Computers in Biology and Medicine 89 (2017) 135–143

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Pre-trained convolutional neural networks as feature extractors for tuberculosis detection

U.K. Lopes [b], J.F. Valiati [a,*]

[a] Artificial Intelligence Engineers - AIE, 262, Vieira de Castro Street, Porto Alegre, RS, Brazil
[b] DevGrid, 482, Italia Avenue, Caxias do Sul, RS, Brazil

ARTICLE INFO

Keywords:
Deep learning
Convolutional neural networks
Tuberculosis
Computer assisted diagnosis
Multiple instance learning
Ensemble learning

ABSTRACT

It is estimated that in 2015, approximately 1.8 million people infected by tuberculosis died, most of them in developing countries. Many of those deaths could have been prevented if the disease had been detected at an earlier stage, but the most advanced diagnosis methods are still cost prohibitive for mass adoption. One of the most popular tuberculosis diagnosis methods is the analysis of frontal thoracic radiographs; however, the impact of this method is diminished by the need for individual analysis of each radiography by properly trained radiologists. Significant research can be found on automating diagnosis by applying computational techniques to medical images, thereby eliminating the need for individual image analysis and greatly diminishing overall costs. In addition, recent improvements on deep learning accomplished excellent results classifying images on diverse domains, but its application for tuberculosis diagnosis remains limited. Thus, the focus of this work is to produce an investigation that will advance the research in the area, presenting three proposals to the application of pre-trained convolutional neural networks as feature extractors to detect the disease. The proposals presented in this work are implemented and compared to the current literature. The obtained results are competitive with published works demonstrating the potential of pre-trained convolutional networks as medical image feature extractors.

1. Introduction

According to the World Health Organization, tuberculosis (along with HIV/AIDS) is the most deadly infectious disease in the world [1]. In its global 2015 annual tuberculosis report, the World Health Organization estimated that 1.5 million people infected by tuberculosis died (1.1 million HIV-negative and 0.4 million HIV-positive) and 9.6 million had fallen ill with the disease in the previous year [1]. Findings in the 2016 report were even worse: 10.4 million new cases of the disease and 1.8 million deaths were reported (1.4 million HIV negative and 0.4 million HIV positive) [2]. Many of those deaths could have been prevented if the disease had been detected at earlier stages. Nowadays there are many highly accurate diagnosis methods based on molecular analysis and bacteriological culture, but unfortunately, most of them are cost prohibitive for mass adoption in developing countries, which are the most affected by the disease. The cheapest and most popular diagnosis techniques, such as sputum smear microscopy, are reported to have sensitivity issues [3].

Another popular diagnosis method uses frontal chest radiographic images, but unfortunately this method is limited by a need for qualified staff to individually check every radiograph. If an automated method capable of detecting the disease were found, then it could support the current methods of diagnosis and be used as a mass detection tool to screen large populations that could not be managed manually [4], thereby greatly diminishing costs and potentially saving many lives.

In recent years, deep learning techniques have achieved outstanding results in a broad range of machine learning tasks [5]. For image classification tasks, Convolutional Neural Networks (CNN) have proved to be specially powerful, being successfully applied in diverse areas such as galaxy morphology prediction [6], development of image-guided autonomous cars [7], face detection [8,9], large-scale video classification [10] and many others [11–13]. There are already many CAD systems applying CNNs to diagnose diseases [14–21], but its fapplication to tuberculosis detection remains limited.

In most published works, CNNs are applied to image classification in four different manners [22]: training the network weights from scratch (usually done only when there is a very large dataset available), fine-tuning the weights of an existing pre-trained CNN (achieves similar

* Corresponding author.
E-mail address: joao.valiati@ai-engineers.com (J.F. Valiati).

http://dx.doi.org/10.1016/j.compbiomed.2017.08.001
Received 12 February 2017; Received in revised form 1 August 2017; Accepted 1 August 2017

results to training from scratch but may work in smaller datasets [23]), using unsupervised pre-training to set initial weights before training the CNN and by using a pre-trained CNN (sometimes called an off-the-shelf or out-of-box CNN) as a feature extractor. The latter method, which is the focus of the present work, usually combines the features extracted from the CNN with preexisting handcrafted features to train a more powerful and accurate classifier.

This work aims to present three different approaches to the use of pre-trained CNNs (using simple and direct feature extraction, multiple instance learning, and ensembles of classifiers) to create a screener to complement the diagnosis of tuberculosis in frontal chest radiographs (CR). Each of these proposed approaches uses three different architectures of pre-trained CNNs (GoogLenet, ResNet, and VggNet) as features extractors, and use the Support Vector Machine (SVM) classifier to identify whether the images contain tuberculosis.

The most important contributions of this work are as follows: the presentation of a comparative analysis of the performance as feature extractors of some of the most important CNN architectures in a tuberculosis dataset, the proposal of a combination of pre-trained CNNs and a MIL (Multiple Instance Learning) algorithm [24], and the evaluation of the use of ensembles of classifiers trained on features extracted from different pre-trained CNNs as an alternative to the combination of handcrafted features and a pre-trained CNN. To our knowledge, this is the first time that any of these contributions is presented in the literature.

This paper is divided as follows: Section 2 presents the most relevant related works. Section 3 explores the datasets and the methodology adopted in the present work. Section 4 explains each proposal in details and presents its results. Section 5 presents a discussion of the results obtained by the present work and compares it to similar proposals in current literature. Finally, in Section 6, the most important findings and lessons are presented along with possible future directions for the improvement of this work.

2. Related works

The next subsections review the most important related works on the application of pre-trained CNNs to medical image classification, as well as the most important recent works on tuberculosis detection.

2.1. Medical image classification using pre-trained CNNs

Since [25], it is known that a CNN trained on the ImageNet dataset [26] learns such a comprehensive set of features that makes it capable of working as a feature extractor for visual recognition on a broad range of different domains, obtaining competitive results and at times outperforming the previous state-of-the-art methods [27–29].

In the medical domain, one of the first applications of pre-trained CNNs was [21] where an ImageNet-trained network is used to detect pleural effusion and enlarged heart condition in x-ray images. The CNN-trained classifier is compared to classifiers trained using features extracted by the algorithms LBP (Local Binary Patterns) [30,31], Gist [32] and PiCoDes (which according to the authors stands for picture codes, but also Pico-Descriptor) [33], a combination of popular algorithms like SIFT (Scale-Invariant Feature Transform) [34], Gist [32] and PHOG (Pyramid Histogram of Oriented Gradients) [35] optimized on a subset of ImageNet. In most tests, the classifier trained using pre-trained CNN features surpassed the alternatives, but the best results were achieved by combining the CNN and PiCoDe features, obtaining an AUC of 0.93 for detection of right pleural effusion and an AUC of 0.89 for enlarged heart condition.

