Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Biomedical Signal Processing and Control 68 (2021) 102656

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Ensemble-based bag of features for automated classification of normal and COVID-19 CXR images

Amira S. Ashour [a,*], Merihan M. Eissa [a], Maram A. Wahba [a], Radwa A. Elsawy [a,b], Hamada Fathy Elgnainy [c], Mohamed Saeed Tolba [d], Waleed S. Mohamed [e]

[a] Department of Electronics and Electrical Communications Engineering, Faculty of Engineering, Tanta University, Egypt
[b] Department of Electronics and Communication Engineering, Alexandria Higher Institute of Engineering &Technology, Egypt
[c] Paxerahealth Company, Smart Village, Cairo, Egypt
[d] Brain Wise, Madinet Nasr, Cairo, Egypt
[e] Department of Internal Medicine, Faculty of Medicine, Tanta University, Tanta, Egypt

ARTICLE INFO

Keywords:
Bag of features
Invariant feature transform
Speeded up robust features detector
K-means
Chest X-ray images
COVID-19
Classification
Ensemble classifiers

ABSTRACT

The medical and scientific communities are currently trying to treat infected patients and develop vaccines for preventing a future outbreak. In healthcare, machine learning is proven to be an efficient technology for helping to combat the COVID-19. Hospitals are now overwhelmed with the increased infections of COVID-19 cases and given patients' confidentiality and rights. It becomes hard to assemble quality medical image datasets in a timely manner. For COVID-19 diagnosis, several traditional computer-aided detection systems based on classification techniques were proposed. The bag-of-features (BoF) model has shown a promising potential in this domain. Thus, this work developed an ensemble-based BoF classification system for the COVID-19 detection. In this model, we proposed ensemble at the classification step of the BoF. The proposed system was evaluated and compared to different classification systems for different number of visual words to evaluate their effect on the classification efficiency. The results proved the superiority of the proposed ensemble-based BoF for the classification of normal and COVID19 chest X-ray (CXR) images compared to other classifiers.

1. Introduction

The COVID-19 pandemic, as announced by the world health organization in 2020, is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which was first informed in Wuhan, China before affecting 218 countries and territories world-wide. Compared to the outbreaks of different coronavirus infections, COVID-19 is considered the most contagious and widespread coronavirus [1]. COVID-19 or coronavirus disease 2019 can spread via several means, most primarily via the droplets and excretions from the infected person, while sneezing, coughing, speaking, or breathing. The recounted symptoms include slight symptoms, such as cough, fatigue, fever, difficulty breathing, and sudden loss of taste and smell to severe complications, such as pneumonia and acute respiratory distress syndrome (ARDS). Molecular test is considered the most common diagnostic test of COVID-19 in comparison to the antigen or antibody tests. However, molecular tests are complex, costly, prone to human errors, and time consuming [2].

Thereby, medical imaging, such as chest X-ray images were approached to assist in the detection of COVID-19 in addition to the clinical symptoms. Chest X-ray (CXR) images allow perceiving the chest pathology via the acquired two-dimensional projection of the patient's chest, which has a pivotal role in the diagnosis of lung diseases and the detection of COVID-19 infection. Compared to the computed tomography (CT) scan, the wide availability and less complexity of the X-ray scan promotes the development of highly applicable computer-aided diagnosis (CAD) frameworks using the acquired CXR images in order to identify and confirm the COVID-19 cases. Accordingly, several studies have adopted machine learning for diagnosing COVID-19 in CXR images. These techniques can be categorized as either deep learning, or traditional machine learning (ML) techniques.

2. Literature review

Several studies have developed deep-learning networks for automated detection of COVID-19. For instance, an optimized convolutional neural network [3] was designed to classify COVID-19, normal, and

* Corresponding author.
E-mail addresses: amirasashour@yahoo.com, amira.salah@f-eng.tanta.edu.eg (A.S. Ashour).

https://doi.org/10.1016/j.bspc.2021.102656
Received 15 January 2021; Received in revised form 30 March 2021; Accepted 16 April 2021
Available online 20 April 2021

pneumonia CXR images, while optimizing the hyperparameters of the convolutional neural network (CNN) using Grey wolf optimization. The results showed 97.78% accuracy, 97.75% sensitivity, and 96.25% specificity. Also, five pre-trained CNN-based models [4], namely the ResNet101, ResNet50, ResNet152, Inception-ResNetV2, and InceptionV3, were proposed for the classification of CXR radiographs into 4classes: COVID-19, bacterial pneumonia, viral pneumonia, and normal leading to classification accuracy ranging between 96.1% and 99.7% among three datasets. The deep CNN CoroNet model [5] was also proposed targeting the same four classes of COVID-19 and pneumonia CXR images using pre-trained Xception network leading to an overall accuracy of 89.6%, 93% precision, and 98.2% recall for COVID-19 detection among the four classes. Moreover, seven different networks of deep convolutional and NN models were included in the COVIDX-Net [6] which targeted the analysis of CXR images into positive or negative COVID-19 cases. The results showed 0.91 and 0.89 F1-scores for normal and COVID-19 cases, using DenseNet and VGG19 models, respectively. Transfer learning was applied in the proposed decompose, transfer, and compose (DeTrac) CNN-based model [7] achieving 95.12% accuracy, 97.91% sensitivity, and 91.87% specificity in the detection of COVID-19 CXR images among normal and severe acute respiratory syndrome cases. Also, adopting transfer learning with several CNN-based models [8] achieved the highest 2-class accuracy of 96.87%, 98.66% sensitivity, and 96.46% specificity using MobileNet v2 for classifying COVID-19 against non COVID-19 cases including normal, VN, and BN. As CNNs are prone to losing spatial information between image instances besides requiring large datasets, COVID-CAPS [9], a capsule networks-based framework was developed to handle relatively small datasets, which achieved 98.3% accuracy, 80% sensitivity, and 98.6% specificity in the four class classification task (i.e. normal, COVID-19, VN, and BN).

