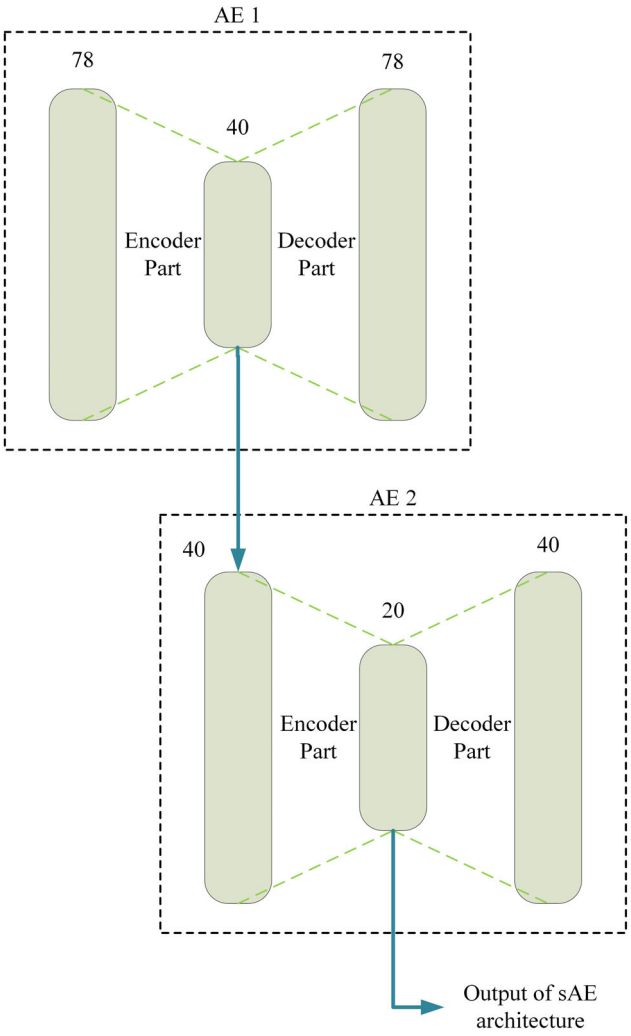

FIGURE 4 The sAE part of the proposed framework [Color figure can be viewed at wileyonlinelibrary.com]

available data. In this way, it ensures that many feature variables are reduced without any loss of information. PCA takes its place in the literature as a linear analysis method. A different coordinate system occurs by rotating the linear combinations of p randomly distributed data ($x1$, $x2$,..., xp) around the original axis. The axes in the new coordinate system show the directions of the highest variability. The primary purpose of performing this coordinate conversion is to provide a better interpretation of the data. In cases where the correlations are quite evident in feature reduction, different spinning techniques may show similar results.[31] Obtained features after the rotation are more meaningful.

2.2.5 | Support vector machines

Support Vector Machines (SVMs) can classify the data with the help of planes. The borders between classes are determined according to these planes. The plane between objects creates a boundary between classes. Linear planes may not show high performance in the classification process. Therefore, it can be necessary to use nonlinear parabolic hyperplanes. Parabolic hyperplanes can be capable of inter-class problem. Figure 5 visualizes the basic working structure of the SVM algorithm. Features are transferred to a different space by using kernel functions. This process is defined as conversion or matching. Thus, features can be distinguished by linear planes in the new space.[32]

3 | EXPERIMENTAL RESULTS

The proposed method is trained on a computer with Intel Core i7-7700K CPU (4.2 GHz), 32 GB DDR4 RAM, and NVIDIA GeForce GTX 1080 graphic card.

The dataset used consists of six classes: ARds, COVID, No finding, pneumocystis-pneumonia, Sars, and streptococcus classes. It contains 126 images in total, including 4 ARds images, 101 COVID images, 2 No finding images, 2 pneumocystis-pneumonia images, 11 Sars images, and 6 streptococcus images. When the image numbers are examined, it is understood that the dataset is a unique dataset for the detection of the COVID virus. However, the numbers of other classes in this dataset cause an overfitting problem in the classification process. For this reason, the number of minority classes should be increased. Because clinical studies related to COVID have just started and the difficulty of labeled data access, it seems impossible to obtain sufficient COVID data for CNN training.