The authors of [36] evaluate the use of CNNs in classifying colonic polyps. In this work, a comprehensive comparison is made of pre-trained CNNs, fine-tuned CNNs, networks trained from scratch, and classical handcrafted features. The combination of pre-trained CNNs and handcrafted features produced the best performance in the tests, outperforming even the CNNs trained specifically for the task.

In Ref. [18], a CNN pre-trained in the ImageNet dataset is used as a feature extractor to detect nodules in pulmonary tomographies. The features extracted using the CNN are fed into a SVM to identify whether the region contains a nodule. Their proposition is then compared to a commercial CAD system approved by the American Food and Drug Administration (FDA), applying both systems to a private dataset containing approximately 1000 images. The commercial CAD system obtained a superior performance, but the combination of the commercial system with the author's proposal increased the maximum obtained sensitivity from 0.68 to 0.71.

In Ref. [20], two different proposals for the application of pre-trained CNNs to detect perifissural nodules are presented and compared to handcrafted features. The proposals obtained an area under the curve (AUC) of 0.847, a result slightly worse than using handcrafted features (AUC of 0.868) but still similar to the results obtained by trained specialists analyzing the images manually.

2.2. Tuberculosis detection

Tuberculosis infections can be perceived in CRs by the detecting the presence of specific patterns in the images. According to [37], among the most common manifestations of tuberculosis in radiographs are: air space consolidation (appears as opacity in the lobes), miliary patterns (a sand-like pattern appearing throughout the lungs), adenopathy (enlargement of the lymph nodes), airways enlargement (appears as tubular rings) and pleural effusion (indistinctness in lateral and medial regions). For an in-depth view of tuberculosis manifestations in radiographs see Ref. [38]. It is important to note that the patterns found in patients infected with tuberculosis are often also found in patients infected with other pulmonary diseases and that in most cases the patterns do not appear in isolation but in a combined form in the same radiograph.

One of the first proposals for a CAD system to detect tuberculosis is [39], wherein a multiscale filter bank is applied on the lung images to extract features. The classification is done using a weighted nearest-neighbor scheme, and the validation is achieved using LOOCV (leave-one-out cross-validation). The technique was applied to two small private datasets, reaching AUC of 0.82 and 0.986. In Ref. [40] the authors propose a tuberculosis detection technique combining a pixel-level textural abnormality analysis with other techniques. They obtained AUC between 0.67 and 0.86. In Ref. [41] a semi-automated method is proposed wherein statistical information about the pixel distribution is used as input to a decision tree classifier. All tests were executed in a small private dataset and the best accuracy obtained was 0.949.

A comparison of the aforementioned studies was hampered by the lack of a common public dataset used by all authors. The first public datasets of chest radiographs for tuberculosis detection were published, almost concomitantly, in Refs. [42] and [43].

Jaeger et al. [4] composed a proposal to combine many standard computer vision algorithms to extract features of radiographic images. The proposal is divided into three steps: segmentation of the region of interest [44], feature extraction using a combination of algorithms (such as histogram of oriented gradients [45], local binary patterns [30], hu moments [46], tamura texture descriptor [47,48] and others), and classification in which the extracted features are fed to a binary classifier to identify the image as healthy or not healthy. The classifiers evaluated were the Multilayer Perceptron (MLP), SVM, decision trees and logistic regression. The authors also applied an unspecified feature selection method to remove unnecessary features. The best results were obtained through logistic regression and linear SVM: AUC of 0.87 in the Montgomery dataset and 0.90 in the Shenzhen dataset. The accuracies were respectively 0.78 and 0.84.

Chauhan et al. [42] proposed a framework to identify tuberculosis in CR images. This framework is composed by a set of modules that must follow a sequence of steps to perform the classification. From the CR dataset, it starts with a preprocessing module based on wavelet

denoising, followed by a feature extraction module based on Gist and PHOG (Pyramid Histogram of Oriented Gradients) features. After that, the relevant features are selected based on chi-square distribution. Finally, the SVM classifier with radial kernel is applied to build the model. The best results attained through Gist and PHOG features are: an accuracy of 0.94 and 0.92 in DA dataset and 0.86 and 0.92 in DB dataset, respectively. Additionally, the authors developed a toolbox that can be installed in x-ray machines to be used for the screening of tuberculosis.

Another important work is [49], which makes a novel proposal to detect tuberculosis using a MIL technique. In Melendez proposal, each CR is divided into unlabeled subregions (called instances), and features based on the moments of pixel intensity distributions are extracted. In this approach the labels of each subregion are unknown, but the label of the whole, undivided CR (called a "bag" in MIL terminology) is known. The standard SVM formulation is not capable of classifying the radiographs at an instance level. Hence, the authors propose the use of a reformulated SVM called MiSVM [50], which enables the classification of groups of samples instead of only individual isolated samples. The authors evaluated the results in three proprietary datasets called Gambia, Tanzania and Zambia. The results attained by the authors are competitive with the current literature, with an AUC between 0.86 and 0.91. Unfortunately the proposal was not evaluated on any of public datasets.

Hwang et al. proposed for the first time the application of a CNN to tuberculosis detection [51]. The proposal includes the creation of a custom network, adapting an existing CNN to the specific problem of tuberculosis detection and recalculating the existing weights. The proposed architecture is a variant of the AlexNet network [52] with a larger input layer (500 × 500 pixels), a new convolutional layer right after the input layer (to process the higher-resolution images) and a new max-pooling layer. The network was trained using a large private dataset consisting of approximately 10,000 images. The network's initial weights were set in two different configurations: random and using the weights learned by the standard AlexNet in the ImageNet dataset. The results attained by randomly initializing the weights were unsatisfactory, producing an accuracy of 0.77 and an AUC of 0.82. In the case of the network trained with pre-learned weights, the accuracy reached 0.90 and the AUC reached 0.96. The trained model was also applied in the classification of the Montgomery and Shenzhen datasets where the obtained results are competitive in most cases. In the Montgomery dataset the accuracy was 0.674 and the AUC 0.884, whilst in the Shenzhen dataset the accuracy was 0.837 and the AUC was 0.926.

3. Materials and methodology

The datasets and the methods used for the development of the CAD system to detect tuberculosis are described in the next subsections.

3.1. Datasets

All experiments of this work were tested in the public datasets Shenzhen and Montgomery, published in Ref. [43], and partially in the datasets named DA and DB, as described in Ref. [42].