Thus, the main limitation of most of the proposed DL-based models the required large datasets that include several alterations of the input images, such as shifting and rotation. However, the availability of large CXR datasets of COVID-19 is still limited. Accordingly, researchers have also adopted traditional machine learning (ML) techniques, for example, a linear support vector machine-based model [10] was proposed for classifying CXR images into healthy or COVID-19. The CXR images were segmented using a multi-thresholding segmentation process into background and several objects of different intensities. Using a dataset of 40 contrast-enhanced CXR images, the suggested system achieved 97.84% accuracy, 99.7% specificity, and 95.76% sensitivity. Also, an ensemble-based support vector machine (SVM) model [11] was implemented for the automated identification of COVID-19 in which the segmentation threshold for the contrast-enhanced CXR images was estimated using Li's method and particle swarm optimization. Subsequently, the texture information was improved using Laws' filter masks, which highlight the micro-structure characteristics, prior to extracting the texture-based feature vector using the gray-level cooccurrence matrix (GLCM). Finally, an ensemble of SVMs using weighted voting was applied in the classification stage yielding to 98.04% accuracy in distinguishing COVID-19 apart of SARS, MERS and ARDS pneumonia. From the previous, despite the few studies adopting traditional ML techniques, promising results have been introduced. Also, to the best of our knowledge, the bag-of-features (BoF) ML models have not been adopted yet in the domain of the CXR image-based COVID-19 diagnosis, despite their efficiency, the BoF has the ability to deal with the changed object's position and orientation. Moreover, from [11], it was deduced that using ensembles in the classification process have led to high classification accuracy in terms of distinguishing COVID-19 apart of other causes of pneumonia, which indicates an expected high performance in classifying CXR images as either COVID-19 or normal, as in our study.

Accordingly, in this paper, we have proposed an automated ensemble-based BoF model with speeded up robust features (SURF) descriptor for the detection of COVID-19 in CXR images using a balanced two-class dataset of normal and COVID-19 cases. The organization of the paper is as follows. Section 2 reports significant related studies in the automated COVID-19 detection based on CXR images. Then, Section 3 introduces the methodology of the proposed ensemble-based BoF framework. In Section 4, the experimental results are reported and interpreted. In Section 5, the proposed system performance is compared to state-of-the-art studies. Finally, the conclusions are presented in Section 6.

3. Methodology of the proposed ensemble-based bag-of-features

Bag of features (BoF), also known as bag of visual words (BoVW) model, is a standalone ML model which is highly efficient in image classification due to its high resistance to the variation in the orientation or the position of the object-of-interest. The main advantage of the BoF is the unneeded segmentation process before the classification stage as it aims to construct a set of visual codewords, also referred to as a codebook or a dictionary, which represents all the possible visual codewords that can be present in the dataset images. The obtained bag of visual words represents a vector of existence counts of a vocabulary of local features in an image without the need of a segmentation process. Hence, given the codebook, an input image can be quantified and represented by a histogram that indicates the occurrence counts of the present visual codewords in the given image. The obtained histograms from the dataset are then used to classify the given images using classification models that are embedded within the BoF model. The BoF model includes three main sequential processes without segmentation, namely local feature extraction, codebook construction and quantization, and classification. The presence of multiple processes in the BoF ML model raises the potential of developing modified models by improving their underlying algorithms. Accordingly, in this study, an ensemble-based BoF for COVID-19 classification-based diagnosis model was proposed by integrating ensemble classifiers at the classification stage of the BoF model. The different processes of the proposed ensemble-based BoF model are discussed below.

3.1. Feature extraction using speeded up robust features descriptor

The feature extraction process in the BoF model is the initial process at which a feature vector is obtained for each determined keypoint without segmentation, which results in a large number of local features for each image. Therefore, the feature extraction process is considered a two-fold process at which: i) the interest points (i.e. keypoints), which represent the feature point locations in each input image are detected, then, ii) feature descriptors are applied to extract the feature vector for each keypoint. In our proposed model, for determining the keypoints and extracting their feature vectors, grid method, and speeded up robust features (SURF) descriptor algorithm [12] were applied, respectively. The interest points were located using the grid method in which a uniform grid with a predefined spacing (i.e. grid step) was applied on the image, such that the intersections of the grid lines determined the locations of the keypoints. In that process, the grid step was set to the size of 8×8.

The SURF descriptor defines the distribution of Haar-wavelet within the locality of the determined interest point. This process is applied using the integral image instead of the original image, which reduces the number of required calculations, which is one of the advantages of using SURF compared to the scale invariant feature transform (SIFT). Thus, given an interest point $X = (x, y)$, where x and $y$represent its $x - axis$ and $y - axis$ coordinates, respectively. The integral image $I_{\sum}(X)$ is calculated as the intensity sum of all the pixels within the rectangular region formed between the pixel and the image origin, as follows:

$$I_{\sum}\left(X\right) = \sum_{i=0}^{i<x}\sum_{j=0}^{j<y} I(i,j) \tag{1}$$

Prior to extracting the descriptor, SURF follows an orientation assignment step at which the horizontal and the vertical Haar wavelet