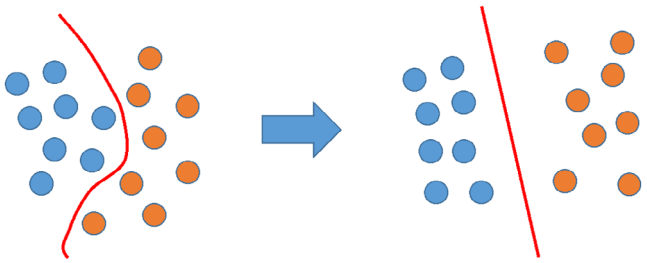

FIGURE 5 Conversion or Matching Process [Color figure can be viewed at wileyonlinelibrary.com]

For this reason, the feature extraction method is most suitable. However, the number of images in the dataset is quite low and the number of sample differences between classes (almost 90% belong to only one class) will create a problem for classifier algorithms. For this purpose, the number of images and data are increased with the image augmentation technique and data over-sampling. Firstly, the number of images in the dataset is increased at the highest possible level, and 260 images in total are obtained. After image augmentation, it consists of a total of 260 images, including 40 ARds, 101 COVID, 24 No findings, 24 pneumocystis-pneumonia images, 43 Sars, and 28 streptococcus images.

Features are extracted from 260 images created using GLCM, LBGLCM, GLRLM, and SFTA feature extraction algorithms. In this case, by selecting the GLCM algorithm offset parameter [2 0], 22 features are extracted for each image, by selecting the LBGLCM algorithm offset parameter [2 0], 22 features are extracted, the GLRLM algorithm generates 7 features, and when the SFTA algorithm feature parameter is selected as 5, it generates 27 features (*6*feature_parameter-3*). When feature vectors of each image are combined, feature vectors consisting of 78 features of each image is obtained. Feature vectors of 260 images consisting of 78 features are increased by using the SMOTE algorithm. The SMOTE algorithm is used according to 6 neighborhood values. Accordingly, the numbers of over-sampled class samples are as follows; for the ARds class, the selected multiplication coefficient is 2, and 80 samples are obtained. For the COVID class, the selected multiplication coefficient is 0, and 101 samples are obtained. For the No findings class, the selected multiplication coefficient is 3, and 72 samples are obtained. For the pneumocystis-pneumonia class, the selected multiplication coefficient is 3, and 72 samples are obtained. For the sars class, the selected multiplication coefficient is 2, and 86 samples are obtained. For the streptococcus, the selected multiplication coefficient is 3, and 84 samples are obtained. As a result of these vector over-sampling operations, the total number of samples is 495.

TABLE 1 The classification results of raw feature vectors

Methods	Features	Sensitivity	Specificity	Accuracy	Precision	F-Score
Raw features with 260 samples	78	75.4%	94.94%	76.92%	74.07%	73.25%
Raw features with 495 samples	78	83.15%	96.96%	86.54%	86.35%	84.55%