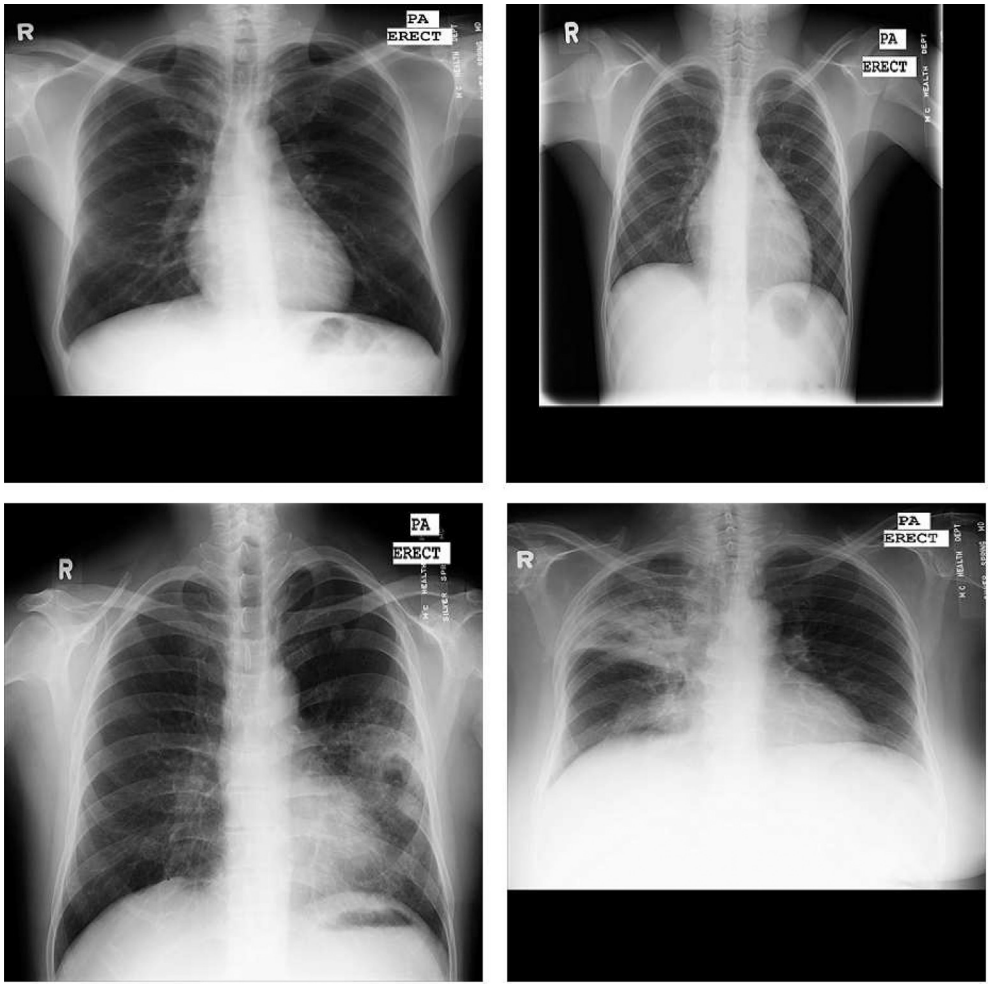

Fig. 1. Examples of CRs in the Montgomery dataset. The top left and top right CRs show normal lungs from a 33-year-old man and from an 8-year-old boy, respectively. Both CRs on the bottom show cases of tuberculosis infection; the left one is from a 54-year-old man with infiltrates in both lungs and a visible cavity in the lingula. The bottom right CR contains a large infiltrate in the right upper lobe and a cavitation plus infiltrate in the right middle lobe. These signs are consistent with active cavitary tuberculosis.

The Montgomery dataset consists of 138 frontal chest x-ray images, 80 of them are CRs of lungs not containing the disease, and 58 contain lungs infected by tuberculosis. All images in this set have been collected by the health department of the Montgomery County in Maryland, USA. The radiographs' resolutions are either 4,020 × 4,892 or 4,892 × 4,020 pixels. This dataset includes additional images containing manually generated lung segmentation masks for every sample in the set. Fig. 1 shows samples of this dataset.

The Shenzhen dataset was collected in the *Guandong* Hospital in Shenzhen, China. The dataset contains 662 frontal CRs, 336 of which are infected by tuberculosis, and 326 of which are not infected by the disease. All image resolutions are around 3,000 × 3,000 pixels. Fig. 2 shows examples of CRs in this dataset.

The datasets DA and DB represent frontal CRs from two different x-ray machines of the National Institute of Tuberculosis and Respiratory Diseases in New Delhi. Both DA and DB datasets contain a balanced set of images with 78 and 75 samples, respectively, for each class (infected and not infected by tuberculosis). The resolutions range from 1024 × 1024 to around 2480 × 2480. No additional image description is provided by the authors.

3.2. Preprocessing

All images present in the datasets used during the development of this work are frontal thoracic CRs and contain regions outside of the lungs that are not relevant to tuberculosis detection. To diminish the risk that features present in the images, but irrelevant to tuberculosis detection distort the final results, we decided to segment the lung regions of the CRs. Two approaches are used to segment the lungs in the CRs. For the Montgomery dataset, the segmentation is trivial because the dataset includes segmentation masks for every image. For the Shenzhen dataset, segmentation masks where generated using the approach proposed in Ref. [44], in which a mapping is calculated between each image and the most similar present in a pre-annotated dataset (called Atlas Set) using the algorithm SIFT Flow [53]. A lung model is then created by averaging the calculated maps. In the last step, a discrete optimization is performed using the graph-cuts algorithm [54] and a customized energy function. The segmentation generated was evaluated by the authors in three different datasets, revealing accuracies of 0.95, 0.94, and 0.92 [55].

After the segmentation of the lung region, the resulting image is cropped to the size of a minimum bounding box containing all the pixels of the lungs. Afterward, three different proposals for tuberculosis detection were evaluated. They are presented in the next subsections.

3.3. Proposal 1 - Simple CNN feature extraction

The first proposal aims to evaluate in simplest way the capacity of three CNN architectures pre-trained in the ImageNet dataset to extract features that are relevant to the classification of radiographic images. The architectures evaluated are GoogLenet [56] (the winner of the classification task in ImageNet Large Scale Visual Recognition Challenge [ILSVRC] 2014), ResNet [57] (winner of ILSVRC 2015), and VggNet [58]

