Health Information Science and Systems

RESEARCH

New bag of deep visual words based features to classify chest x-ray images for COVID-19 diagnosis

Chiranjibi Sitaula* and Sunil Aryal

Abstract

Purpose: Because the infection by Severe Acute Respiratory Syndrome Coronavirus 2 (COVID-19) causes the Pneumonia-like effect in the lung, the examination of Chest X-Rays (CXR) can help diagnose the disease. For automatic analysis of images, they are represented in machines by a set of semantic features. Deep Learning (DL) models are widely used to extract features from images. General deep features extracted from intermediate layers may not be appropriate to represent CXR images as they have a few semantic regions. Though the Bag of Visual Words (BoVW)-based features are shown to be more appropriate for different types of images, existing BoVW features may not capture enough information to differentiate COVID-19 infection from other Pneumonia-related infections.

Methods: In this paper, we propose a new BoVW method over deep features, called Bag of Deep Visual Words (BoDVW), by removing the feature map normalization step and adding the deep features normalization step on the raw feature maps. This helps to preserve the semantics of each feature map that may have important clues to differentiate COVID-19 from Pneumonia.

Results: We evaluate the effectiveness of our proposed BoDVW features in CXR image classification using Support Vector Machine (SVM) to diagnose COVID-19. Our results on four publicly available COVID-19 CXR image datasets (D1, D2, D3, and D4) reveal that our features produce stable and prominent classification accuracy (82.00% on D1, 87.86% on D2, 87.92% on D3, and 83.22% on D4), particularly differentiating COVID-19 infection from other Pneumonia.

Conclusion: Our method could be a very useful tool for the quick diagnosis of COVID-19 patients on a large scale.

Keywords: Bag of visual words (BoVW), Bag of deep visual words (BoDVW), Chest X-ray, COVID-19, Deep features, SARS-CoV-2

Introduction

The disease caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [28, 33, 48], commonly known as COVID-19, was originated in Wuhan city of China in late 2019 [50]. It is believed to be originated from bats [29, 39]. The virus has been transmitting from human to human all around the world [3, 11, 17]. It has spread over 200 countries in the world at present and become a pandemic that has killed 2,184,120 people[1] and 909 people in Australia alone[2] so far. While analyzing the effect of the SARS-CoV-2 virus in human body, it has been known that it causes the Pneumonia-like effect in the lungs. Thus, the study of Chest X-Ray (CXR) images could be an alternative to a swab test for early quick diagnosis of COVID-19. An automated CXR image analysis tool can be very useful to health practitioners for mass screening of people quickly. Also, recent studies show that AI-enabled techniques produce higher performance (e.g., precision, specificity, and sensitivity) than rapid tests [12].

For automatic analysis of images using algorithms, they are represented in machines by a set of semantic features, which are either traditional vision-based features [21] or deep learning-based features [22]. Large artificial neural networks, also known as Deep Learning (DL) models, are widely used to extract features from images and shown

*Correspondence: c.sitaula@deakin.edu.au
School of Information Technology, Deakin University, 75 Pigdons Rd, Waurn Ponds, Melbourne, VIC 3216, Australia

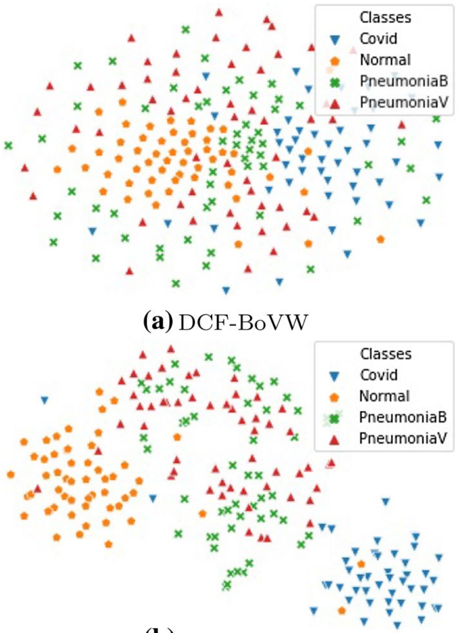
[1] https://www.worldometers.info/coronavirus/ (accessed date: 28/01/2021).

[2] https://www.health.gov.au/news/health-alerts/novel-coronavirus-2019-ncov-health-alert/coronavirus-covid-19-current-situation-and-case-numbers. (accessed date: 28/01/2021).

to work well in various types of images [14, 38, 51, 53–55]. Several studies have used DL models to analyze CXR images for coronavirus diagnosis, too. For instance, two recent works [34, 38] include the fine-tuning approach of transfer-learning on pre-trained DL models such as AlexNet [26], ResNet-18 [15], GoogleNet [57], etc. These methods normally require a massive amount of data to learn the separable features in addition to extensive hyper-parameter tuning tasks. However, most of the biomedical images (e.g., COVID-19 CXR images) are normally limited because of privacy issues. Thus, working with limited amount of data is always a challenging problem in DL models. Similarly, unlike other types of images, existing feature extraction methods such as GAP (Global Average Pooling) features achieved from pre-trained models may not provide accurate representation for CXR images because of their sparsity (i.e., having fewer semantic regions in them). Also, CXR images of lungs infected by COVID-19 and other Pneumonia look similar (i.e., there is a high degree of inter-class similarities). There might be subtle differences at very basic level, which, in our understanding, may be captured using the Bag of Words approach over deep features.

Bag of Visual Words (BoVW)-based features are shown to be more appropriate in images with the characteristics discussed above (sparsity and high inter-class similarity). They consider visual patterns/clues (known as visual words) in each image in the collection, thereby capturing sparse interesting regions in the image, which are useful in dealing with the inter-class similarity problem to some degree. BoVW-based feature extraction approach is popular not only in traditional computer vision-based methods such as Scale Invariant Features Transform (SIFT) [35] but also in DL-based methods due to its ability to capture semantic information extracted from the feature map of pre-trained DL models. The Bag of Deep Visual Words (BoDVW) features designed for one domain may not work well for another domain due to the varying nature of the images. For example, the Bag of Deep Convolutional Features (DCF-BoVW) [60] designed for satellite images may not work exactly for biomedical images such as CXR images. This is because of the fact that satellite image contains numerous semantic regions scattered in the image (dense) and thus DCF-BoVW could capture enough semantic regions of such images. However, the CXR images contain fewer semantic regions (sparse), which may not be captured accurately by DCF-BoVW.