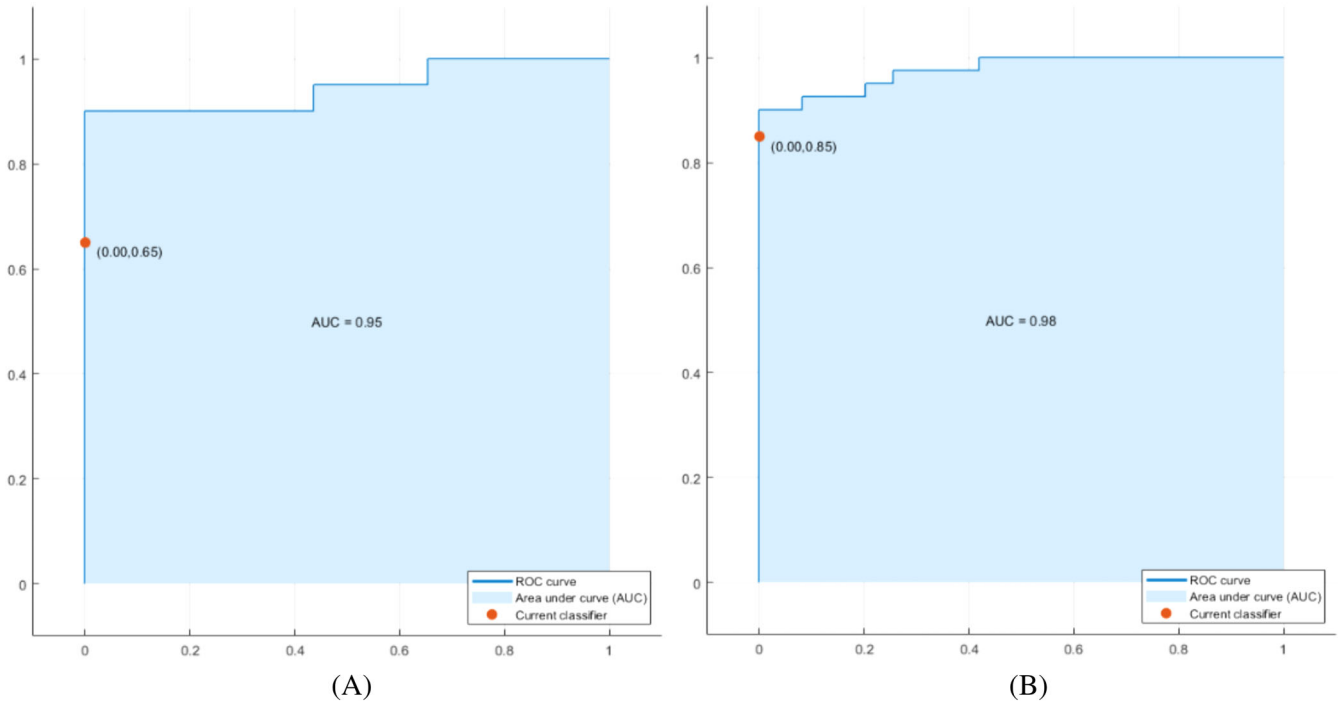

FIGURE 6 AUC curves of raw features, A, 260 samples, B, 495 samples [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 The classification results of shrunken features with sAE

Methods	Features	Sensitivity	Specificity	Accuracy	Precision	F-Score
shrunken features with 260 samples	20	61.04%	92.65%	65.77%	62.62%	61.45%
shrunken features with 495 samples	20	68.91%	93.91%	71.92%	69.89%	69.13%

3.1 | Classification results of raw feature vectors

The proposed framework includes three stages: feature extraction, over-sampling, and shrunken features. For this purpose, experiments are carried out in these three stages. Firstly, the classification results with 260 samples and 495 samples are examined for 78 features in raw form. The contribution of the over-sampling method is investigated by comparing the performance of the classification processes with the SVM algorithm. Table 1 shows the performance parameters of raw feature vectors consisting of 78 features of 260 samples and raw feature vectors of 495 samples created with the SMOTE algorithm. It is seen that increasing the minority classes with synthetic samples has a positive effect on classification accuracy.

Table 1 shows that the classification made with the addition of synthetic classes creates a performance contribution of more than 10%. One of the most important reasons for this is the elimination of imbalance between classes. Another reason is that the number of samples almost doubles. Figure 6 shows the AUC curves of the experiments.