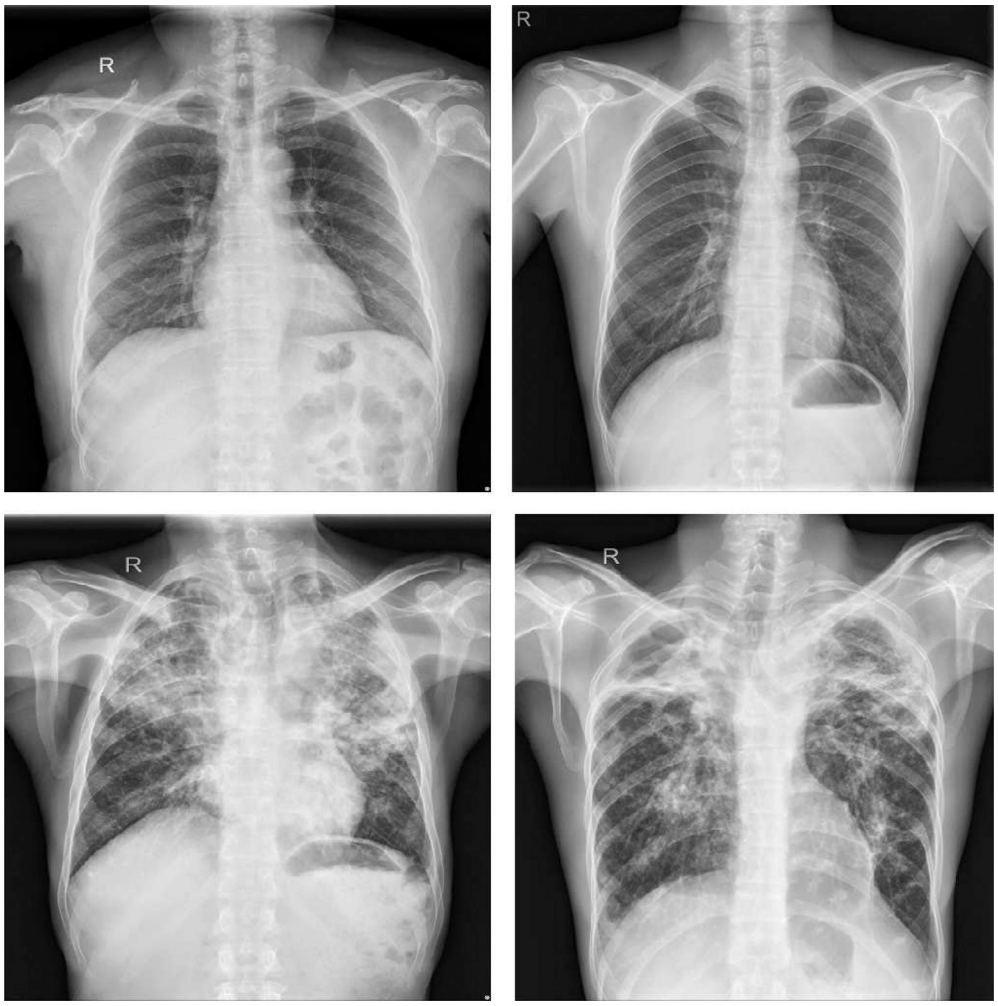

Fig. 2. Examples of CRs in the Shenzhen dataset. The top left CR is from a 48-year-old woman, and the top right from a 24-year-old man. Neither CR contains signs of tuberculosis. The bottom two images are cases of bilateral secondary tuberculosis: the left CR is from a 56-year-old male and the right CR is from a 26-year-old male.

(winner of ILSVRC's localization task in 2014).

After the preprocessing, the image containing the segmented lungs is resized to the dimensions of the input layer of each CNN (224 × 224 for GoogLenet and VggNet and 227 × 227 for ResNet), and then propagated in the network. After this, the output of the last layer before classification is extracted and used as an input to the training of a SVM. The choice of layer to extract the features is based on experiments made by Ref. [25]. Fig. 3 shows all the steps for this proposal.

For the training of the SVM, both linear and RBF kernels were evaluated. The validation was done using a nested [59] scheme with a 5-fold validation in the outer loop and in the inner loop a LOOCV in Montgomery dataset and 10-fold in Shenzhen. The best parameters for the classifier were selected by a grid search wherein the C parameter ranged from 1 to 1000 and the gamma from 1/4096 to 1.

3.4. Proposal 2 - Bag of CNN features

The first proposal has a crucial disadvantage. Since the lung images have to be resized to fit the CNN input layer, a lot of potentially useful information for identifying tuberculosis signs may be lost. The most straightforward way to avert this problem would be a sliding window approach, wherein each window is classified separately. Unfortunately, this approach is not viable because the datasets do not include positional information indicating the regions where signs of the disease are found. To avoid this, the second proposal formulates the disease classification as a multiple instance learning problem [24]. In a standard supervised classification problem, a class label is associated with each sample. In a MIL problem, the class labels are associated with groups of samples called bags. Each bag contains many individual samples (which are called instances in MIL terminology) whose class labels are unknown [60].

In this proposal, the same three CNN architectures (GoogLenet, VggNet, and ResNet) are used as features extractors, with the difference being that the CRs are not resized. Instead, each CR is divided into subregions (with 50% overlapping to prevent loss of relevant information) whose size is equal to the networks input layer. Using the MIL terminology, each subregion is an instance and each radiograph is a bag.

The first step is the generation of a dictionary of visual characteristics by clustering the vectors extracted from the instances of the bags. Each subregion is fed into the network and, as in the first proposal, the output of last layer before classification is extracted. The extracted vectors are clustered using the K-means algorithm [61], thus creating a dictionary of visual characteristics. The dictionary is later used to generate the global bag descriptor: a histogram of occurrence counts of the dictionary's visual characteristics. After obtaining the global bag descriptor, the problem is now a conventional supervised classification problem, since every bag has a single feature vector and a single label. Essentially, this method is similar to the bag-of-words (BOW) model [62] used for image classification. Figs. 4 and 5 contain diagrams showing, respectively, the steps for the generation of the dictionary and for the classification of the CRs.

In the creation of the dictionary, the parameter K (which indicates the size of the dictionary) was heuristically set to four different values: 100, 200, 300, and 500. Each dictionary was created using the same methodology presented in Ref. [24]. In the classification stage, as in Proposal 1, the SVM classifier is used with both linear and RBF kernels, a grid search for parameter optimization. For validation, a nested cross validation was used. In the outer loop a 5-fold validation was applied and in the inner loop a LOOCV was used in the Montgomery dataset and, for performance reasons, a 10-fold validation in the Shenzhen dataset.

3.5. Proposal 3 - Ensembles

As stated in previous sections, there are many applications of CNNs as feature extractors in medical imaging classification. But, to our knowledge, there are no published works examining the results of applying ensembles of classifiers trained using features extracted from multiple pre-trained CNNs.

It is well known that using ensemble methodology, it is possible to combine multiple individual classifiers to create a new, more accurate classifier [63–65]. The aim of the third and last proposal is to create ensemble classifiers by combining the SVMs trained using the features extracted from GoogLenet, ResNet, and VggNet.

Ensemble classifiers were created for Proposal 1 (one for each dataset) and Proposal 2 (one for Montgomery and one for Shenzhen). All ensembles combine three classifiers (one using features from GoogLenet, one using features from ResNet, and one using features from VggNet) and obtain an output through a simple soft-voting scheme.

The methodology applied to obtain the best models for each proposal and each network architecture, as mentioned in the previous sections 3.3 and 3.4, served as the base to implement the ensemble proposition. Based on the best models obtained in this early stage, the ensembles were created and the results were calculated using 5-fold cross validation.

