IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 37, No. 6, December, 2020, pp. 1075-1084
Journal homepage: http://iieta.org/journals/ts

Spatial Pyramid Pooling in Deep Convolutional Networks for Automatic Tuberculosis Diagnosis

Pike Msonda[1], Sait Ali Uymaz[1*], Seda Soğukpınar Karaağaç[2]

[1] Dept. of Computer Engineering, Faculty of Engineering and Natural Sciences, Konya Technical University, Konya 42250, Turkey
[2] Konya Education and Research Hospital, 97 New Meram Avenue, Meram Konya 42040, Turkey

Corresponding Author Email: sauymaz@ktun.edu.tr

https://doi.org/10.18280/ts.370620

Received: 20 October 2020
Accepted: 8 December 2020

Keywords:
automated diagnosis, deep convolutional neural networks, image classification, spatial pyramid pooling, tuberculosis

ABSTRACT

In recent decades, automatic diagnosis using machine-learning techniques have been the focus of research. Mycobacterium Tuberculosis (TB) is a deadly disease that has plagued most developing countries presents a problem that can be tackled by automatic diagnosis. The World Health Organization (WHO) set years 2030 and 2035 as milestones for a significant reduction in new infections and deaths although lack of well-trained professionals and insufficient or fragile public health systems (in developing countries) are just some of the major factors that have slowed the eradication of the TB endemic. Deep convolutional neural networks (DCNNs) have demonstrated remarkable results across problem domains dealing with grid-like data (i.e., images and videos). Traditionally, a methodology for detecting TB is through radiology combined with previous success DCNN have achieved in image classification makes them the perfect candidate to classify Chest X-Ray (CXR) images. In this study, we propose three types of DCNN trained using two public datasets and another new set which we collected from Konya Education and Research Hospital, Konya, Turkey. Also, the DCNN architectures were integrated with an extra layer called Spatial Pyramid Pooling (SPP) a methodology that equips convolutional neural networks with the ability for robust feature pooling by using spatial bins. The result indicates the potential for an automated system to diagnose tuberculosis with accuracies above a radiologist professional.

1. INTRODUCTION

Tuberculosis (TB) is a deadly endemic that ranks as one of the deadliest diseases in the world. TB caused by a pathogen called Mycobacterium tuberculosis commonly victimizes people in developing countries [1]. The deadliness of TB is reflected in the estimate provided by the World Health Organization (WHO) that found 1.3 million Human Immunodeficiency Virus (HIV)-negative people fell victim together with an additional 300 thousand HIV-positive people in just the year 2017 alone [1]. Efforts to reduce the number of people infected by TB are increasing, however, the lack of expert clinical care in developing countries reduces the effectiveness of these attempts. The existence of various types of TB medically requires different strategies in diagnosis and treatment. Pulmonary Tuberculosis, given an International Classification of Disease (ICD) of A15.0 is the most common type and the primary concern of this paper. This type mostly affects the chest cavity more especially the lungs, which in most cases results in prolonged coughing, chest pains, fatigue, and fever among the many symptoms associated with this type of TB. As stipulated by the WHO, people living with HIV are more likely to contract TB due to negative effects HIV has on the body's immune system [1]. The rise of deep learning technologies has increased relative to the advancement of computer hardware and software. In recent decades, the increase in computational power of Graphical Processing Units (GPU), as well as Computer Processing Units (CPU), has massively propelled research in computer vision, natural language processing and voice recognition [2, 3]. The influence of deep learning has led into creating innovative solutions that perform tasks at the same competence as a human being. Deep learning methodologies go a step further than conventional machine learning algorithms, usually by adding more layers as well as a combination of different methods. However, convolutional neural networks (CNNs) are the most suited for problems that involve images and videos [2, 4]. In line with the attempts by the WHO to completely eradicate TB by the year 2030, computer aided diagnosis is one of the methods that is helping achieve this goal.