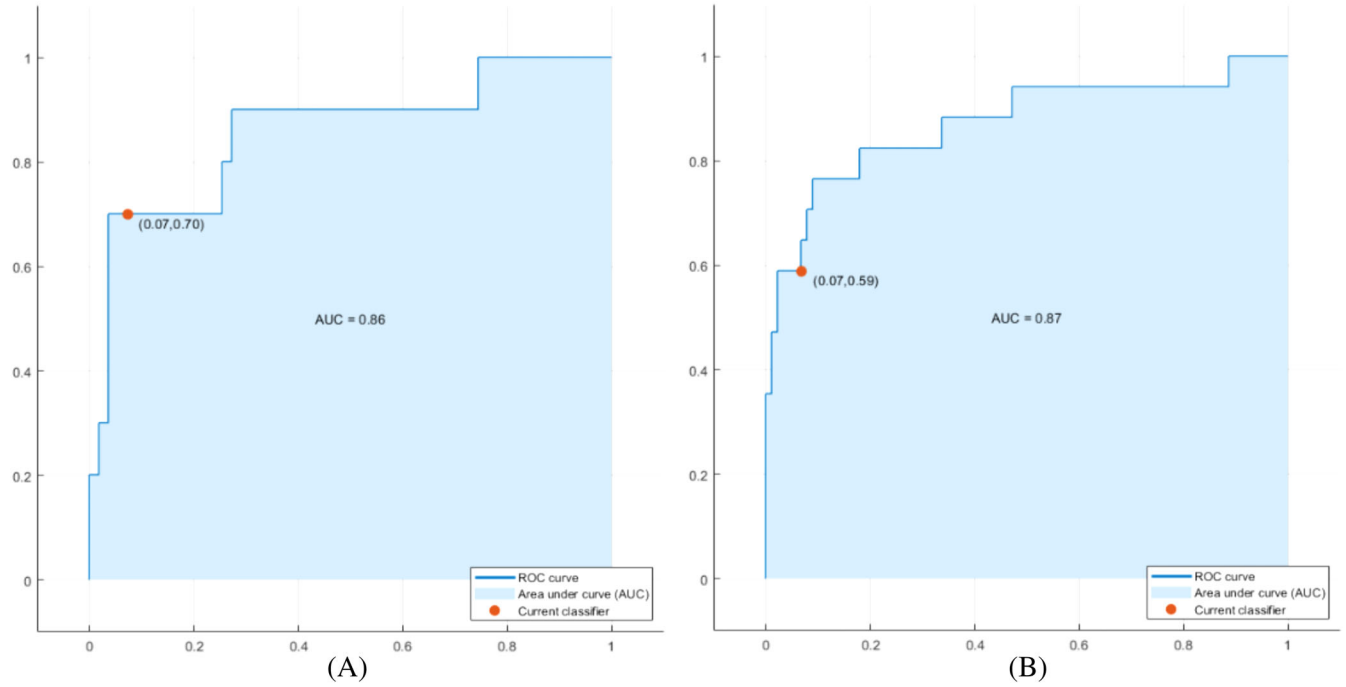

FIGURE 7 AUC curves of shrunken features with sAE, A, 260 samples, B, 495 samples [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 The classification results of shrunken features with PCA

Methods	Features	Sensitivity	Specificity	Accuracy	Precision	F-Score
shrunken features with 260 samples	20	85.08%	97.32%	88.46%	89.75%	87.12%
shrunken features with 495 samples	20	91.88%	98.54%	94.23%	96.73%	93.99%

3.2 | Classification results of sAE

The feature vector in the raw state is quite long, and long feature vectors pose a variety of problems for a limited number of observations. The most important of these is the learning of noises and the problem of sticking to irrelevant points of interest. Other problems are the storage problem and computational complexity. To overcome these problems, the feature vector is narrowed by the sAE method. In this method, which we call shrunken features, sAE reduces the length of the high-length feature vectors. Twenty features are obtained at the recommended sAE output. With these features, two experiments were carried out. In the first experiment, only the results related to 260 samples reproduced with image augmentation, and in the second experiment, 495 sample results obtained with the SMOTE algorithm are examined. Table 2 shows the classification performances obtained using SVM.

When Tables 1 and 2 are compared, it is seen that the classification performance decreases considerably. It is understood that the training process leads to overfitting to training data in sAE architecture with its low number of samples and synthetic data. As seen in the AUC curves in Figure 7, the sAE method cannot be used with these data due to the overfitting problem.

3.3 | Classification results of PCA

Since the sAE method is supervised, it is understood that feature narrowing operation with the insufficient number of samples has failed. For this reason, the PCA algorithm, which performs feature reduction in unsupervised style, is tried. PCA algorithm transforms 78 features to 20 features according to their relations. To examine the balanced dataset effect, classification experiments are applied to the PCA algorithm with 260 samples and 495 samples. Classification accuracy and AUC results obtained are shown in Table 3. Comparing these results with Tables 1 and 2, it is understood that the performance of the PCA algorithm is higher. Considering the