In this paper, we propose a new BoDVW-based feature extraction method to represent CXR images. Our method eliminates some of the intermediate steps present in DCF-BoVW [60] and adds new steps because of the nature of CXR images. For this, we adopt the following steps. First, we extract the raw feature map from the mid-level (4th pooling layer) of the VGG16 pre-trained DL model [49] for each input image. We prefer the 4th pooling layer in our work, which has been chosen by empirical study and suggestion from the recent work by Sitaula et al. [52]. Next, we perform L2-normalization of each deep feature vector over the depth of the feature map. Using the training set, we design a codebook/dictionary over such deep features extracted from all the training images. Next, based on the codebook, we achieve our proposed features using a bag of visual words method for each input image. Last, such features based on the bag of visual words method is normalized by L2-norm, which results in the final representation of the input image. Because our final features are based on patterns extracted from mid-level features from training images, they capture more discriminating clues of sparse CXR images. The comparison of two-dimensional projections of features produced by DCF-BoVW and our proposed method on the COVID-19 image dataset [8] based on the t-SNE visualization [37] is shown in Fig. 1. It reveals that our features impart the higher separability between different classes.



(a) DCF-BoVW

(b) Our method

Fig. 1 Scatter plot of two dimensional projection of features produced by DCF-BoVW and our proposed method based on t-SNE visualization on chest x-ray images of Dataset 4 [8, 24]

The main **contributions** in our work are listed below:

(a) Propose to use the improved version of a bag of visual words method over deep features to work for the COVID-19 CXR image representation.
(b) Analyze the classification performance of our method across deep features extracted from five different pooling layers of the VGG16 model. Due to higher discriminability of deep features extracted from mid-level VGG16 model (see details in Sect. 4.4 and Sitaula et al. [52]), we leverage the fourth pooling layer (p_4) for feature extraction in our work. To design a codebook from deep features in our work, we use unsupervised clustering with the simple k-means algorithm.
(c) Evaluate our method on four datasets against the state-of-the-art methods based on pre-trained DL models in the COVID-19 CXR classification task using the Support Vector Machine (SVM) classifier. The results show that our method produces stable and state-of-the-art classification performance.

The remainder of the paper is organized as follows. In Sect. 2, we review some of the recent related works on CXR image representation and classification. Similarly, we discuss our proposed method in Sect. 3 in a stepwise manner. Furthermore, Sect. 4 details the experimental setup, performance comparison, and ablative study associated with it. Finally, Sect. 5 concludes our paper with potential directions for future research.

Related works

Deep Learning (DL) has been a breakthrough in image processing producing significant performance improvement in tasks such as classification, object detection, etc. A DL model is a large Artificial Neural Network (ANN), which has been designed based on the working paradigm of the human brain. If we design our DL model from scratch and train it, it is called a user-defined DL model. Similarly, if we use existing deep learning architectures pre-trained on large datasets, such as ImageNet [10] or Places [61], they are called pre-trained DL models. The features extracted from intermediate layers of DL models, either user-defined or pre-trained, provide rich semantic features to represent images that result in significantly better task-specific performance than traditional computer vision methods such as Scale Invariant Feature Transform (SIFT) [35], Generalized Search Tree (GIST)-color [41], Generalized Search Trees (GIST) [40], Histogram of Gradient (HOG) [9], Spatial Pyramid Matching (SPM) [30], etc.

Thus, in this section, we review some of the recent works in CXR image classification using DL models [2, 7, 20, 34, 36, 38, 42, 43, 47, 52, 56, 59]. We categorize them into two groups: 2.1 standalone deep learning algorithms and 2.2 ensemble learning algorithms

Standalone deep learning algorithms

At first, Stephen et al. [56] presented a new model for the detection of Pneumonia using DL and machine learning approach. They trained a Convolutional Neural Network (CNN) from scratch using a collection of CXR images. Their method produces the validation accuracy of 93.73% on such dataset. Islam et al. [20] devised a Compressed Sensing (CS)-based DL model for the automatic classification of CXR images for Pneumonia disease. Their method imparts 97.34% classification accuracy for the detection of Pneumonia. Similarly, Ayan et al. [2] used DL models on CXR images for early diagnosis of Pneumonia. They used Xception [5] and VGG16 [49] pre-trained models. Their results unveil that the VGG16 model outperforms the Xception model in terms of classification accuracy (87.00% versus 82.00%). This strengthens the efficacy of VGG16 model for CXR image representation and classification. Thus, the use of a pre-trained model became widespread in the representation and classification CXR images. For example, Varshni et al. [59] leveraged several pre-trained models such as VGG16 [49], Xception [5], ResNet50 [15], DenseNet121 [18], and DenseNet169 [18] individually as the features extractors and trained four classifiers separately using SVM [16], Random Forest [4], k-nearest neighbors [1], and Naïve Bayes [31] for the classification purpose. Among all those models used in their work, features extracted from DenseNet-169 model with SVM yields the highest area under curve (AUC) score of 80.02% in the classification. Furthermore, Loey et al. [34] used Generative Adversarial Networks (GAN) [13] and fine-tuned on AlexNet [26], ResNet18 [15], and GoogleNet [57] for the classification of the COVID-19 CXR images, where images belong to 4 categories (Covid, Normal, Pneumonia viral, and Pneumonia bacteria). For 2-class problem (Covid vs Normal) in their work, all of three methods (AlexNet, ResNet18, and GoogleNet) produce 100% classification accuracy. Similarly, for 3-class problem (Covid vs Normal vs Pneumonia bacteria) and 4-class problem (Covid vs Normal vs Pneumonia viral vs Pneumonia bacteria), AlexNet and GoogleNet produce the accuracies of 85.19% and 80.56%, respectively. In their method, GAN was exploited to augment the x-ray images to overcome the over-fitting problem during the training phase. Moreover, Khan et al. [25] devised a new deep learning model using the Xception [5] model, where they performed fine-tuning on CXR images. Their method imparts the overall classification accuracy of 89.60% in 4-class problem (Covid vs Pneumonia bacteria vs Pneumonia viral vs Normal), whereas it produces 95% accuracy in 3-class problem (Covid vs Pneumonia vs Normal).

Moreover, Ozturk et al. [42] established a new DL model for the categorization of COVID-19 related CXR images that uses DarkNet19 [45]. Their method provides the classification accuracy of 98.08% in 2-class problem (Covid vs No_Findings) and 87.02% in multi-class problem (Covid vs No_Findings vs Pneumonia). Furthermore, Luz et al. [36] devised another novel DL model, which uses the EfficientNet [58] model that adopts transfer learning over CXR images for the classification task. Their method yields the overall classification accuracy of 93.90%. Furthermore, Panwar et al. [43] established a new model, which is called nCOVnet, using the VGG16 model, which imparts a prominent accuracy for COVID-19 CXR image analysis. This further claims that the VGG16 model, which was quite popular in the past, is still popular in CXR image analysis. Their method imparts 97.62% true positive rate for the prediction of COVID-19 cases. Recently, Sitaula et al. [52] established an attention module on top of the VGG16 model (AVGG) for the CXR images classification. Their method outperforms several state-of-the-art methods. Their method produces the classification accuracy of 79.58% in 3-class problem (Covid vs No_findings vs Pneumonia), 85.43% in 4-class problem (Covid vs Normal vs Pneumonia bacteria vs Pneumonia viral), and 87.49% in 5-class problem (Covid vs No_findings vs Normal vs Pneumonia bacteria vs Pneumonia viral).