4. Obtained results

In this section, the results obtained in datasets Shenzhen and Montgomery are given more attention than DA's and DB's results. This is due to their popularity in current literature which allowed us to better compare the obtained results of all proposals. Also due to the lower resolution of some the images present in DA and DB datasets the application of Proposal 2 was not implemented.

Table 1 shows the results for Proposal 1 in terms of accuracy and AUC. For the Montgomery dataset, the best results were attained by GoogLenet

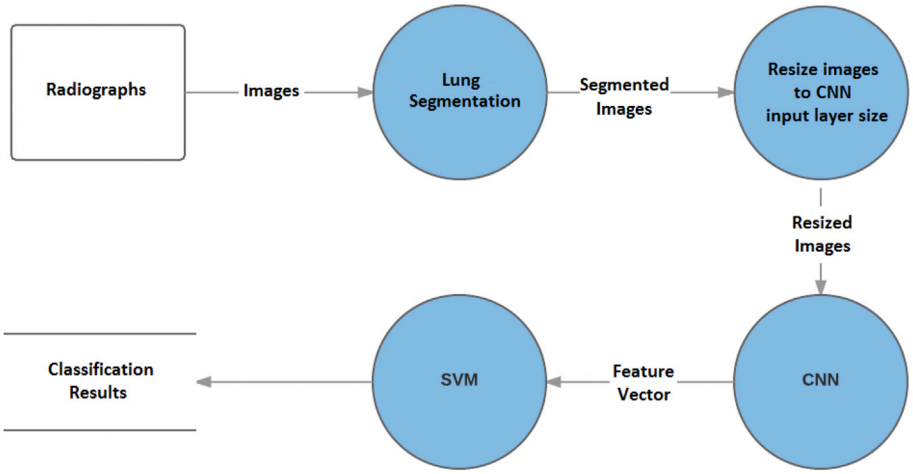

Fig. 3. Diagram showing the steps for Proposal 1.

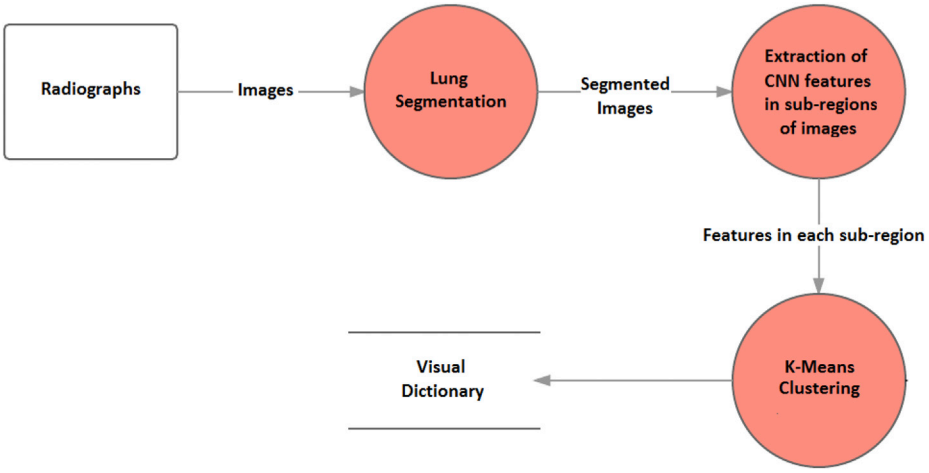

Fig. 4. Diagram showing the steps for the generation of the dictionary of visual characteristics in Proposal 2.

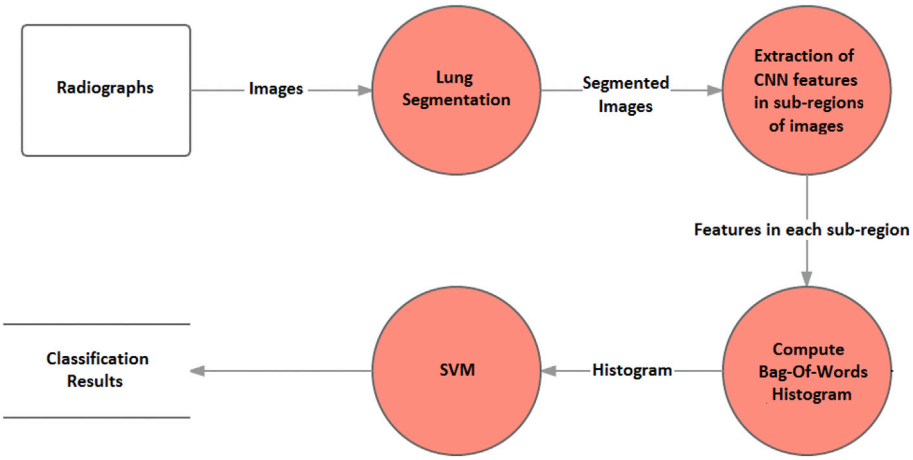

Fig. 5. Diagram showing the steps for classification in Proposal 2.

with an accuracy of 0.782 and AUC of 0.838. In the Shenzhen dataset, the network ResNet obtained the best accuracy, with 0.834, while the best AUC was attained by VggNet, with 0.912.

It can be noted from viewing this table that the results attained in the Shenzhen dataset are superior to the results of the Montgomery dataset. The same pattern is seen in all tables presented in this work, and also in related published works [4,51]. The most important factor impairing the results in Montgomery is probably the limited size of the dataset, containing only 138 samples, while Shenzhen has 662 samples. Also, the class imbalance found in the Montgomery dataset (in which 60% of the samples are negative and 40% are positive) in comparison to Shenzhen (which has an almost 50/50 ratio) could be another possible factor.

Tables 2 and 3 show the results for the second proposal, revealing the best accuracy and AUC for each dictionary size (parameter K) value.

ResNet presents more stable accuracies considering the different values of K. In the Montgomery dataset (Table 2), the best accuracy reached 0.826 for $K = 200$ in the GoogLenet network. The best accuracy for ResNet and VggNet was attained for $K = 500$. The best AUC in Montgomery dataset was reached when $K = 500$. Note that GoogLenet and ResNet achieve very similar results.

In the Shenzhen dataset (Table 3), ResNet gives consistently inferior accuracies when compared to the other CNNs for all K parameter values. GoogLenet achieved the best performance in this dataset with an accuracy of 0.847 for $K = 300$. For VggNet the best results, considering the

Table 1
Proposal 1 - Results for different architectures.

Datasets	Accuracy			AUC		
	GoogLenet	ResNet	VggNet	GoogLenet	ResNet	VggNet
Montgomery	0.782	0.746	0.753	0.838	0.776	0.777
Shenzhen	0.822	0.834	0.828	0.903	0.900	0.912

Table 2
Proposal 2 - Montgomery.