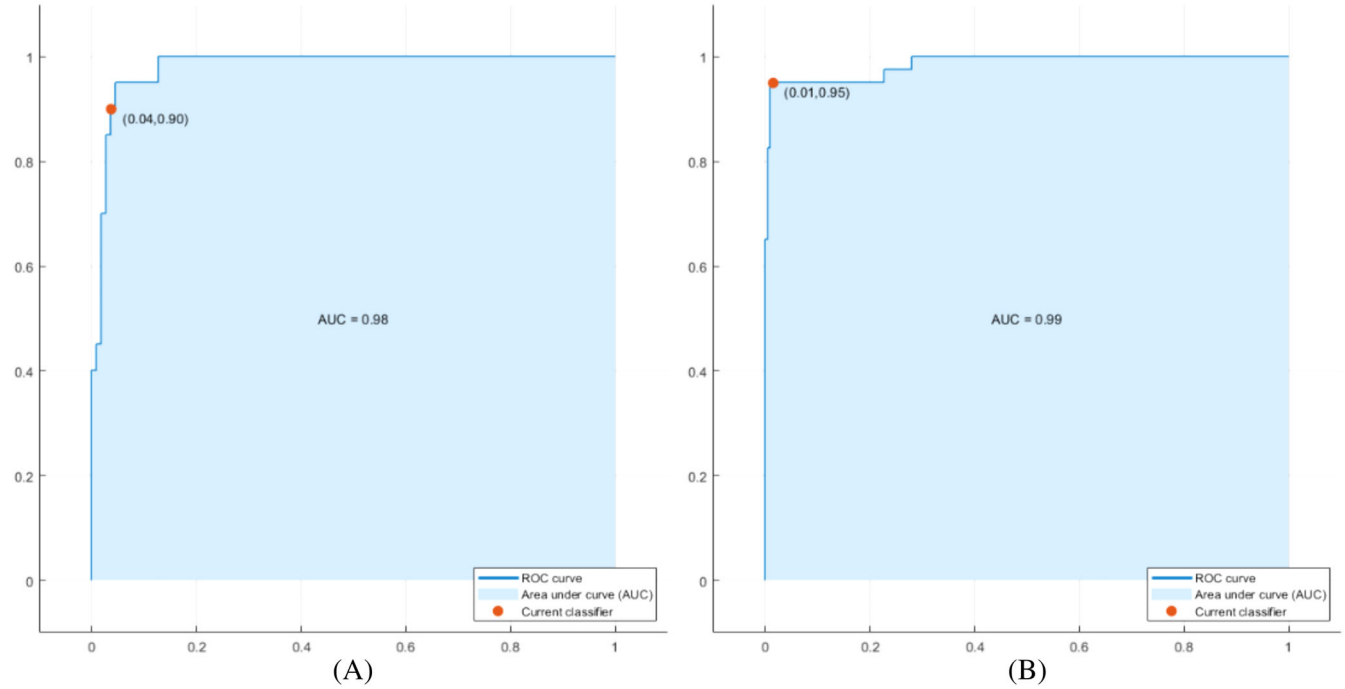

FIGURE 8 AUC curves of shrunken features with PCA. A, 260 samples, B, 495 samples [Color figure can be viewed at wileyonlinelibrary.com]

proposed framework and dataset status, it is seen that the PCA algorithm is more suitable for feature selection. It is thought that its effect will be high, especially in the investigation of viruses such as COVID that started suddenly and in studies with a small number of data.

Table 3 shows that in experiments with 495 samples, it produces more successful results than other experiments, regardless of the number of features. After oversampling, the classification performance increases even more due to the class balance. Figure 8 shows the AUC curve of the features obtained by the PCA method.

4 | DISCUSSION

The proposed method aims to detect sudden outbreaks such as Coronavirus automatically. However, in such cases, it is difficult to find a sufficient number of labeled data. In this case, it is shown that 90% of accuracy performances can be achieved by using features rather than using CNN-based methods. Besides, the contributions of image augmentation and data over-sampling methods are examined for datasets, where the number of samples between classes is quite unstable. The second rows of Tables 1-3 show that the contribution of these methods to performance is almost 10%. Of course, the similarities between real data and synthetic data are quite high, but synthetic data generally only uses the in-class variation. For this reason, classifier performance can be misleading. However, when all comparisons are examined, the common contribution cannot be ignored. The results of the sAE method, which produces favorable results for many studies in the literature, are surprising. The main reasons for this are the low number of samples, the imbalance between classes, and the closeness in synthetic data. When this situation causes in-class affinity, it creates a code generation problem for sAE. It appears that it is not appropriate to use a CNN architecture for the training of such datasets, which is insufficient even for a shallow sAE architecture. Similar to the approach to feature extraction techniques, PCA architecture was considered instead of the sAE architecture. Since the PCA architecture can operate independently from the number of samples, it has provided very successful classification performance.