Ensemble learning algorithms

Ensemble learning methods have also been used in CXR image representation and classification, where different types of features are combined for better discrimination of images. Zhou et al. [62] proposed an ensemble learning approach of several ANNs for the lung cancer cell identification task. Their method provides the encouraging performance on several experimental sets in detecting the cancer cells compared to standalone models. For example, the ensemble model provides 17.3% rate of overall false identification, whereas standalone model provides 48.2% rate of overall false identification on the same experimental set. Sasaki et al. [47] established an ensemble learning approach using DL on CXR images. In their method, they performed several filtering and pre-processing operations on images and then ensembled them using DL for the detection of abnormality in CXR images. Their model yields the area under curve (AUC) value of 0.99 in the classification of CXR images. Li et al. [32] also utilized multiple CNNs (E-CNNs) to reduce the false positive results on lung nodules of CXR images. E-CNNs model attains the highest sensitivity of 94% on CXR images. Moreover, Islam et al. [19] designed an ensemble method to aggregate different pre-trained deep learning models for abnormality detection (Tuberculosis and Cardiomegaly) in lung images. Their model provides 90% classification accuracy in Tuberculosis detection and 93% accuracy in Cardiomegaly detection. Chouhan et al. [7] introduced a model, where the outputs of 5 pre-trained deep learning models, namely AlexNet, ResNet18, DenseNet121, GoogleNet, and Inception-V3, were ensembled for the detection of Pneumonia using transfer learning (TL) approach. This helps to learn multiple types of information achieved from various pre-trained DL models to bolster the classification performance. Their ensemble model provides the classification accuracy of 96.40% in CXR images classification, which is superior to the performance of standalone models. Nevertheless, ensemble learning algorithms are arduous for which we need to be vigilant in hyper-parameter tuning in addition to the over-fitting problem.

Most existing methods in the literature need a huge amount of data for fine-tuning DL models and most of them extract high-level features, which may not be sufficient for CXR images. They require mid-level features that are neither more generic nor more specific. In the next section, we introduce our proposed approach to extract such mid-level features.

Proposed method

The mid-level features of CXR images can be achieved from the feature maps extracted from the intermediate layers of pre-trained models using a Bag of Visual Words (BoVW) method. Since CXR images are sparse (having few semantic regions), existing bag of visual words methods that have been applied to represent other images (e.g., satellite images) may not work accurately in this domain. To this end, we propose an improved version of a bag of visual words method on deep features to represent CXR images more accurately. In this section, we discuss the steps involved in our proposed feature extraction method. There are three main steps in our method: deep features extraction (Sect. 3.1), unsupervised codebook (dictionary) design (Sect. 3.2), and proposed features extraction (Sect. 3.3). The overall pipeline of the proposed method is shown in Fig. 2.

Deep features extraction

At first, we extract the deep features from the feature map of the 4th pooling (p_4) layer from VGG16 [49], which is a deep learning model pre-trained on the ImageNet [10] dataset. We prefer VGG16 in our work because of three reasons. First, it has a unrivalled performance in recent biomedical image analysis works such as COVID-19 CXR image analysis [52], breast cancer image analysis [51], etc.

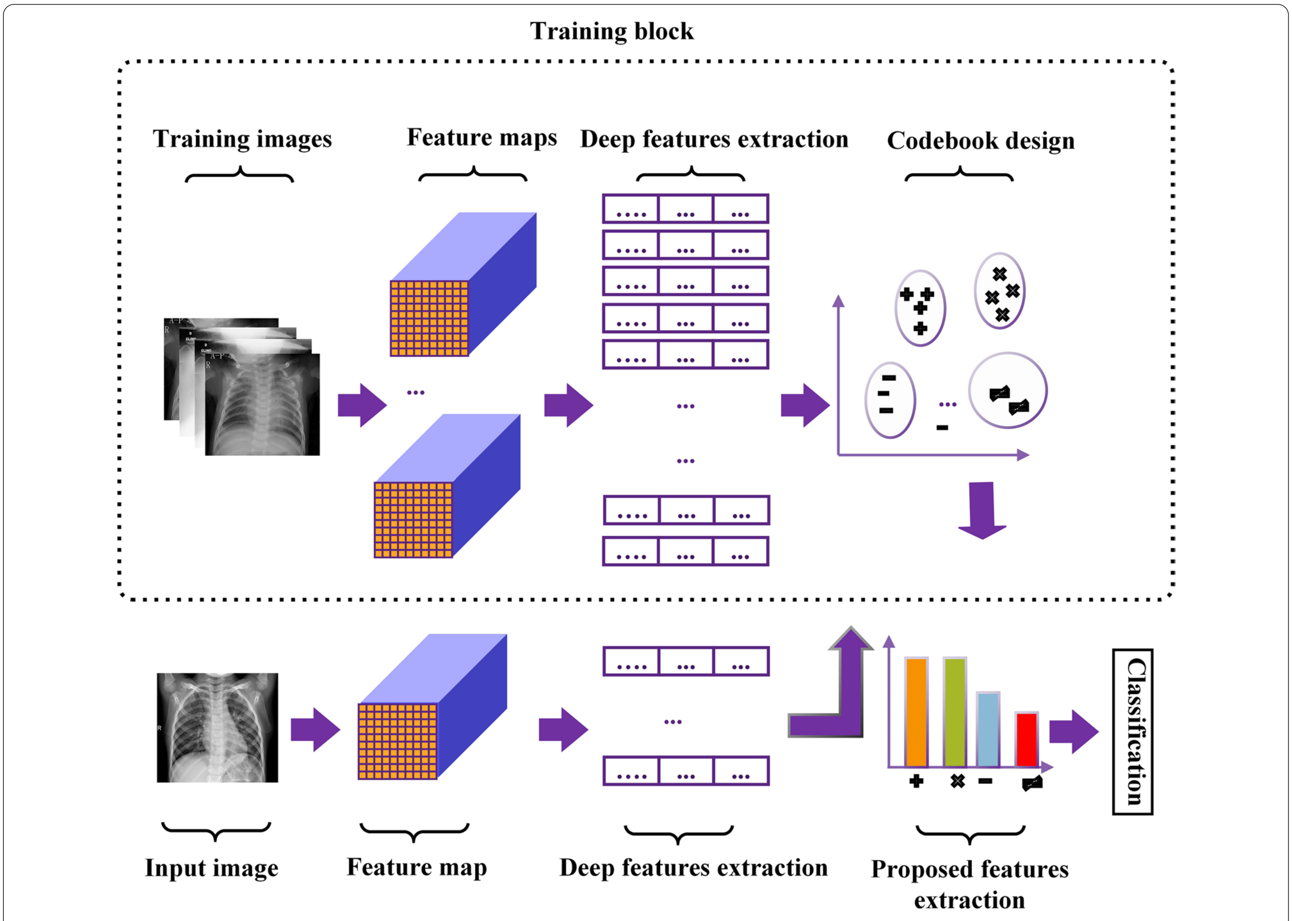


Fig. 2 The overall pipeline of the proposed method. Based on the codebook/dictionary achieved from training block, the proposed features vector is extracted for each input image using the bag of visual features approach

Second, it is easy to analyze and experiment with its five pooling layers. Third, it uses smaller-sized kernels, which could learn distinguishing features of biomedical images at a smaller level.