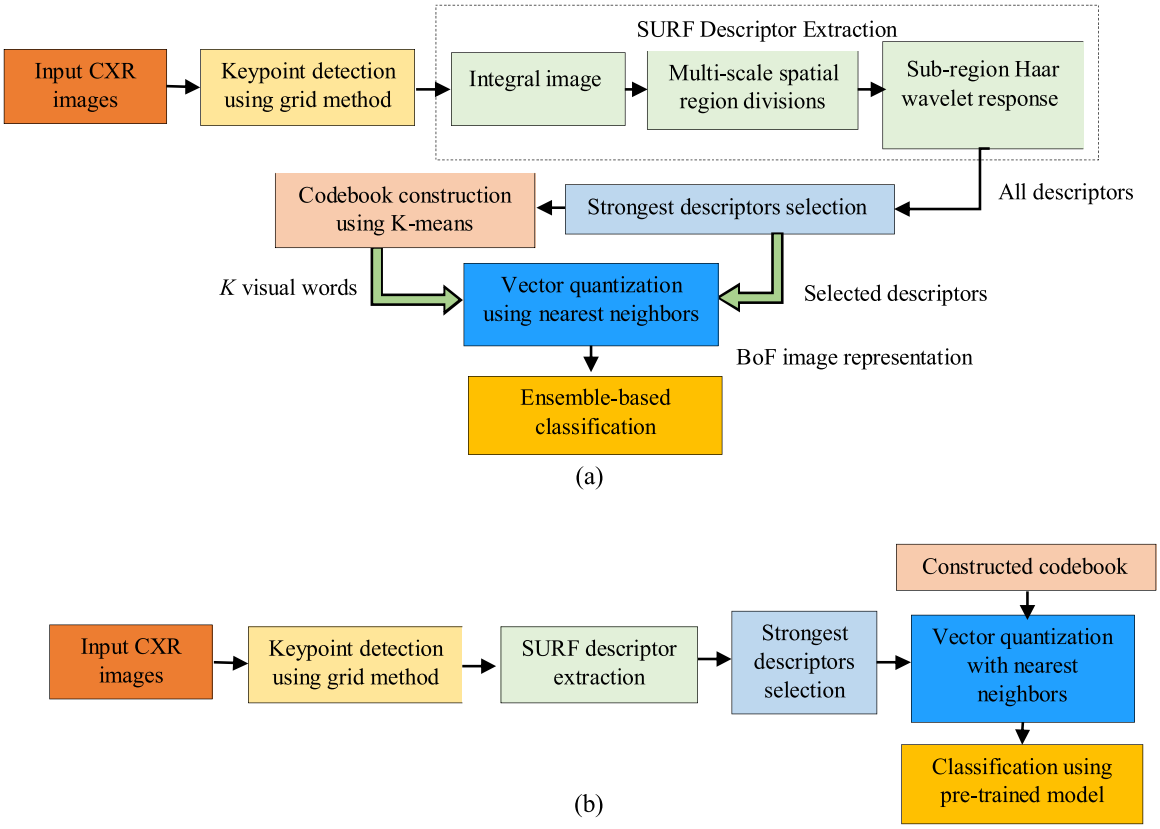

Fig. 1. Block diagram of the proposed ensemble-based bag-of-features classifier for COVID-19 diagnosis model: (a) training phase; (b) testing phase.

responses are calculated in the preset neighborhood of the interest point. Then, the Gaussian weighted sum of these responses is estimated in a sliding orientation window covering an angle of 60° to obtain the dominant orientation. However, this process can be neglected to improve the computational speed. Accordingly, the upright-SURF (U-SURF) was applied in our study in which the image is robust-up to $\pm 15^{\circ}$ change in orientation, while having faster calculations [12].

For feature extraction, first, square patches (i.e. regions) were centered around the detected keypoints. Multiple sized regions were applied in order to extract multiscale features, namely 32×32, 64×64, 96×96, and 128×128. Per each scale, regions were divided into 8×8 square sub-regions, where the vertical and horizontal Haar wavelet responses $d_x$and d_y, respectively were calculated. Thus, forming the four-dimensional sub-region feature vector:

$$FV = \left[\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|\right] \tag{2}$$

where both horizontal and vertical responses were summed over the sub-region in the first two entries. Also, the sums of the absolute responses $\sum|d_x|$and $\sum|d_y|$ are considered to represent the intensity changes' polarity. Thus, per each scale, a.

3.2. Codebook construction & quantization using K-means algorithm

The keypoint feature extraction process has generated a vast number of features by having a 64-dimensional feature vector for each region per each scale. These features are initially reduced prior to constructing the codebook. For feature-space reduction, the variance of the extracted descriptors is computed, then, the 80% of the strongest descriptors (i.e. having the highest score) are selected. Next, these features were quantized using K-means clustering algorithm to construct the visual vocabulary, which comprises *K*visual words. Using K-means clustering, the obtained descriptors are grouped into *K*clusters, such that the cluster centers represent the *K*visual words. For obtaining the K visual words, first, *K*initial cluster centers are randomly selected based on the inputted *N*descriptors. Then, the Euclidean distance between each of the *N* descriptors (points $\boldsymbol{d}_i$) and each of the *K*initial cluster center is calculated, as expressed by the following equation:

$$D_E = \underset{\substack{j=1,\ldots\ldots,K \\ i=1,\ldots\ldots,N}}{\boldsymbol{argmin}} \left\|\boldsymbol{d}_i - \boldsymbol{C}_j\right\|^2 \tag{3}$$

where $\boldsymbol{C}_j$ represents the cluster center. Thus, the descriptor is then assigned to the nearest cluster center. Subsequently, the new cluster centers are calculated in addition to the Euclidean distances between the descriptors and the new cluster centers. This process is repeated iteratively, while reducing the sum of the squared Euclidean distances, until the cluster centers became steady. The final obtained cluster centers represent the *K*visual words (i.e. codewords) of the BoF codebook.

After constructing the codebook, input images are represented by a histogram that indicates the frequency of occurrence of the *K*visual words within the image. This vector quantization process is performed using the nearest neighbors algorithm based on the Euclidean distance measure, which assigned each extracted descriptor to its nearest codeword. Hence, the histograms of the input images represented the distribution of visual content using the constructed codebook. Hereby, these histograms were then exploited by ML algorithm to classify the input CXR images into either normal or COVID-19 case.

3.3. Classification using ensemble-based models

Ensemble-based models integrate a set of classifiers for producing a superior classification performance compared to the performance of each individual classifier, thus, reducing the poor selection possibility. The ensemble-based models are classified into classifier selection models, where only the output of the best performing classifier is

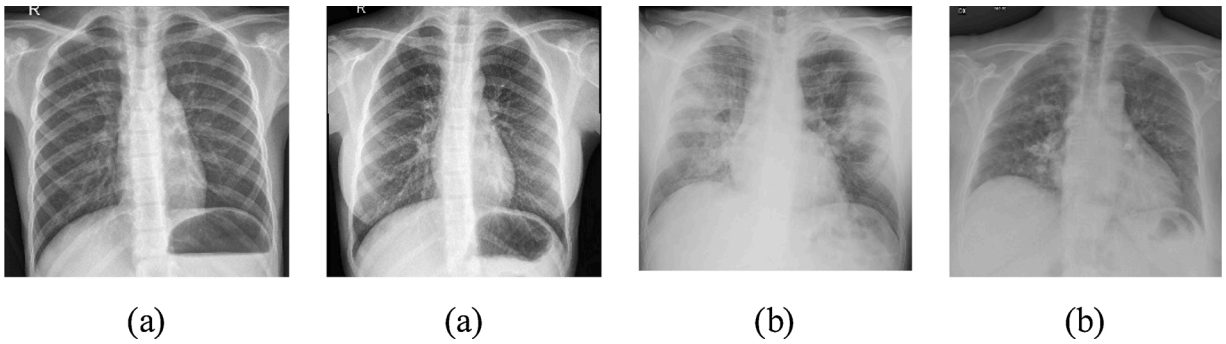

Fig. 2. Sample CXR images from the dataset: (a) normal CXR images; (b) COVID-19 CXR images.