K	Accuracy			AUC		
	GoogLenet	ResNet	VggNet	GoogLenet	ResNet	VggNet
100	0.782	0.804	0.725	0.874	0.874	0.821
200	0.826	0.804	0.733	0.884	0.884	0.827
300	0.762	0.804	0.746	0.887	0.887	0.815
500	0.782	0.810	0.803	0.925	0.926	0.831

Table 3
Proposal 2 - Shenzhen.

K	Accuracy			AUC		
	GoogLenet	ResNet	VggNet	GoogLenet	ResNet	VggNet
100	0.818	0.811	0.816	0.885	0.884	0.900
200	0.834	0.799	0.835	0.892	0.882	0.904
300	0.847	0.804	0.832	0.885	0.889	0.891
500	0.823	0.808	0.837	0.887	0.894	0.899

AUC, are achieved with K under 300. It is interesting to note that all values of K give very similar AUC results in all the network architectures.

Table 4 shows the results obtained by the ensembles built using classifiers from Proposals 1 and 2. An interesting and unexpected finding here is that the ensemble of Simple CNN Features (Proposal 1) managed to achieve the best accuracy of 0.846 in the Shenzhen dataset, which is the same result achieved by the ensemble of Proposal 2 classifiers. The best accuracy for Montgomery dataset was attained with Bag of Features, 0.826. The best AUC for Montgomery was also found with Bag of Features, 0.908 and the best AUC for Shenzhen reached 0.926 with Simple Features proposal.

In addition to the main tests executed in the Shenzhen and Montgomery datasets, extra experiments were executed in the smaller DA and DB datasets made public by Ref. [42]. Both Proposal 1 and part of Proposal 3 (the ensemble of simple CNN classifiers) were applied to the databases and the best results were achieved by the ensembles of classifiers. For validation, a nested cross-validation was applied with a 5-fold outer loop and a LOOCV inner loop. The obtained accuracy was 0.801 for the dataset DA and 0.828 for dataset DB and the AUC was 0.868 and 0.912 for datasets DA and DB respectively. The results obtained by Chauhan et al. [42] can be seen in Section 2.2. No experiments were conducted applying Proposal 2 due to the lower resolution of some of the images present in both DA and DB datasets. The main idea behind the second proposal is to extract features from a statistically significant number of sub-windows in a high resolution image without losing any information, conditions which are not always met in DA and DB datasets.

Table 4
Ensemble of CNNs.

Datasets	Simple Features		Bag of Features	
	Accuracy	AUC	Accuracy	AUC
Montgomery	0.760	0.834	0.826	0.908
Shenzhen	0.846	0.926	0.846	0.910

Table 5
Comparison - Accuracy.

Datasets	Literature approaches		Our approach			
	[4]	[51]	P1	P2	EP1	EP2
Montgomery	0.783	0.674	0.782	**0.826**	0.760	**0.826**
Shenzhen	0.840	0.837	0.834	**0.847**	0.846	0.846

The bold values mean the more elevated results obtained for each dataset and each approach.

Table 6
Comparison - AUC.

Datasets	Literature approaches		Our approach			
	[4]	[51]	P1	P2	EP1	EP2
Montgomery	0.869	0.884	0.838	**0.926**	0.834	0.908
Shenzhen	0.900	**0.926**	0.912	0.904	**0.926**	0.910

The bold values mean the more elevated results obtained for each dataset and each approach.

5. Discussion

The last two Tables 5 and 6, compare the results obtained in all the proposals presented in this paper and in similar works about tuberculosis detection [4,51]. The columns named P1 and P2 refer to Proposal 1 and Proposal 2, and columns EP1 and EP2 refer to the ensembles created using the classifiers trained in Proposals 1 and 2, respectively.

In terms of accuracy (Table 5), the proposals P2 and EP2 outperform both [4] and [51] with the exact same result of 0.826. For Shenzhen dataset, 3 out of 4 approaches surpassed the literature, with highlights to the proposition P2 that attained 0.847.

In terms of AUC (Table 6), P2 and EP1 achieved very good performance. For the Montgomery dataset, the P2 was the winner with a result of 0.926 and surpassed the literature approaches. In the Shenzhen dataset, our ensemble proposal attained the same result reached by Hwang et al., with an AUC of 0.926.

As shown in the comparative tables, our proposals achieved results that are superior in many cases and, at worst, are competitive with proposals of similar published works.

The results achieved in Proposal 1, in almost all cases, are inferior to both the published literature approaches and our other presented proposals. The main reason for that is, most likely, the fact that all radiographs must be resized to the dimensions of the input layer of the networks. In some cases that is less than 1/20 of the original size.

Proposal 2's results were superior when compared to Proposal 1 in most cases, as expected. Here the decision to avoid the loss of information by using the radiographs in their original size brought important performance improvements. It is interesting to note that Proposal 2 also managed to obtain superior results even when compared to Hwang's proposal [51], which uses a fine-tuned CNN. In this case again, the most likely reason for the superior result is the use of the radiographs in a high resolution. Even though Hwang's proposal resizes the images to a higher resolution than our first proposal (500×500 vs 224×224), it probably still lost important information that was successfully used by our second Proposal to correctly detect the disease. The main disadvantage of our second approach is the complexity it introduces to the disease detection process, with additional stages of clustering and histogram generation before the classification. Another possible problem, depending on the hardware where the process is executed, is the longer execution time: on average, each radiography is divided into more than 60 subregions, and that may considerably slow the feature extraction process.

The results obtained by Proposal 3 were underwhelming, to say the least. Ensembles of pre-trained CNNs had already been applied in a different domain to improve upon the performance of individual CNNs [66] but unfortunately that was not the case in our experiments. One of the most important requirements in the creation of ensembles of classifiers is a diversity of errors; i.e., the errors of the base classifiers must have a low correlation [67]. Since all CNNs were trained on the same dataset (ImageNet), there was a risk that their outputs would be too similar, with a high correlation in their errors, thus impairing the creation of the ensembles. As seen in the results tables, that is exactly what happened. The ensembles results are not, in the majority of cases, superior to the base classifier results and in some cases the results even are inferior.

Among the CNN architectures evaluated in this paper, none is clearly superior to the others. For Proposal 1 GoogLenet achieved the best results in the Montgomery dataset and VggNet and ResNet did so in Shenzhen (Table 1). In Proposal 2 ResNet's results are the most stable in Montgomery in terms of accuracy and GoogLenet and VggNet are the winners in Shenzhen (Tables 2 and 3). It could be expected that ResNet would beat all other architectures since its results are clearly superior in ImageNet, but that was not the case. ResNet's very deep architecture, comprising 152 layers, seems to be overkill for binary medical image classification tasks. For ImageNet benchmarks it makes a lot of sense to create deeper and deeper networks because the dataset is so vast and diverse and requires the knowledge of a large number of abstractions to account for all possible classes. For disease detection in medical images the data variability is many orders of magnitude smaller and there doesn't seem to be a need for very deep networks to improve the performance.