We believe that 4th layer of such a model has a higher level of discriminability than other layers as seen in Fig. 3. The detailed discussion about the efficacy of the 4th pooling layer is also presented in Sect. 4.4. Furthermore, we use the VGG16 model due to its simple and prominent features extraction capability in various types of image representation tasks [14, 27, 55]. Authors in [51, 52] highlighted the importance of 4th pooling layer compared to other layers in biomedical imaging for separable feature extraction. The size of the features map from the p_4 layer of the VGG16 model is 3-D shape having $H = 14$ (height), $W = 14$ width, and $L = 512$ (length). From each feature map, we achieve 14×14 features, each of size 512. Then, each feature vector is L2-normalized. This normalization helps to preserve the separability of deep features of images [14]. Let us say that an input image yields feature map with $14 \times 14 = 196$ number of features vectors that are represented by x_0, x_1, x_2,$\cdots$,x_{196}. Each features vector x_i is of 512-D size (i.e., $|x_i| = 512$), which is then normalized by L2-norm as seen in Eq. (1).

$$x_i' = \frac{x_i}{||x_i||_2 + \epsilon} \tag{1}$$

In Eq. (1), the features vector x_i' represents the ith normalized deep features vector extracted from the corresponding feature map. While achieving such feature vector, we add $\epsilon = 1e-08$ with denominator to avoid the divide by zero exception because the feature map obtained for chest x-ray images is sparse and it is more likely to encounter the divide by zero exception in most cases.

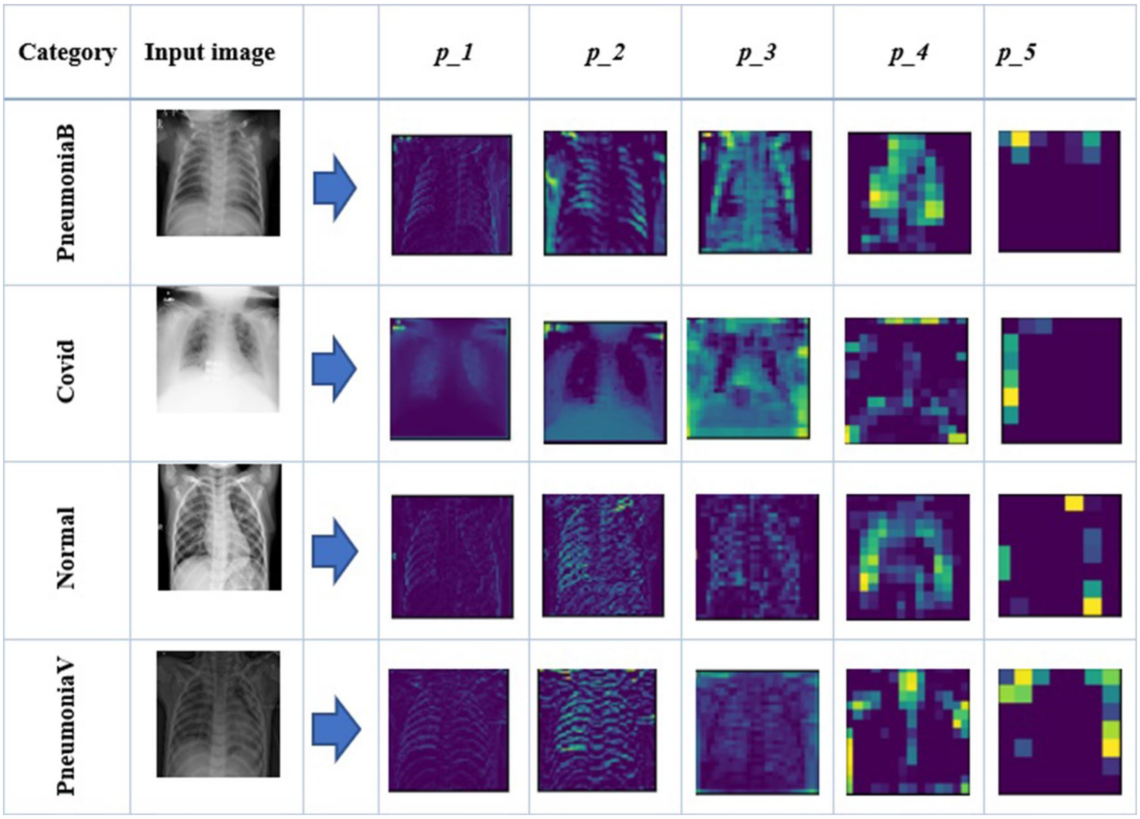


Fig. 3 Feature maps of an input image from each of the four categories in the COVID-19 dataset extracted from the five pooling layers of VGG16. *p_i* ($i = 1, 2, \cdots, 5$) represents the *i*th polling layer

Unsupervised dictionary (codebook) design

We use deep features (extracted from the VGG16 model as discussed above in Sect. 3.1) of all training images to design a dictionary or codebook. Each image provides $\{x_i'\}_{i=1}^{196}$ deep features and let's say there are m training images. Thus, the total number of deep features to design our codebook is $196 \times m$. To design the codebook or dictionary, we utilize a simple, yet popular unsupervised clustering algorithm called k-means [23] that groups deep features having similar patterns into clusters. Given a parameter k, k-means algorithm provides k groups or clusters ($\{c_1, c_2, \cdots, c_k\}$) of deep features, where deep features in each group are similar (i.e., they capture similar patterns of images). We use such k cluster centroids as a dictionary or codebook of deep visual words, which is used to extract features for each input image.

Proposed feature extraction

To extract features of each input image y, we first follow step 3.1 to achieve 196 normalized deep features of y and then, design a histogram based on the dictionary defined in step 3.2. The size of histogram is k (the dictionary size), where each code (cluster centroid) in the dictionary c_j has a weight w_j. All 196 deep features of y are assigned to their nearest centroids. The weight w_j is the number of deep features assigned to the cluster c_j. In other words, histogram is a bag of visual words (centroids), where weights are their frequencies. The resulting features of y is a k-D vector $\{w_1, w_2, \cdots, w_k\}$. The extracted bag of visual words features vector is, finally, normalized as in Eq. (1), which acts as our proposed features of the corresponding input image.

Difference between our BoVW and DCF-BoVW features

The main differences between our BoVW and DFC-BoVW features are explained in three different aspects.

Firstly, the L1-normalization used by the DCF-BoVW method is more suitable for dense images such as satellite images. However, since the chest x-ray images are sparse in nature, such normalization becomes counterproductive as it masks some discriminating clues. Thus, we eliminate this normalization in our method due to the nature of chest x-ray images.