Studies in the literature generally carry out training with a limited number of examples. The transfer learning approach is used because the number of samples required for the training of studies involving CNN architecture is not sufficient. However, high performance cannot be achieved due to the transfer of trained features with general images. Some studies reduce real image test performance with hard data augmentation. The proposed framework does not use CNN structures due to the COVID dataset containing an insufficient number of samples. Instead, a very powerful feature extraction framework is presented. On the other hand, the proposed

bilateral sample duplication process is the first in the literature. As a result, it is not often appropriate to use today's methods in the supervised style to examine newly emerged datasets containing an insufficient number of samples. Similarly, datasets with an unbalanced class structure are not suitable for training. In cases where it is not possible to find or wait for more labeled data, these datasets can be successfully represented by features. For the unbalanced class problem, performing two-stage data replication instead of a single-sided data replication provides higher performance.

5 | CONCLUSION

Early diagnosis and control of infectious diseases such as COVID-19 are vital for public health. Therefore, automated pre-diagnosis systems are needed to help diagnose the disease quickly and accurately. In this study, we developed a machine learning-based system that can analyze both X-ray and CT images. These days when the access to the Corona data is very limited, experiments are carried out with a dataset with very little data and inter-class imbalance. Due to the limitations of the dataset, feature extraction methods are preferred over deep learning-based methods. The proposed framework performance, which is based on combining feature vectors produced by four feature extraction methods and then reproducing with oversampling and augmentation methods, is very inspiring. It produces beneficial results, especially in terms of comparing sAE and PCA performances. The results of the sAE and PCA algorithms are presented simultaneously, to provide the study to be useful for the researchers who want to work in this field. Besides, the image augmentation and data over-sampling effect are demonstrated in experiments. In future studies, a broader dataset will be produced with more balanced and more labeled COVID-19 data. Also, CNN architectures that can produce a high performance on such datasets will be tried to be developed.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

ETHICS STATEMENT

The paper does not contain any studies with human participants or animals performed by any of the authors.

ORCID

Şaban Öztürk https://orcid.org/0000-0003-2371-8173
Umut Özkaya https://orcid.org/0000-0002-9244-0024
Mücahid Barstuğan https://orcid.org/0000-0001-9790-5890

REFERENCES

1. Yang, Y., Lu, Q., Liu, M., Wang, Y., Zhang, A., Jalali, N. (2020). Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China. medRxiv.
2. Li Q, Guan X, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N Engl J Med*. 2020;382:1199-1207.
3. Zhang N, Wang L, Deng X, et al. Recent advances in the detection of respiratory virus infection in humans. *J Med Virol*. 2020; 92:408-417.
4. Chung M, Bernheim A, Mei X, et al. CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology*. 2020;295(1): 202-207.
5. Negassi M, Suarez-Ibarrola R, Hein S, Miernik A, Reiterer A. Application of artificial neural networks for automated analysis of cystoscopic images: a review of the current status and future prospects. *World J Urol*. 2020. https://doi.org/10.1007/s00345-019-03059-0.
6. Koo HJ, Lim S, Choe J, Choi SH, Sung H, Do KH. Radiographic and CT features of viral pneumonia. *Radiographics*. 2018;38(3):719-739.
7. Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA*. 2020;323(11):1061-1069.
8. Murala S, Wu QJ. Local ternary co-occurrence patterns: a new feature descriptor for MRI and CT image retrieval. *Neurocomputing*. 2013;119:399-412.
9. Gjesteby L, Yang Q, Xi Y, Zhou Y, Zhang J, Wang G. Deep learning methods to guide CT image reconstruction and reduce metal artifacts. *Proc SPIE*. 2017;10132:101322W.
10. Sørensen, L., Loog, M., Lo, P., Ashraf, H., Dirksen, A., Duin, R. P., & De Bruijne, M. (2010). Image dissimilarity-based quantification of lung disease from CT. Paper presented at international conference on medical image computing and computer-assisted intervention, pp. 37–44.
11. Zhang, W. L., & Wang, X. Z. (2007). Feature extraction and classification for human brain CT images. Paper presented at international conference on machine learning and cybernetics, pp. 1155–1159.
12. Homem MRP, Mascarenhas NDA, Cruvinel PE. The linear attenuation coefficients as features of multiple energy CT image classification. *Nucl Instrum Methods Phys Res, Sect A*. 2000;452(1–2):351-360.
13. Albrecht A, Hein E, Steinhöfel K, Taupitz M, Wong CK. Bounded-depth threshold circuits for computer-assisted CT image classification. *Artif Intell Med*. 2002;24(2):179-192.
14. Yang X, Sechopoulos I, Fei B. Automatic tissue classification for high-resolution breast CT images based on bilateral filtering. *Proc SPIE*. 2011;7962:79623H.
15. Khobahi, S., Agarwal, C., & Soltanalian, M. (2020). CoroNet: A deep network architecture for semi-supervised task-based identification of COVID-19 from chest X-ray images. medRxiv.
16. Agrawal, T., Gupta, R., & Narayanan, S. (2019). On evaluating CNN representations for low resource medical image classification. Paper presented at IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 1363–1367.
17. Özyurt F, Tuncer T, Avci E, Koc M, Serhatlioğlu İ. A novel liver image classification method using perceptual hash-based convolutional neural network. *Arab J Sci Eng*. 2019;44(4):3173-3182.