selected as the final classification output; or classifier fusion models, where predefined rules are applied for combining the outputs of the individual classifiers to obtain the final decision. The number of classifiers in the ensemble, i.e. the ensemble size, is selected based on the tradeoff between the classification accuracy and speed, where large ensembles produce better classification results on the expense of longer training and prediction time. Ensemble learning applies several approaches in integrating the several models, such as: i) bagging (bootstrap aggregation) at which a set of models are trained using randomly sampled subset of the training points, while containing all the feature set, then the obtained predictions are aggregated using averaging for obtaining the final output, ii) boosting which weights the constructed models based on their performance by focusing on the misclassified data and assigning them higher weights to reduce classification errors, and iii) random subspace which uses random feature subsets to train each learner, then, the outputs of the different models are combined by majority vote or by combining the posterior probabilities.

In bagging or bootstrap aggregation, weak learners, such as decision trees are trained using random subsamples of the training points. Decision trees are sensitive to the input data, hence, if the training data is changed, the obtained predictions will significantly change. Thus, the bagging approach aims to reduce the high variance of the decision trees and reduce their bias by training a number of decision trees with different data samples and then combining the predictions of the multiple trees into a final decision instead of depending on one individual tree. Nevertheless, random subspace algorithm requires less computational time compared to the bagging and boosting methods, as each learner is trained using a subset of the feature space instead of all the features. Subspace discriminant ensemble engage linear discriminant analysis (LDA) to determine a low-dimensional discriminant subspace [13]. In this study, ensemble-based classification models using bagged trees ensembles and subspace discriminant ensembles were investigated in the classification layer of the proposed BoF model.

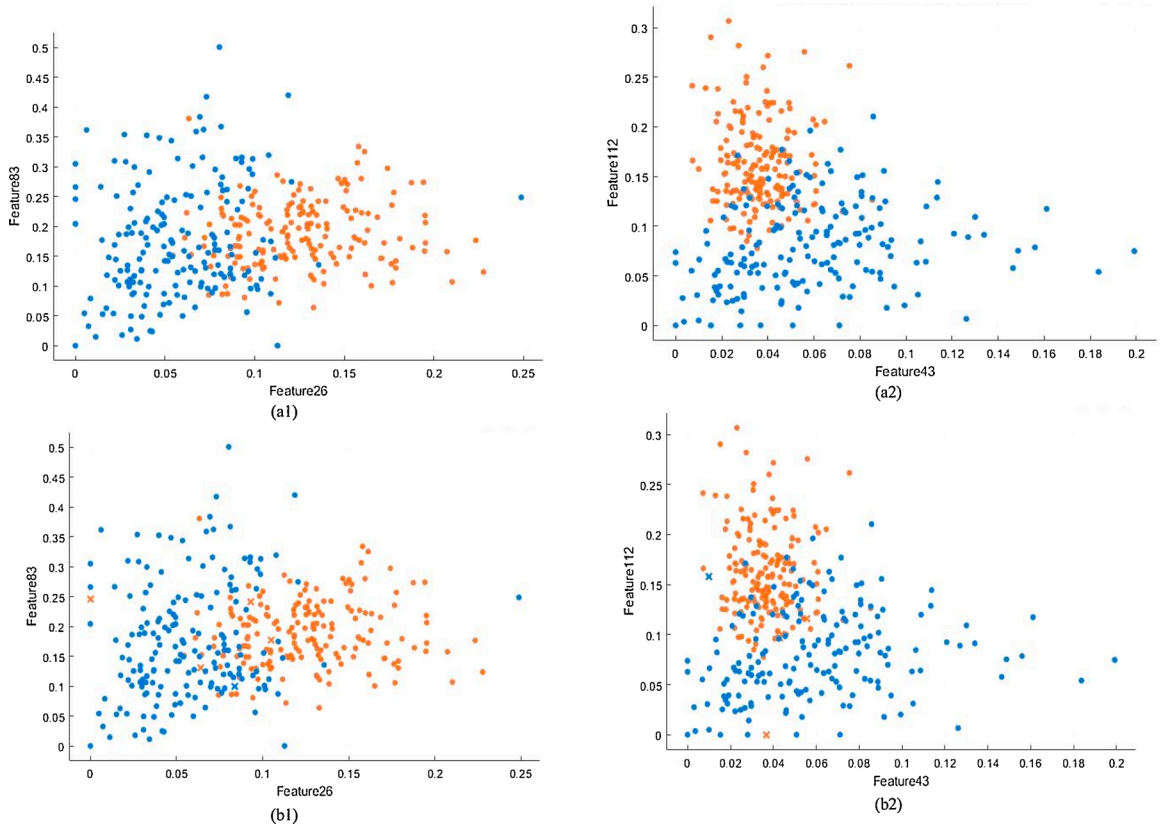

Fig. 3. Scatter plot for two different feature pairs using $K = 200$: (a) original feature pairs; (b) classified feature pairs using the ensemble subspace discriminant model.

3.4. Proposed ensemble-based bag-of-features COVID-19 detection model

The proposed ensemble-based bag-of-features COVID-19 diagnosis model initially undergoes a training phase using the labeled training CXR images to construct the codebook of the BoF model in addition to train the classification model for setting its parameters. Fig. 1(a) demonstrates the sequential processes that were carried out during the training phase, which encompassed the keypoint detection using grid method followed by the SURF descriptor extraction. The descriptor extraction process was based on the calculated integral image, which was divided into spatial regions and sub-regions using multi-scale block sizes. Accordingly, Haar wavelet response vector was extracted from each sub-region, which resulted in a vast number of descriptors per each image. Hereby, 80% of the top (i.e. strongest) features were selected based on their variance, as the increase in the descriptor variance among classes indicates a high distinctive descriptor. Afterward, the codebook was constructed using the K-means clustering algorithm to obtain a preset number of visual words K. Subsequently, the selected descriptors were quantized using the nearest neighbors algorithm to assign each descriptor to its nearest visual word and produce the histogram indicating the frequency of occurrence of each visual word in the image. Finally, the obtained histograms for the training images were used to train the classification model to discriminate the COVID-19 CXR images and the normal cases.