Secondly, we apply L2-normalization to the deep features extracted from the unnormalized feature maps to exploit the property of cosine similarity in the k-means clustering. Note that Euclidean distance on the L2-normalized feature is equivalent to using cosine distance. The directions of deep features are more important than their lengths to group vectors with similar patterns into clusters to define our codebook. This will help us to detect sparse patterns in images, which can be useful in discriminating abnormalities in x-ray images.

Finally, we replace the L1-normalization of the final BoVW features used in the DCF-BoVW method by L2-normalization. Again, this allows us to exploit the property of cosine similarity in the SVM's RBF kernel. Because BoVW features are sparse as many vector entries are zeros, cosine similarity is more appropriate than the Euclidean distance.

Complexity analysis

In this subsection, we analyze the time complexity of our method. Since our proposed method is based on k-means clustering algorithm over training deep features, it takes $O(m \times k \times d \times i)$ time complexity, where m, k, d, and i denote number of input feature vectors, number of clusters, features size of each vector, and number of iterations, respectively. After that, each testing feature vector for each corresponding input image takes $O(l \times k)$ time complexity during the proposed feature extraction, where l denote the total number of deep features representing the testing image.

Experimental setup and comparison

Dataset

We utilize four COVID-19 CXR image datasets that are publicly available.

Dataset 1 [42] comprises of 3 categories: Covid-19, Pneumonia, and No_findings. Here, each category has at least 125 images. The No_findings category has several ambiguous and challenging CXR images.

Dataset 2 [25] comprises of 4 categories: Covid, Normal, Pneumonia viral (PneumoniaV), and Pneumonia bacteria (PneumoniaB).

Dataset 3 [25, 42] includes 5 categories: Covid, No_findings, Normal, Pneumonia bacteria (PneumoniaB), and Pneumonia viral (PneumoniaV). Dataset 3 is the combination of No_finding category from Dataset 1 and other categories from Dataset 2. Here, each category includes at least 320 CXR images.

Dataset 4 [8, 24] has 4 categories: Covid, Normal, PneumoniaV, and PneumoniaB, where each category contains at least 69 images. This dataset has been used by [34], which can be downloaded from the link[3]

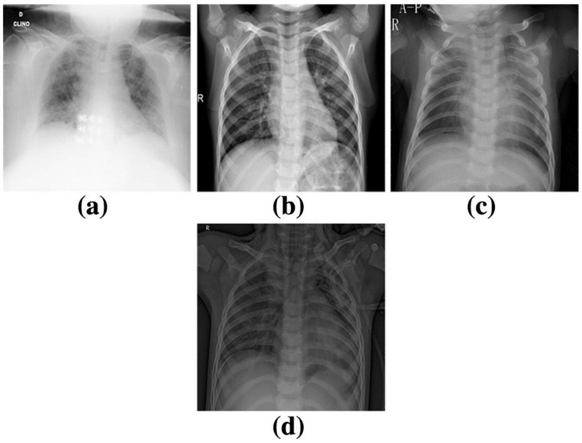


Fig. 4 Example images of chest x-ray images from Dataset 4 [8, 24] for four classes: **a** Covid, **b** Normal, **c** PneumoniaB, and **d** PneumoniaV

Example images of COVID-19 are shown in Fig. 4. Also, further detailed information of all datasets are provided in Table 1.

We divide the images of each dataset into 70:30 ratio for the train:test sets for each category (class). We compare the average accuracy of five different runs.

Implementation

To implement our work, we use Keras [6] implemented in Python [46]. Keras is used to implement the pre-trained model in our work. We use the number of clusters $k = 400$ in k-means clustering to define the dictionary to extract proposed features. For the classification purpose, we use the Support Vector Machine (SVM) classifier implemented in Scikit-learn [44]. We normalize and standardize our features to feed into the SVM classifier. Moreover, we fix the kernel as the radial basis function (*RBF*) kernel with the γ parameter as $1e-05$. We automatically tune the SVM cost parameter C in the range of $\{1, 10, 20, \cdots, 100\}$ on the training set using a 5-fold cross-validation method and use the optimal setting to train the model using the entire training set. We execute all our experiments on a workstation with NVIDIA Geforce GTX 1050 GPU and 4 GB RAM.

Comparison with state-of-the-art methods

We present the results of the experiments conducted to compare our method with five recent state-of-the-art methods (one method uses the BoW approach over deep features and four methods adopt transfer-learning approach) that are based on pre-trained models on four CXR image datasets (D1, D2, D3, and D4) in Table 2. In the table, the second, third, fourth, and fifth columns enlist the accuracies of contending methods on D1,

[3] COVID-19 Dataset Available online: https://drive.google.com/uc?id=1coM7x3378f-Ou2l6Pg2wldaOI7Dntu1a (accessed on Apr 17, 2020).

Table 1 Description of datasets used in our work

Dataset	# of images	Categories	Ref.
Dataset 1 (D1)	1,125	Covid-19, Pneumonia, and No_findings	[42]
Dataset 2 (D2)	1,638	Covid, Normal, PneumoniaB, and PneumoniaV	[25]
Dataset 3 (D3)	2,138	Covid, Normal, No_findings, PneumoniaB, and PneumoniaV	[25, 42]
Dataset 4 (D4)	320	Covid, Normal, PneumoniaB, and PneumoniaV	[8, 24]

Table 2 Comparison with previous methods on four datasets (D1, D2, D3, and D4) using average classification accuracy (%) over five runs

Method	D1 (%)	D2 (%)	D3 (%)	D4 (%)
DCF-BoVW, 2018 [60]	75.31	81.53	83.72	72.46
CoroNet, 2020 [25]	76.82	80.60	83.41	-
Luz et al., 2020 [36]	47.51	84.29	79.96	-
nCOVnet, 2020 [43]	62.95	70.62	67.67	-
AVGG, 2020 [52]	79.58	85.43	87.49	-
Ours	82.00	87.86	87.92	83.22

Note that '-' represents unavailable results because of the over-fitting problems in existing DL-based methods using transfer learning on D4

D2, D3, and D4, respectively. Note that the accuracies reported in the table are averaged accuracy of five runs for each method.

Results in the second column of Table 2 show that our method outperforms all five contenders with the accuracy of 82.00% on D1. This further highlights that it imparts the performance increment of 2.542% from the second-best method (AVGG [52]) and 35% from the worst method (Luz et al. [36]). Similarly, on D2 in the third column of Table 2, we notice that our method outperforms all five methods with an accuracy of 87.86%, which is 2.43% higher than the second-best method (AVGG [52]) and 17% higher than the worst-performing method (nCOVnet [43]). In the fourth column of Table 2 on D3, we observe that our method, which yields 87.92% accuracy, is superior to the second-best method (AVGG [52]) with a slim margin of 0.43%, whereas it imparts over 20% accuracy against the worst performing method (nCOVnet [43]). Last but not the least, in the fifth column of Table 2 on D4, we notice that our method, which produces 83.22%, outperforms the DCF-BoVW [60] with the margin of over 10% accuracy. Please note that for D4, we only compare our method with DCF-BoVW [60], which can work for a limited amount of data only, and do not compare with other DL-based methods that use transfer learning because this dataset has a very limited number of CXR images.