6. Conclusion

In this paper, three different proposals were presented for the application of pre-trained CNNs in tuberculosis detection. In the first proposal, three different CNN architectures are used to extract features from a

resized radiographic image. The extracted features are then used to train a SVM classifier. In the second proposal, the same three CNN architectures are used to extract features from subregions of the CR. The extracted features are then combined to create a single global descriptor that is used to train a SVM. In the last proposal, the best SVMs trained on Proposals 1 and 2 are used to create ensembles of classifiers.

Pre-trained networks possess important characteristics that make them natural candidates when applying deep learning to medical image classification tasks: they do not require an expensive and time-consuming training step nor do they need a large dataset to achieve reasonably good results. Of course, their performance usually is inferior to a network trained specifically for the desired task. In some cases, it has been shown that CNNs with fine-tuned weights consistently outperform pre-trained networks, even in smaller datasets [68]. For those reasons, the use of pre-trained networks as feature extractors is usually not the favored way to apply CNNs to CAD classification tasks. If that is the case, then does it make sense to use pre-trained CNNs at all? The most common use of pre-trained networks for CAD tasks in current literature is in combination with previously hand-crafted features [21,36,68] to train a new classifier that is able to attain superior results.

Based on the results obtained in the present work, we believe that pre-trained networks can be a very useful and powerful tool in the toolbox for other cases. The most interesting use case, in our opinion, is to apply techniques similar to our Proposal 2 in high-resolution datasets. High-resolution images present a more difficult problem to CNNs since the time required to train the weights of a deep network could be prohibitive. In this case, the use of a pre-trained CNN to extract features of subregions of the images in combination with a MIL technique is a more viable way to apply deep networks to the task at hand.

As for future advancements in tuberculosis detection, recent works such as [69] indicate that the future lies in building increasingly large datasets. In the aforementioned work, a more recent version of GoogLenet is trained using approximately 130,000 dermatological images for the detection of skin cancer, and it achieves results on par with human experts. This suggests that the way forward for all CAD systems is to create ever larger datasets that can be used to train deep networks. If a dataset of annotated lung radiographs as large as the one used in the aforementioned work were built, then it could be expected with a high degree confidence that the results obtained on skin cancer detection could be repeated in tuberculosis detection, thus enabling the creation of better CAD systems for the disease. Of course, building up such large datasets is a long and expensive endeavor. For the foreseeable future, the research will probably need to be conducted in smaller datasets.

6.1. Future works

Many improvements could be made to the techniques presented in this work. The most important would probably be to further study the performance of ensembles of pre-trained CNNs in medical imaging. The ensembles created in this work use a very simple majority voting scheme. It could be interesting to evaluate different voting methods and analyze different CNN architectures.

In the second proposal presented here, the K-means algorithm is used to generate a dictionary of visual characteristics present in the radiographic images. There are many other clustering algorithms that could be used in the dictionary generation stage, such as the Expectation Maximization (EM) algorithm. At least one published work exists wherein EM obtained superior results as a dictionary generation algorithm when compared to K-means [24].

Another possible improvement could be made by using samples from other public datasets. As of 2016, Montgomery and Shenzhen are the only public datasets containing chest radiographs with tuberculosis annotations. However, there are other chest radiograph datasets, such as the JSRT Digital Image Database [70], contain annotations indicating the presence of lung tumors. By using the additional samples, perhaps a generic lung anomaly classifier could be created and used to detect tuberculosis.

Acknowledgments

We thank CAPES (Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior) for the financial support. CAPES had no involvement in the study design, in the collection, analysis and interpretation of data; in the writing of the manuscript; and in the decision to submit the manuscript for publication.