In the testing phase, as illustrated in Fig. 1(b), the input CXR images were applied to the keypoint detection algorithm for extracting the SURF descriptors from the detected keypoints. The selected strongest features were quantized using the constructed codebook to obtain the frequency of occurrence histograms for the testing images. Finally, the pre-trained classification model produced the classification decisions regarding the input testing CXR images.

4. Experimental results and discussion

In this study, an open publicly-available dataset of CXR images was applied for training and testing the proposed system [14]. A 4 GB GPU, Intel Core i7 desktop was used to execute the MATLAB software for evaluating the proposed system. At the access date, the dataset consisted of 400 CXR images, including 200 COVID-19 cases and 200 normal cases, which were collected from public sources in addition to hospitals and physicians. Fig. 2 displays a sample of the dataset images.

The CXR images were separated into training and testing sets using the five-fold cross-validation technique by splitting the dataset into five equal folds to find the classifier's overall performance as the average of the five runs. The training images followed the process demonstrated in Fig. 1(a) for keypoint detection and SURF descriptor extraction. These features were quantized into Kvisual words according to the number of codewords that were defined at the codebook construction process. In our study, we investigated the effect of changing Kon the proposed system performance. The classification metrics of accuracy, sensitivity, precision, specificity, F-measure, and area under curve (AUC) were estimated at $K = 150$ and $K = 200$. Generally, increasing the number of codewords lead to better classification performance due to the presence of more distinctive features. However, this comes on the expense of the computational time. Also, different classification models were compared to the proposed ensemble-based classification in the BoF model, which includes the default linear SVM, and several nearest neighbor KNN algorithms. Also, two different ensemble-based classification models were studied: the ensemble bagged trees model and the ensemble subspace discriminant model. Fig. 3 demonstrates the scatter plot of a two different feature pairs using $K = 200$ in the first row, and their classification results using the ensemble subspace discriminant classifier are visually demonstrated in the second row, where the cross mark represents the misclassified samples.

From Fig. 3, the foremost aim of the selected classification method is to decrease the number of the misclassified samples in the given two

Fig. 4. Classification performance metrics of subspace discriminant ensemble at $K = 150$ and $K = 200$.

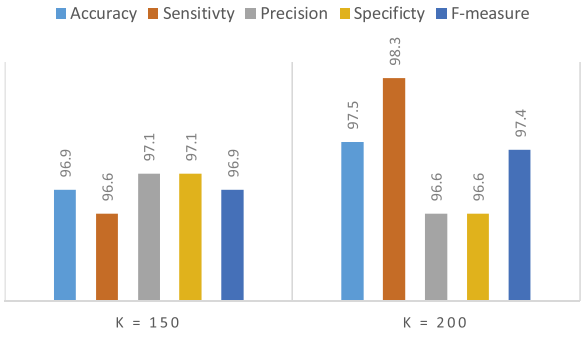

Fig. 5. Classification performance metrics of bagged trees ensembles at $K = 150$ and $K = 200$.

classes. Different classification models were investigated at different numbers of visual words to select the classification model having the optimal performance metrics. Figs. 4 to 11 demonstrates the effect of increasing the number of visual words from $K = 150$ and $K = 200$ on the different classification models, namely subspace discriminant ensemble, bagged trees ensemble, linear SVM, cosine KNN, fine KNN, medium KNN, coarse KNN, and weighted KNN, respectively. The performance metrics of the classification models were evaluated according to the true negative (TN), true positive (TP), false positive (FP), and false negative (FN) rates. The positive class represented the COVID-19 class, where the negative class represented the normal cases. The accuracy was calculated as the percentage of the truly detected images (TP and TN) relative to the entire number of input images. The sensitivity or recall or true positive rate was calculated as the percentage of the positive cases which were truly detected. Precision determined the percentage of the truly detected positive cases relative to all the detected positives. On the other hand, the specificity determined the true negative rate, which is the percentage of the truly identified normal cases. Finally, the F-measure was calculated to represent the weighted average of precision and recall.

Fig. 4 displays that increasing the number of visual words from $K = 150$ to $K = 200$ increases the classification accuracy, precision, specificity, and F-measure metrics of the subspace discriminant ensemble by nearly 2% each, reaching 98.6%, 97.7%, 97.7%, and 98.6%, respectively. Also, an approximate increase of 1% was achieved in the classifier sensitivity reaching 99.4% at $K = 200$. However, for bagged trees ensembles in Fig. 5, the effect of the increase of $K$was obvious only positive on the accuracy, sensitivity, and F-measure metrics. On contrary, a negative effect was intercepted on the precision and specificity metrics, which indicates the increase in the number of false positive cases, while the reduction in the number of false negative cases.

Thus, Fig. 5 illustrates that the rise in the number of visual words from $K = 150$ to $K = 200$ increases the classification accuracy and F-measure metrics of the bagged trees ensemble by nearly 0.7% each reaching 97.5% and 97.4%, respectively, while, increasing the

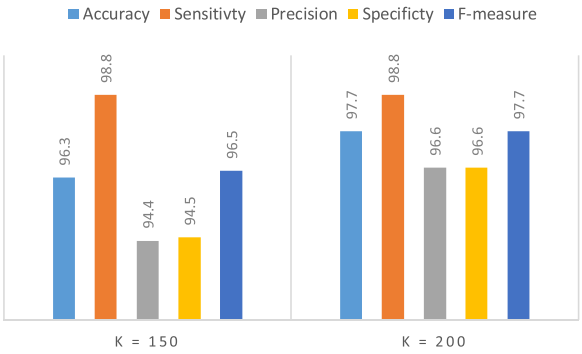

Fig. 6. Classification performance metrics of linear SVM at $K = 150$ and $K = 200$.