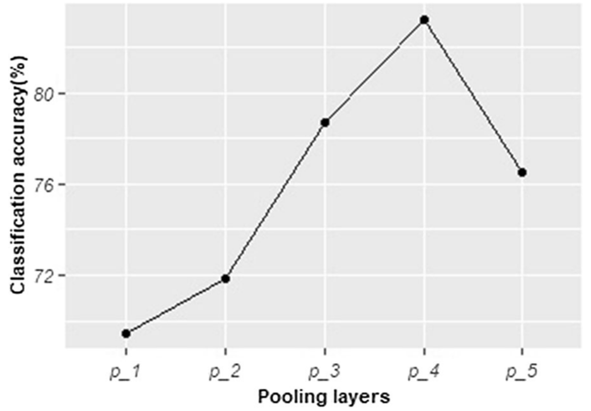


Fig. 5 Average classification accuracy (%) achieved by our method on D4 using deep features extracted from the five pooling layers (*p*_1 to *p*_5) of the VGG16 model

The comparison of our method against five different recent DL-based methods on four datasets unveils that our method provides a stable and prominent performance. This result further underscores that the classification performance of the bag of words approach, which captures the more detailed spatial information of deteriorated regions more accurately than other methods, seems more appropriate to CXR image analysis (e.g., COVID-19 CXR images) than other DL-based methods using transfer learning approach.

Ablative study of pooling layers

In this subsection, we present the results of an ablative study on D4, which is the smallest dataset, to analyze the effect on the classification accuracy of using deep features from the five different pooling layers of VGG16 in our method. The detailed results are presented in Fig. 5. While observing the line graph, we notice that the 4th pooling layer of the VGG16 model produces highly separable features than other pooling layers on the COVID-19 dataset. Furthermore, the lower pooling layers (p_1, p_2, and p_3), which provide the the generic low-level information of the image, and higher pooling layer (p_5), which provides the specific high-level information of the image, are not appropriate to CXR images. This could be because

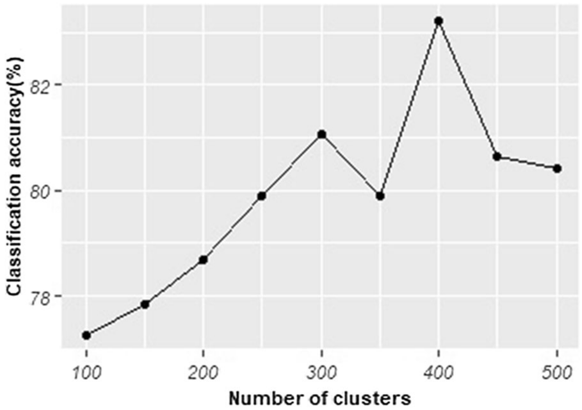


Fig. 6 Average classification accuracy (%) with different cluster number on D4. Note that deep features from the 4th pooling layer (p_4) were used

of nature of CXR image, which neither prefer generic nor specific information for better class separability.

Ablative study of cluster numbers

We analyze different number of unsupervised patterns to be used in our experiments on D4. For this, we vary the cluster numbers from 100 to 500 using the interval of 50 and present the results in Fig. 6. From the line graph, we notice that the appropriate number of clusters that produce the best result is $k = 400$. Thus, we believe that both lower and higher number of clusters than 400 are not useful to discriminate CXR image because lower cluster numbers may not cover complete discriminating patterns to represent the CXR images and higher cluster numbers may repeat the discriminating patterns.

Ablative study of class-wise performance

We study the average class-wise performance of our method on D4. The average class-wise performance are reported using precision, recall, and f1-score, which are defined in Eqs. (2),(3), and (4), respectively.

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

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

$$\text{F1-score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}, \tag{4}$$

where *TP*, *FP*, and *FN* represent true positive, false positive, and false negative results, respectively. We present the average precision, recall, and f1-score in Table 3. The results show the discriminability of our proposed method in all four classes. It shows that our method can distinguish the Covid and normal class well and there is some confusion among two Pneumonia classes.

Table 3 Average class-wise study (%) over five runs of our method on D4 using precision, recall, and f1-score

Class	Precision (%)	Recall (%)	F1-score (%)
Covid	100.00	97.20	98.40
Normal	94.20	93.60	93.80
PneumoniaB	75.80	67.60	71.00
PneumoniaV	68.00	76.80	71.80

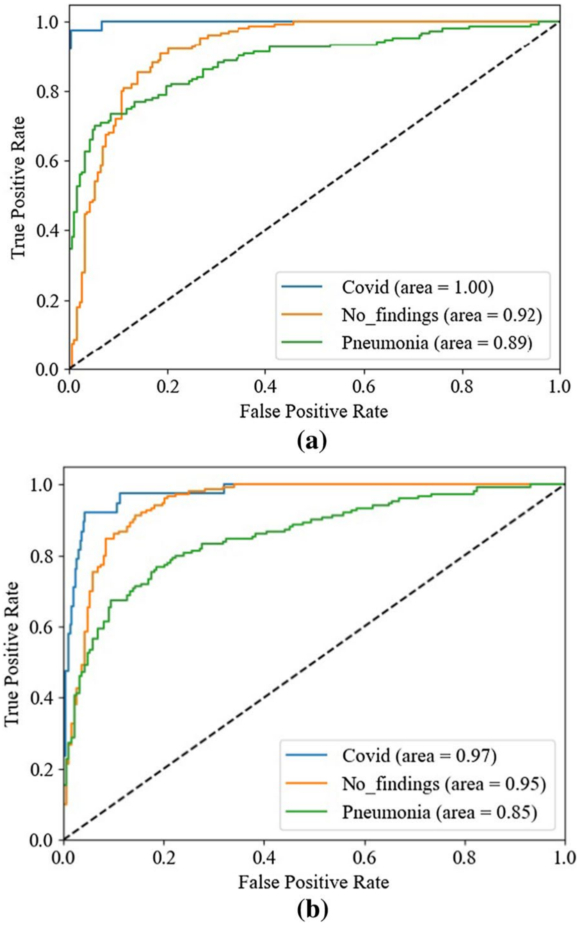


Fig. 7 ROC plot of our method (**a**) and recent method, AVGG [52] (**b**)

We also compare our method with one recent method for the class-wise analysis using Receiver Operating Characteristic (ROC) curve, which plots the graph based on true positive rate and false positive rate. As an example for class-wise analysis, we utilize third train/test split

Table 4 Analysis of different train/test splits using classification accuracy (%) on D1

Train/test	Accuracy (%)
50/50	80.24
60/40	81.55
70/30	85.79
80/20	81.77
90/10	83.18

Note that the results are based on the randomly designed single train/test set on D1 for each of five different ratios

(f3) of D1. The ROC curves of two methods on this set are shown in Fig. 7. While looking at both ROC plots, we observe that our method produces superior performance in terms of area for two classes (Covid and Pneumonia) to one of the recent methods (AVGG [52]) and comparable performance for No_findings class.