References

[1] Global Tuberculosis Report 2015, 2015.
[2] Global Tuberculosis Report 2016, 2016.
[3] C.C. Leung, Reexamining the role of radiography in tuberculosis case finding, Int. J. Tuberc. Lung Dis. Official J. Int. Union Against Tuberc. Lung Dis. 15 (2011) 1279.
[4] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R.K. Singh, S. Antani, et al., Automatic tuberculosis screening using chest radiographs, Medical Imaging, IEEE Trans. 33 (2014) 233–245.
[5] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117.
[6] S. Dieleman, K.W. Willett, J. Dambre, Rotation-invariant convolutional neural networks for galaxy morphology prediction, Mon. Notices R. Astron. Soc. 450 (2015) 1441–1459.
[7] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, R. Cheng-Yue, F. Mujica, A. Coates, et al., An Empirical Evaluation of Deep Learning on Highway Driving, arXiv Prepr. arXiv:1504.01716, 2015.
[8] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325–5334.
[9] S.S. Farfade, M.J. Saberian, L.-J. Li, Multi-view face detection using deep convolutional neural networks, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, 2015, pp. 643–650.
[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[11] J. Zbontar, Y. LeCun, Computing the stereo matching cost with a convolutional neural network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1592–1599.
[12] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE Trans. Pattern Analysis Mach. Intell. 35 (2013) 221–231.
[13] S. Sudholt, G.A. Fink, Phocnet: A Deep Convolutional Neural Network for Word Spotting in Handwritten Documents, arXiv preprint arXiv:1604.00187, 2016.
[14] G. Apou, N.S. Schaadt, B. Naegel, G. Forestier, R. Schönmeyer, F. Feuerhake, C. Wemmert, A. Grote, Detection of lobular structures in normal breast tissue, Comput. Biol. Med. 74 (2016) 91–102.
[15] K.-L. Hua, C.-H. Hsu, S.C. Hidayati, W.-H. Cheng, Y.-J. Chen, Computer-aided classification of lung nodules on computed tomography images via deep learning technique, OncoTargets Ther. 8 (2015).
[16] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, H. Larochelle, Brain Tumor Segmentation With Deep Neural Networks, arXiv preprint arXiv:1505.03540, 2015.
[17] J. Kawahara, A. BenTaieb, G. Hamarneh, Deep features to classify skin lesions, in: Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on, IEEE, 2016, pp. 1397–1400.
[18] B. van Ginneken, A.A. Setio, C. Jacobs, F. Ciompi, Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans, in: Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, IEEE, 2015, pp. 286–289.
[19] Y. Miki, C. Muramatsu, T. Hayashi, X. Zhou, T. Hara, A. Katsumata, H. Fujita, Classification of teeth in cone-beam ct using deep convolutional neural network, Comput. Biol. Med. 80 (2017) 24–29.
[20] F. Ciompi, B. de Hoop, S.J. van Riel, K. Chung, E.T. Scholten, M. Oudkerk, P.A. de Jong, M. Prokop, B. van Ginneken, Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box, Med. Image Anal. 26 (2015) 195–202.
[21] Y. Bar, I. Diamant, L. Wolf, H. Greenspan, Deep learning with non-medical training used for chest pathology identification, in: SPIE Medical Imaging, International Society for Optics and Photonics, 2015, 94140V–94140V.
[22] H.C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, Deep convolutional neural networks for computer-aided detection: cnn architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (2016) 1285–1298.
[23] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging 35 (2016) 1299–1312.
[24] J. Amores, Multiple instance classification: review, taxonomy and comparative study, Artif. Intell. 201 (2013) 81–105.
[25] A. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: an astounding baseline for recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 806–813.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 248–255.
[27] G.D. Juraszek, A.G. Silva, A.T. da Silva, Reconhecimento de produtos por imagem utilizando palavras visuais e redes neurais convolucionais, 2014.
[28] O. Penatti, K. Nogueira, J. Santos, Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015.
[29] F. Bousetouane, B. Morris, Off-the-shelf cnn features for fine-grained classification of vessels in a maritime environment, in: International Symposium on Visual Computing, Springer, 2015, pp. 379–388.
[30] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 971–987.
[31] T. Ahonen, A. Hadid, M. Pietikäinen, Face recognition with local binary patterns, in: European Conference on Computer Vision, Springer, 2004, pp. 469–481.
[32] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (2001) 145–175.
[33] A. Bergamo, L. Torresani, A.W. Fitzgibbon, Picodes: learning a compact code for novel-category recognition, in: Advances in Neural Information Processing Systems, 2011, pp. 2088–2096.
[34] D.G. Lowe, Object recognition from local scale-invariant features, in: Computer Vision, 1999, The Proceedings of the Seventh IEEE International Conference on, vol. 2, IEEE, 1999, pp. 1150–1157.
[35] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, ACM, 2007, pp. 401–408.
[36] E. Ribeiro, A. Uhl, G. Wimmer, M. Häfner, Exploring deep learning and transfer learning for colonic polyp classification, Comput. Math. Methods Med. 2016 (2016).
[37] S. Antani, Automated Detection of Lung Diseases in Chest X-rays, Technical Report to the LHNCBC Board of Scientific, 2015.
[38] C.L. Daley, M. Gotway, R. Jasmer, Radiographic manifestation of tuberculosis, A Primer Clin. 1 (2003).
[39] B. Van Ginneken, S. Katsuragawa, B.M. ter Haar Romeny, M.A. Viergever, et al., Automatic detection of abnormalities in chest radiographs using local texture analysis, Medical Imaging, IEEE Trans. 21 (2002) 139–149.
[40] L. Hogeweg, C. Mol, P.A. de Jong, R. Dawson, H. Ayles, B. van Ginneken, Fusion of local and global detection systems to detect tuberculosis in chest radiographs, in: Medical Image Computing and Computer-assisted Intervention–miccai 2010, Springer, 2010, pp. 650–657.
[41] J.H. Tan, U.R. Acharya, C. Tan, K.T. Abraham, C.M. Lim, Computer-assisted diagnosis of tuberculosis: a first order statistical approach to chest radiograph, J. Med. Syst. 36 (2012) 2751–2759.
[42] A. Chauhan, D. Chauhan, C. Rout, Role of gist and phog features in computer-aided diagnosis of tuberculosis without segmentation, PLoS One 9 (2014) e112980.
[43] S. Jaeger, S. Candemir, S. Antani, Y.-X.J. Wáng, P.-X. Lu, G. Thoma, Two public chest x-ray datasets for computer-aided screening of pulmonary diseases, Quant. Imaging Med. Surg. 4 (2014) 475–477.
[44] S. Candemir, S. Jaeger, K. Palaniappan, S. Antani, G. Thoma, Graph-cut based automatic lung boundary detection in chest radiographs, in: IEEE Healthcare Technology Conference: Translational Engineering in Health & Medicine, 2012, pp. 31–34.
[45] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005, IEEE Computer Society Conference on, vol. 1, IEEE, 2005, pp. 886–893.
[46] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Trans. Inf. Theory 8 (1962) 179–187.
[47] H. Tamura, S. Mori, T. Yamawaki, Textural features corresponding to visual perception, IEEE Trans. Syst. Man, Cybern. 8 (1978) 460–473.
[48] P. Howarth, S. Ruger, Robust texture features for still-image retrieval, IEE Proc. Vision, Image Signal Process. 152 (2005) 868–874.
[49] J. Melendez, B. van Ginneken, P. Maduskar, R.H. Philipsen, K. Reither, M. Breuninger, I.M. Adetifa, R. Maane, H. Ayles, C.I. Sanchez, A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays, Medical Imaging, IEEE Trans. 34 (2015) 179–192.
[50] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Advances in Neural Information Processing Systems, 2003, pp. 561–568.
[51] S. Hwang, H.-E. Kim, J. Jeong, H.-J. Kim, A novel approach for tuberculosis screening based on deep convolutional neural networks, in: SPIE Medical Imaging, International Society for Optics and Photonics, 2016, 97852W–97852W.
[52] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[53] C. Liu, J. Yuen, A. Torralba, J. Sivic, W.T. Freeman, Sift flow: dense correspondence across different scenes, in: European Conference on Computer Vision, Springer, 2008, pp. 28–42.
[54] Y. Boykov, G. Funka-Lea, Graph cuts and efficient nd image segmentation, Int. J. Comput. Vis. 70 (2006) 109–131.
[55] S. Candemir, S. Jaeger, K. Palaniappan, J.P. Musco, R.K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, C.J. McDonald, Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration, IEEE Trans. Med. Imaging 33 (2014) 577–590.
[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[57] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, arXiv preprint arXiv:1512.03385, 2015.
[58] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-scale Image Recognition, CoRR abs/1409.1556, 2014.
[59] G.C. Cawley, N.L. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (2010) 2079–2107.
[60] L. Dong, A Comparison of Multi-instance Learning Algorithms, Ph.D. thesis, Citeseer, 2006.
[61] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297. Oakland, CA, USA.
[62] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Workshop on Statistical Learning in Computer Vision, vol. 1, ECCV, Prague, 2004, pp. 1–2.
[63] D. Opitz, R. Maclin, Popular ensemble methods: an empirical study, J. Artif. Intell. Res. 11 (1999) 169–198.
[64] R. Polikar, Ensemble based systems in decision making, IEEE Circuits Syst. Mag. 6 (2006) 21–45.
[65] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev. 33 (2010) 1–39.
[66] S. Pouyanfar, S.-C. Chen, Semantic event detection using ensemble deep learning, in: Multimedia (ISM), 2016 IEEE International Symposium on, IEEE, 2016, pp. 203–208.
[67] T.G. Dietterich, Ensemble methods in machine learning, in: International Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15.
[68] H.-C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, Deep convolutional neural networks for computer-aided detection: cnn architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (2016) 1285–1298.
[69] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118, http://dx.doi.org/10.1038/nature21056.
[70] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-I. Komatsu, M. Matsui, H. Fujita, Y. Kodera, K. Doi, Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules, Am. J. Roentgenol. 174 (2000) 71–74.