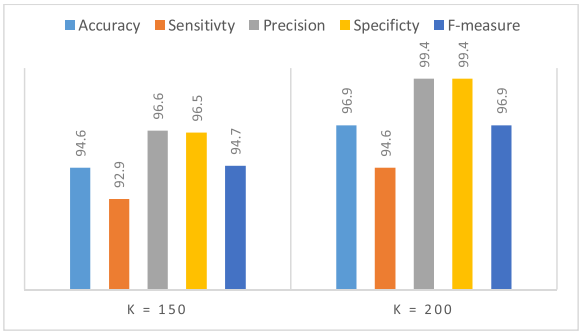

Fig. 7. Classification performance metrics of cosine KNN at $K = 150$ and $K = 200$.

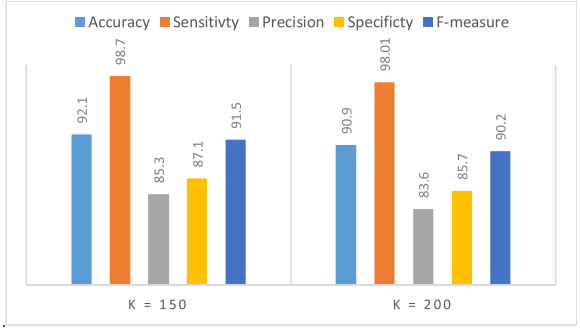

Fig. 8. Classification performance metrics of fine KNN at $K = 150$ and $K = 200$.

sensitivity by nearly 2% reaching 98.3%. However, both the precision and specificity scored an approximate reduction of 0.7% from 97.1%–96.6%.

Fig. 6 demonstrates the classification metrics of linear SVM at both $K = 150$ and $K = 200$, which reveals an approximate 2% increase in both the classification precision and specificity from 94.5%–96.6%. Also, the accuracy and F-measure metrics increased by nearly 1.5% each, thus, both reached 97.7%. While, the sensitivity remained steady at 98.8%, which reveals that the number of false negative and true positive remained steady with the increase of K. Figs. 7 to 11 demonstrate the classification metrics at both $K = 150$ and $K = 200$ for different types of K-nearest neighbor classifiers, namely cosine, fine, medium, weighted, and cubic KNNs.

Fig. 7 demonstrates that all the classification metrics of the cosine KNN increased with the increase of $K$from 150 to 200. The accuracy, sensitivity and F-measure have increased by an average of 2% reaching 96.9%, 94.6% and 96.9%, respectively. Moreover, both the precision and specificity have increased from nearly 96.5%–99.4% with the

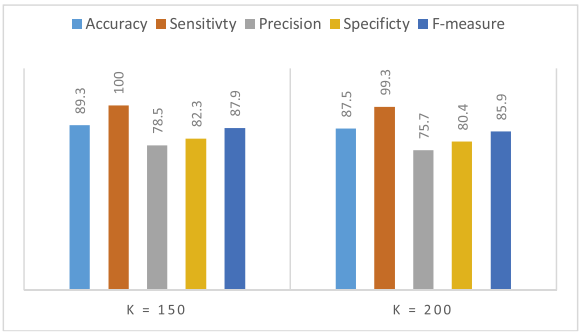

Fig. 9. Classification performance metrics of medium KNN at $K = 150$ and $K = 200$.

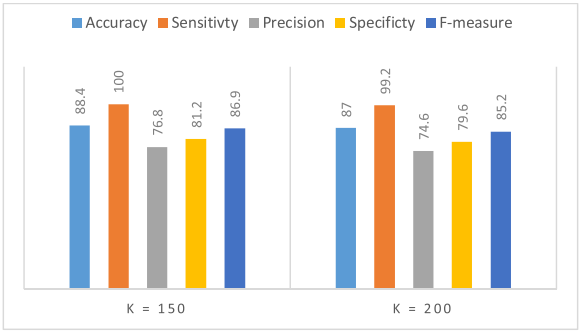

Fig. 10. Classification performance metrics of weighted KNN at $K = 150$ and $K = 200$.

increase of KFrom Figs. 4 to 7, it can be concluded that the increase of K resulted in improving the performance of the ensemble-based, linear SVM and cosine KNN classification models. This is because as the value of Kincreases, the number of cluster centers increases, and fewer descriptors will belong to the cluster. Thus, the average distortion in assigning the descriptors will decrease and descriptors would be more precisely assigned to clusters leading to more discriminant cluster points (i.e. visual words). Nonetheless, Figs. 8 to 11 representing the fine, medium, weighted and cubic KNNs indicate the decrease in classification performance with the increase of K This result occurs due to the increase of the number of visual words leads to less occurrence levels in the histogram of the BoF, which leads to closer sample points. The KNNs categorize the given sample points into different classes based on their distance to neighboring points. Hence, the closer proximity occurring with the increase of Kleads to less distinction in the classification process, with the exception of the cosine KNN which achieved better performance with the increase of K In that case, the sample points are assigned to the nearest neighbors based on a cosine distance metric, which calculates the included angle between the sample points instead of the Euclidean distance.

Fig. 8 illustrates that the rise in the number of visual words from $K = 150$ to $K = 200$, decreases the accuracy, precision, specificity, and F-measure of the fine KNN by nearly 2% reaching 90.9%, 83.6%, 85.7% and 90.2%, respectively. However, the sensitivity has roughly remained steady at 98%. As the fine KNN considers the fine detailed distinctions between classes, better results were obtained using the fine KNN compared to the medium KNN in Fig. 9 at which the coarser distinctions are observed between classes.

Fig. 9 demonstrates that increasing the number of visual words from $K = 150$ to $K = 200$ decreased the accuracy, sensitivity, precision, specificity, and F-measure of the medium KNN reaching 87.5%, 99.3%, 75.7%, 80.4% and 85.9%, respectively with an average decrease of more than 2%.