Analysis of different train/test splits

In this subsection, we study the classification performance of different train/test ratios. For this study, we utilize five different random train/test splits on D1 using five different ratios (50/50, 60/40, 70/30, 80/20, and 90/10). The results are listed in Table 4. While observing the table, we notice that each ratio has its own importance in the classification; however, our chest x-ray classification performs better on train/test split of 70/30 ratio. This helps to maintain the number of training and testing images sufficiently for the classification. Thus, we use this split ratio throughout our work.

Analysis of hyper-parameters

In this subsection, we study the effect of different hyper-parameters used in our work. For such study, we choose one split (e.g., third set) of D1 and analyze the effects of two main hyper-parameters, C and Gamma (γ), used in SVM during classification. The sample results are listed in Table 5. While observing the table, we notice that the best C and Gamma values of the current set for higher classification accuracy (%) (80.20 ± 0.03) are 40 and 1e-05, respectively. Note that we perform such operation for each split of each dataset and select C values automatically keeping Gamma fixed to 1e-05. This results in variation of C values from one split to another during classification for each dataset used in our work.

Table 5 Sample analysis of hyper-parameters used with RBF kernel in SVM based on classification accuracy (%) ± standard deviation in our work

	Gamma				
C	1e-01	1e-02	1e-03	1e-04	1e-05
1	44.60 ± 0.01	71.70 ± 0.04	79.50 ± 0.03	69.60 ± 0.08	68.40 ± 0.07
10	44.60 ± 0.01	71.40 ± 0.04	78.40 ± 0.04	79.50 ± 0.05	69.60 ± 0.07
20	44.60 ± 0.01	71.40 ± 0.04	78.40 ± 0.04	79.40 ± 0.05	75.40 ± 0.05
30	44.60 ± 0.01	71.40 ± 0.04	78.40 ± 0.04	78.00 ± 0.08	78.90 ± 0.04
40	44.60 ± 0.01	71.40 ± 0.04	78.40 ± 0.04	76.90 ± 0.07	**80.20 ± 0.03**
50	44.60 ± 0.01	71.40 ± 0.04	78.40 ± 0.05	76.40 ± 0.08	80.10 ± 0.04
60	44.60 ± 0.01	71.40 ± 0.04	78.40 ± 0.04	75.70 ± 0.08	80.20 ± 0.05
70	44.60 ± 0.01	71.40 ± 0.04	78.40 ± 0.04	75.60 ± 0.07	80.10 ± 0.05
80	44.60 ± 0.01	71.40 ± 0.04	78.40 ± 0.04	75.40 ± 0.07	79.50 ± 0.07
90	44.60 ± 0.01	71.40 ± 0.04	78.40 ± 0.04	75.20 ± 0.06	79.40 ± 0.05
100	44.60 ± 0.01	71.40 ± 0.04	78.40 ± 0.04	75.10 ± 0.06	79.50 ± 0.05

Note that we perform 5-fold cross validation over corresponding training set to choose the best hyper-parameters for the SVM classification

Conclusion and future works

In this paper, we propose a new feature extraction method based on Bag of Deep Visual Words (BoDVW) to represent chest x-ray images. Empirical results in the classification of chest x-ray images using the COVID-19 dataset show that our method is more appropriate to represent chest x-ray images. This is mainly because our features can capture a few interesting regions (sparse markers) indicating abnormalities well. Our features are extracted using a visual dictionary defined by the clustering of deep features from all training images. Therefore, they can capture patterns in each training image and thus help to capture potential markers for various lung infections such as COVID-19 and Pneumonia. Also, the size of our proposed features is relatively very small compared to other existing methods and our method runs faster than other existing methods.

Though the evaluation is done on a relatively small dataset, our method shows promising results to detect and distinguish lung infection due to Pneumonia and COVID-19. COVID-19 is a relatively new disease and there are not a lot of chest x-ray images available. Nevertheless, given the current crisis with the COVID-19 pandemic, our method, which is accurate and fast, can be very useful for health professionals for mass screening of people for COVID-19. Accurate detection and distinction of lung infections due to COVID-19 and Pneumonia are very important for COVID-19 diagnosis as people infected by these diseases show similar symptoms.

In the future, it would be interesting to verify our results in a large study with more sample images including other types of lung infection such as Tuberculosis. Another potential direction is to investigate if a similar approach can be used to represent other types of medical images such as CT scans, histopathological images, colonoscopy images, etc.

Acknowledgements
Authors would like to thank anonymous reviewers for their constructive comments to improve the quality of the paper. Sunil Aryal is supported by the joint grant by US Air Force Office of Scientific Research and Office of Naval Research under award number FA2386-20-1-4005.