18. Xu G, Cao H, Udupa JK, et al. A novel exponential loss function for pathological lymph node image classification. *Proc SPIE*. 2020;11431:114310A.
19. Lakshmanaprabu SK, Mohanty SN, Shankar K, Arunkumar N, Ramirez G. Optimal deep learning model for classification of lung cancer on CT images. *Future Gen Comput Syst*. 2019;92:374-382.
20. Gao M, Bagci U, Lu L, et al. Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Comput Methods Biomech Biomed Eng*. 2018;6(1):1-6.
21. Narin, A., Kaya, C., & Pamuk, Z. (2020). Automatic detection of coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks. arXiv preprint arXiv:2003.10849.
22. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*. 2020;395(10223):507-513.
23. Cohen, J. P., Morrison, P., & Dao, L. (2020). COVID-19 image data collection. arXiv preprint arXiv:2003.11597.
24. Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill*. 2020;25(3):2000045.
25. Alam FI. Optimized calculations of Haralick texture features. *Eur J Sci Res, Fauquier*. 2011;50(4):543-553.
26. Öztürk Ş, Akdemir B. Application of feature extraction and classification methods for histopathological image using GLCM, LBP, LBGLCM, GLRLM and SFTA. *Procedia Comput Sci*. 2018;132:40-46.
27. Sohail, A.S.M., Bhattacharya, P., Mudur, S.P., Krishnamurthy, S. (2011). Local relative GLRLM-based texture feature extraction for classifying ultrasound medical images. Paper presened at 24th Canadian conference on electrical and Computer Engineering, pp. 001092–001095.
28. Traina, C. Jr., Traina, A. J. M., Wu, L., & Faloutsos, C. (2000). Fast feature selection using fractal dimension. Paper presented at Brazilian Symposium on Databases (SBBD), pp. 158–171.
29. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artificial Intel Res*. 2002;16:321-357.
30. Yuan C, Chen X, Yu P, et al. Semi-supervised stacked autoencoder-based deep hierarchical semantic feature for real-time fingerprint liveness detection. *J Real-Time Image Process*. 2020;17(1):55-71.
31. Ringnér M. What is principal component analysis? *Nat Biotechnol*. 2008;26(3):303-304.
32. Kulkarni SR, Harman G. Statistical learning theory: a tutorial. *Wiley Interdiscip Rev: Comput Stat*. 2011;3(6):543-556.

How to cite this article: Öztürk Ş, Özkaya U, Barstuğan M. Classification of Coronavirus (COVID-19) from X-ray and CT images using shrunken features. *Int J Imaging Syst Technol*. 2021;31:5–15. https://doi.org/10.1002/ima.22469