Fig. 10 demonstrates that increasing the number of visual words

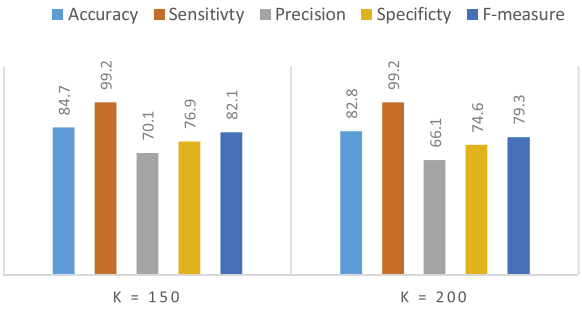

Fig. 11. Classification performance metrics of cubic KNN at $K = 150$ and $K = 200$.

from $K = 150$ to $K = 200$ decreased the accuracy, sensitivity, precision, specificity, and F-measure of the weighted KNN reaching 87%, 99.2%, 74.6%, 79.6% and 85.2%, respectively with an average decrease of less than 2%. Conversely, in Fig. 11, the accuracy, sensitivity, precision, specificity, and F-measure of the cubic KNN have reached 82.8%, 99.2%, 66.1%, 74.6% and 79.3%, respectively at $K = 200$ with an average reduction of nearly 4% compared to the case at $K = 150$.

From Figs. 4 to 11, it is clear that the best overall classification performance was obtained from the subspace discriminant ensemble model compared to the bagged trees ensemble, the default linear SVM and the KNN algorithms at $K = 200$. Figs. 12 and 13 compares the classification performance metrics of these classification models at $K = 150$ and $K = 200$, respectively.

Fig. 12 establishes the highest classification performance with $K = 150$ was obtained using the ensembles subspace discriminant, which achieved 96.9% accuracy, 98.3% sensitivity, 95.5% precision, 95.6% specificity, and 96.8% F-measure. In the second rank, the ensembles bagged trees, which achieved 96.9% accuracy, 96.6% sensitivity, 97.1% precision, 97.1% specificity, and 96.9% F-measure. While in the third rank, the linear SVM achieved 96.3% accuracy, 98.8% sensitivity, 94.5% specificity, 94.4% precision, and 96.5% F-measure.

Fig. 13 establishes the highest classification performance with $K = 200$ was obtained using the ensembles subspace discriminant, which achieved 98.6% accuracy, 99.4% sensitivity, 97.7% precision, 97.7% specificity, and 98.6% F-measure. In the second rank, the linear SVM, which achieved 97.7% accuracy, 98.8% sensitivity, 96.6% precision, 96.6% specificity, and 97.7% F-measure. However, in the third rank, the ensembles bagged trees achieved 97.5% accuracy, 98.3% sensitivity, 96.6% specificity, 96.6% precision, and 97.4% F-measure. Table 1 reports the area under the receiver operating characteristics curve (ROC), which known as the AUC for the different classifiers at both $K = 150$ and $K = 200$.

Accordingly, the proposed BoF model for diagnosing COVID-19 based on CXR images establishes its best performance using the ensembles subspace discriminant model with 200 visual words (i.e. $K = 200$).

Table 1
Area under receiver operating characteristics curve for different classification models using $K = 150$ and.

Classification Model	AUC using $K = 150$	AUC using $K = 200$
Ensembles Subspace Discriminant	**1.00**	**1.00**
Ensembles Bagged Trees	**1.00**	**0.99**
Linear SVM	1.00	1.00
Cosine KNN	0.99	0.99
Fine KNN	0.92	0.91
Medium KNN	0.99	0.99
Weighted KNN	0.99	0.99
Cubic KNN	0.99	0.99

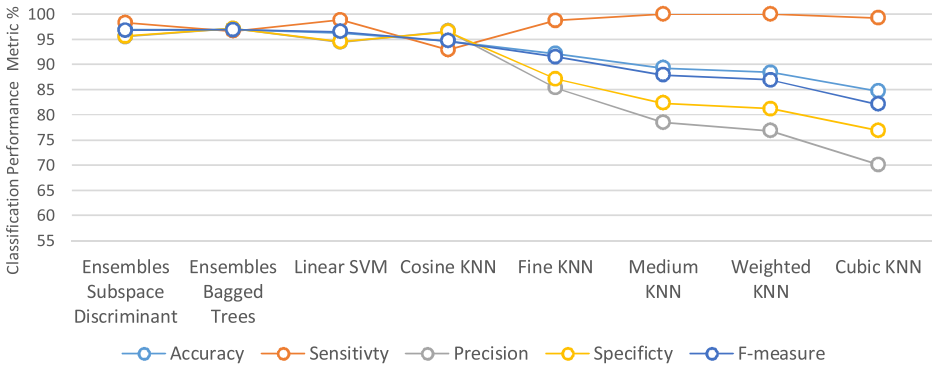

Fig. 12. Classification performance metrics of different classification models for the proposed BoF at $K = 150$.

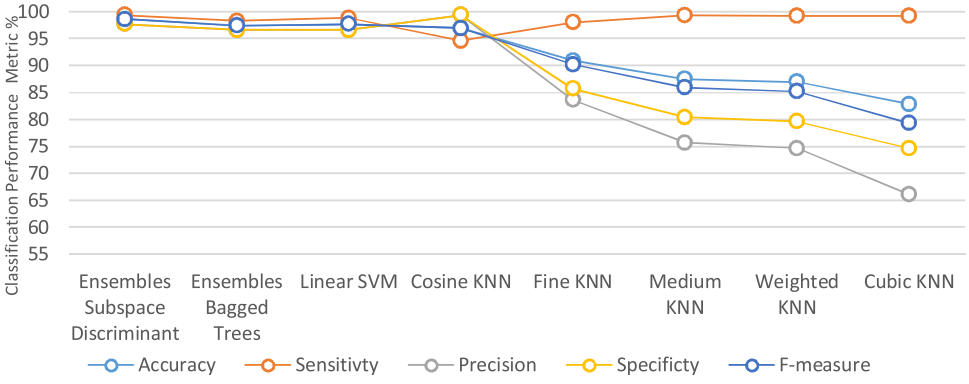

Fig. 13. Classification performance metrics of different classification models for the proposed BoF at $K = 200$.

Table 2
Performance metrics comparative study against the state-of-the-art studies.