Received: 29 January 2021 Accepted: 20 April 2021
Published online: 18 June 2021

References

1. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat. 1992;46(3):175–85.
2. Ayan E, Ünver HM. Diagnosis of pneumonia from chest x-ray images using deep learning. In: In Proc. Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), pp 1–5 (2019)
3. Bastola A, Sah R, Rodriguez-Morales AJ, Lal BK, Jha R, Ojha HC, Shrestha B, Chu DK, Poon LL, Costello A, et al. The first 2019 novel coronavirus case in Nepal. Lancet Infect Dis. 2020;20(3):279–80.
4. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
5. Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 1251–1258 (2017)
6. Chollet F, et al. Keras. https://github.com/fchollet/keras (2015)
7. Chouhan V, Singh SK, Khamparia A, Gupta D, Tiwari P, Moreira C, Damaševičius R, de Albuquerque VHC. A novel transfer learning based approach for pneumonia detection in chest x-ray images. Appl Sci. 2020;10(2):559.
8. Cohen JP, Morrison P, Dao L. Covid-19 image data collection. arXiv preprint arXiv:200311597 (2020)
9. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 886–893 (2005)
10. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2009)
11. Giovanetti M, Benvenuto D, Angeletti S, Ciccozzi M. The first two cases of 2019-ncov in italy: Where they come from? J Med Virol 2020. https://doi.org/10.1002/jmv.25699
12. Gomes J, Barbosa V, Santana M, Bandeira J, Valença M, de Souza R, Ismael A, dos Santos W. Ikonos: An intelligent tool to support diagnosis of covid-19 by texture analysis of x-ray images. Research on Biomedical Engineering pp 1–14 (2020)
13. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. Advances in Neural Information Processing Systems, pp 2672–2680 (2014)
14. Guo Y, Liu Y, Lao S, Bakker EM, Bai L, Lew MS. Bag of surrogate parts feature for visual recognition. IEEE Trans Multimedia. 2018;20(6):1525–36.
15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp 770–778 (2016)
16. Hearst MA. Support vector machines. IEEE Intell Syst. 1998;13(4):18–28.
17. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, Spitters C, Ericson K, Wilkerson S, Tural A, et al. First case of 2019 novel coronavirus in the United States. N Engl J Med. 2020;382(10):929–36.
18. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 4700–4708 (2017)
19. Islam MT, Aowal MA, Minhaz AT, Ashraf K. Abnormality detection and localization in chest x-rays using deep convolutional neural networks. arXiv preprint arXiv:170509850 (2017)
20. Islam SR, Maity SP, Ray AK, Mandal M. Automatic detection of pneumonia on compressed sensing images using deep learning. In: In Proc. Canadian Conference of Electrical and Computer Engineering (CCECE), pp 1–4 (2019)
21. Ismael A, Şengür A. The investigation of multiresolution approaches for chest x-ray image based covid-19 detection. Health Inf Sci Syst. 2020;8(1):1–11.
22. Ismael A, Şengür A. Deep learning approaches for covid-19 detection based on chest x-ray images. Exp Syst Appl. 2021;164:114054.
23. Jin X, Han J. K-Means Clustering. Boston, MA: Springer; 2010. p. 563–4.
24. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172(5):1122–31.
25. Khan A, Shah J, Bhat M. Coronet: a deep neural network for detection and diagnosis of covid-19 from chest x-ray images. Comput Methods Programs Biomed. 2020;196:105581.
26. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proc. Adv. Neural Inf. Process. Syst. (NIPS), pp 1097–1105 (2012)
27. Kumar A, Singh SK, Saxena S, Lakshmanan K, Sangaiah AK, Chauhan H, Shrivastava S, Singh RK. Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. Inf Sci. 2020;508:405–21.
28. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (sars-cov-2) and corona virus disease-2019 (covid-19): the epidemic and the challenges. Int J Antimicrobial Agents. 2020;55(3):105924.
29. Latinne A, Hu B, Olival KJ, Zhu G, Zhang L, Li H, Chmura AA, Field HE, Zambrana-Torrelio C, Epstein JH, Li B, Zhang W, Wang LF, Shi ZL, Daszak P. Origin and cross-species transmission of bat coronaviruses in china. Nat Commun. https://doi.org/10.1038/s41467-020-17687-3 (2020)
30. Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp 2169–2178 (2006)
31. Lewis DD. Naive (bayes) at forty: The independence assumption in information retrieval. In: Proc. European Conference on Machine Learning, pp 4–15 (1998)
32. Li C, Zhu G, Wu X, Wang Y. False-positive reduction on lung nodules detection in chest radiographs by ensemble of convolutional neural networks. IEEE Access. 2018;6:16060–7.
33. Li J, Li JJ, Xie X, Cai X, Huang J, Tian X, Zhu H. Game consumption and the 2019 novel coronavirus. Lancet Infect Dis. 2020;20(3):275–6.
34. Loey M, Smarandache F, Khalifa M. Within the lack of chest covid-19 x-ray dataset: a novel detection model based on gan and deep transfer learning. Symmetry. 2020;12(4):651.
35. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis. 2004;60(2):91–110.
36. Luz E, Silva PL, Silva R, Moreira G. Towards an efficient deep learning model for covid-19 patterns detection in x-ray images. arXiv preprint arXiv:200405717 (2020)
37. Maaten L, Hinton G. Visualizing data using t-sne. J Mach Learn Res. 2008;9:2579–605.
38. Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. arXiv preprint arXiv:200310849 (2020)
39. Nguyen TT, Abdelrazek M, Nguyen DT, Aryal S, Nguyen DT, Khatami A. Origin of novel coronavirus (covid-19): A computational biology study using artificial intelligence. bioRxiv https://doi.org/10.1101/2020.05.12.091397 (2020)
40. Oliva A. Gist of the scene. In: Oliva A, Schyns PG, editors. Neurobiology of Attention. Burlington, MA: Elsevier; 2005. p. 251–6.

41. Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis. 2001;42(3):145–75.
42. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR. Automated detection of covid-19 cases using deep neural networks with x-ray images. Comput Biol Med. 2020;121:103792.
43. Panwar H, Gupta P, Siddiqui MK, Morales-Menendez R, Singh V. Application of deep learning for fast detection of covid-19 in x-rays using ncovnet. Chaos Solitons Fractals. 2020;138:109944.
44. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
45. Redmon J, Farhadi A. Yolo9000: better, faster, stronger. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 7263–7271 (2017)
46. Rossum G. Python reference manual. Amsterdam, The Netherlands: Tech. rep (1995).
47. Sasaki T, Kinoshita K, Kishida S, Hirata Y, Yamada S. Ensemble learning in systems of neural networks for detection of abnormal shadows from x-ray images of lungs. J Signal Proces. 2012;16(4):343–6.
48. Sharfstein JM, Becker SJ, Mello MM. Diagnostic testing for the novel coronavirus. Jama. 2020;323(15):1437–8.
49. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 (2014)
50. Singhal T. A review of coronavirus disease-2019 (covid-19). Indian J Pediatr. 2020;87(4):281–6.
51. Sitaula C, Aryal S. Fusion of whole and part features for the classification of histopathological image of breast tissue. Health Inf Sci Syst. 2020;8(1):1–12.
52. Sitaula C, Hossain M. Attention-based vgg-16 model for covid-19 chest x-ray image classification. Appl Intell 2020. https://doi.org/10.1007/s10489-020-02055-x
53. Sitaula C, Aryal S, Xiang Y, Basnet A, Lu X (2020a) Content and context features for scene image representation. arXiv preprint arXiv:200603217
54. Sitaula C, Xiang Y, Aryal S, Lu X.Scene image representation by foreground, background and hybrid features. arXiv preprint arXiv:200603199 (2020b)
55. Sitaula C, Xiang Y, Basnet A, Aryal S, Lu X. Hdf: hybrid deep features for scene image representation. In: Proc. International Joint Conference on Neural Networks (IJCNN), pp 1–8 (2020c)
56. Stephen O, Sain M, Maduh UJ, Jeong DU. An efficient deep learning approach to pneumonia classification in healthcare. J Healthcare Eng 2019;2019:4180949 .
57. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp 1–9 (2015)
58. Tan M, Le QV. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:190511946 (2019)
59. Varshni D, Thakral K, Agarwal L, Nijhawan R, Mittal A. Pneumonia detection using cnn based feature extraction. In: In Proc. International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp 1–7 (2019)
60. Wan J, Yilmaz A, Yan L. Dcf-bow: Build match graph using bag of deep convolutional features for structure from motion. IEEE Geosci Remote Sens Lett 2018;15(12):1847–51.
61. Zhou B, Khosla A, Lapedriza A, Torralba A, Oliva A. Places: an image database for deep scene understanding. arXiv preprint arXiv:161002055 (2016)
62. Zhou ZH, Jiang Y, Yang YB, Chen SF. Lung cancer cell identification based on artificial neural network ensembles. Artif Intell Med 2002;24(1):25–36.

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