Reference	Model	Dataset	Accuracy	Sensitivity	Precision	Specificity	F-measure
Proposed model	Ensembles subspace discriminant-based BoF model with $K = 200$	400 CXR images, including 200 COVID-19 cases and 200 normal cases (publicly available)	98.6%	99.4%	97.7%	97.7%	98.6%
[6]	COVIDX-NET VGG19 CNN model	50 CXR images including 25 positive COVID-19 cases	–	–	–	–	89%
	DenseNet CNN model		–	–	–	–	91%
[8]	MobileNet v2	224 CXR images of COVID-19 cases, 504 of normal cases, and 714 of viral and bacterial pneumonia cases (publicly available)	Two class classification: 96.87% Three class classification: 94.72%	98.66%	–	96.46%	–
[11]	SVM ensembles model	51 CXR images including 39 positive COVID-19 cases and 12 negative cases (MERS, SARS, and ARDS viral pneumonia)	98.04%	100%	–	91.67%	–

5. Performance evaluation of the proposed system against other studies

Table 2 demonstrates a comparative analysis that highlights the proposed system performance in contradiction of other state-of-the-art studies. The reported studies used CXR images including confirmed COVID-19 cases and negative COVID-19 cases. It is worth noting that the continuous update of the publicly available COVID-19 datasets makes it difficult to compare to proposed study to other studies which applied the exact same dataset. However, Table 2 indicates the superiority of the proposed model against the deep learning techniques in [6,8] and the SVM ensembles model in [11].

6. Conclusion

The occurrence of COVID-19 pandemic has imposed major pressure on healthcare facilities, which hinders providing efficient healthcare services without the risk of infections. Computer-aided diagnostic systems present an automated risk-free solution to diagnose COVID-19 using CXR images. Although several automated detection systems were proposed in literature, most of these systems relied on deep-learning techniques, which require large datasets for accurate performance. However, this condition was hardly achieved in several studies due to availability limitations. Accordingly, in our study, we have investigated the BoF classification model, which is one of the most promising traditional ML models. In our proposed model, the effect of the used number of visual words on the classification performance was studied using $K = 150$ and $K = 200$.Accordingly, it was concluded that the increase in the number of visual words boosts the classification accuracy due to the presence of more distinctive features, while reducing the computational time.

In the proposed BoF model, ensembles were employed in the classification process for the efficient classification of CXR images into normal or COVID-19 cases. Two several ensembles were investigated namely, the ensemble subspace discriminant and the ensemble bagged trees, which were compared to other classifiers including the linear SVM, which represent the default classifier in the bag-of-features, and different KNN classifiers. Results have indicated the superiority of the ensemble subspace discriminant at $K = 200$, which achieved 98.6% accuracy, 99.4% sensitivity, and 97.7% precision.

Funding acquisition

All Authors are members in the ASRT, Egypt funding project: ID: 7157.

CRediT authorship contribution statement

Amira S. Ashour: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - review & editing. **Merihan M. Eissa:** Software, Methodology, Formal analysis. **Maram A. Wahba:** Formal analysis, Visualization, Writing - original draft. **Radwa A. Elsawy:** Resources, Formal analysis, Visualization, Investigation. **Hamada Fathy Elgnainy:** Resources, Formal analysis, Visualization, Investigation. **Mohamed Saeed Tolba:** Resources, Formal analysis, Visualization, Investigation. **Waleed S. Mohamed:** Formal analysis, Investigation, Writing - review & editing, Supervision.

Acknowledgement

This work is funded and supported by the national program for supporting society innovation “Ideation Fund “of the Academy of Scientific Research & Technology (ASRT), Egypt [Project ID: 7157]. The authors are thankful to ASRT for this funding.

Declaration of Competing Interest

The authors report no declarations of Interest.

References

[1] N. Petrosillo, G. Viceconte, O. Ergonul, G. Ippolito, E. Petersen, COVID-19, SARS and MERS: are they closely related? Clin. Microbiol. Infect. (2020).
[2] C. Li, C. Zhao, J. Bao, B. Tang, Y. Wang, B. Gu, Laboratory diagnosis of coronavirus disease-2019 (COVID-19), Clin. Chim. Acta (2020).
[3] T. Goel, R. Murugan, S. Mirjalili, D.K. Chakrabartty, OptCoNet: an optimized convolutional neural network for an automatic diagnosis of COVID-19, Appl. Intell. (2020) 1–16.
[4] A. Narin, C. Kaya, Z. Pamuk, Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks, arXiv preprint arXiv 2003 (2020) 10849.
[5] A.I. Khan, J.L. Shah, M.M. Bhat, Coronet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images, Comput. Methods Programs Biomed. (2020) 105581.
[6] E.E.-D. Hemdan, M.A. Shouman, M.E. Karar, Covidx-net: a framework of deep learning classifiers to diagnose covid-19 in x-ray images, arXiv preprint arXiv 2003 (2020) 11055.
[7] A. Abbas, M.M. Abdelsamea, M.M. Gaber, Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network, arXiv preprint arXiv 2003 (2020) 13815.
[8] I.D. Apostolopoulos, T.A. Mpesiana, Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks, Phys. Eng. Sci. Med. (2020) 1.
[9] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K.N. Plataniotis, A. Mohammadi, Covid-caps: a capsule network-based framework for identification of covid-19 cases from x-ray images, arXiv preprint arXiv 2004 (2020) 02696.

[10] A.E. Hassanien, L.N. Mahdy, K.A. Ezzat, H.H. Elmousalami, H.A. Ella, Automatic x-ray covid-19 lung image classification system based on multi-level thresholding and support vector machine, medRxiv (2020).
[11] S. Mohammed, F. Alkinani, Y. Hassan, Automatic computer aided diagnostic for COVID-19 based on chest X-Ray image and particle swarm intelligence, Int. J. Intell. Eng. Syst. 13 (2020) 63–73.
[12] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. 110 (2008) 346–359.
[13] J. Ye, R. Janardan, Q. Li, Two-Dimensional Linear Discriminant Analysis,‖ in Advances in Neural Information Processing Systems, vol. 17, MIT Press, Cambridge, MA, 2005.
[14] J.P. Cohen, P. Morrison, L. Dao, K. Roth, T.Q. Duong, M. Ghassemi, Covid-19 image data collection: prospective predictions are the future, arXiv preprint arXiv 2006 (2020) 11988.