In this study, we present a methodology that utilizes the DCNN in classifying TB affected patients using Chest X-Rays (CXR) one of the most common methods employed in radiology. This study aims to produce results with higher accuracy than other CNN models and traditional machine learning methods by using DCNN models with SPP technique. CXRs are a result of a controlled dose of ionization that assists in creating snapshots of internal body organs in the chest i.e., the lungs and heart.

2. RELATED WORK

CXRs provide one of the most common ways to diagnose

TB however; this usually requires a trained expert to be able read correctly. It is worth noting that in this paper an accuracy of 84% was obtained from a trained expert on a portion of the data used. Attempts by machine learning algorithms to classify CXR have produced promising results with algorithms like Support Vector Machines (SVM) performing considerably well. In this particular instance, data was segmented using graph cut algorithm before finally feeding the features into SVM [5]. A combination of clinical data together with X-ray based computer-aided detection attempt produced remarkable results given in terms of operating characteristic curve of (0.84, 0.74, 0.72) [6]. By experimenting on a patient database containing about 392 patients, the researchers first extracted clinical information related to every patient and their associated CXR. The CXRs features using CAD, the best features are ranked and subsequently given to a multiple learner fusion reads the CXRs that performs classification [6]. Similarly, a comparison study between CAD4TB (computer-aided diagnosis for TB) against clinical officers is done in Lusaka, Zambia one of the countries that have a large number of TB victims [7]. The study is drawing its results from 161 subjects where CAD4TB computes an abnormality score between 0 and 100. Four clinical officers also scored the 161 subjects [7]. Among the available 161 patients, 97 were positive for bacterial TB as well 120 patients had abnormal CXR. This study points out that CAD4TB system obtained compared results to a trained clinical officer by scoring an AUC of 0.73 which is in the range of 0.65-0.75 as scored by the clinical officers [7]. Further in the realm of machine learning, automatic TB screening achieved by using a Support Vector Machine (SVM) as a classifier on a public dataset from Shenzhen Hospital, China [8] and Montgomery County (MC), USA [8, 9]. Firstly, the lungs segmented by a common method called Graph Cut Based where the results of the segmentation go through feature computation and finally classification by SVM [6]. Similar to the goal of automatic TB diagnosis, CNN is in detecting TB bacilli from microscopic sputum smear images. This approach achieves 86.76% F-Score [10].

Used CNNs in several researches have shown remarkable performance on grid-like data including images and videos. An advantage of using CNNs is that they require no feature extraction as well as the ability to be transferrable through weights makes CNN the perfect algorithm to approach the problem of automatic tuberculosis from CXRs [4]. A recently published report by Lakhani and Sundaram [11] Radiological Society of North America (RSNA) confirms the potential of DCNN in the classifying CXRs. They present two DCNN architectures; AlexNet, GoogLeNet, both pre-trained on ImageNet and trained models on a public TB dataset. They also selected the best performing models were used to create ensembles [11]. Their results are remarkable achieving 99% accuracy with an expert radiologist augmentation [11]. Increased computation capacity both mobile and remotely provide the best opportunity to speed up TB diagnosis in areas where medical health resources are scant [12]. With mobile computing in Healthcare (m-Health) as the core motivation, a DCNN is deployed at an endpoint where TB diagnosis requests are processed [12]. This approach has the capacity that can greatly assist the effort in poor resource areas in speeding the chest radiography more especially about TB. It is worth mentioning that contrary to our approach, other methods exist that can be used to automatically diagnose TB. One such method uses CNN to detect tuberculosis bacilli from 22 microscopic sputum smear images. By using a CNN architecture comprised of 2 convolutional layers with filters of 32 (3 x 3), 3 convolutional layers with 128 (3 x 3) and 1 convolutional layer with 128 filters, they were able to achieve 0.9713, 0.784, 0.8676 recall, precision and F-score respectively [10]. Another more recently research approaches this problem by utilizing a pre-trained AlexNet architecture on a total 10,848 CXR obtained from the Korean Institute of Tuberculosis (KIT), 138 National Institutes of Health (NIH), USA and 662 images from Shenzhen No 3 People's Hospital, China. AlexNet is a famous DCNN architecture that achieved great renown in the 2012 ImageNet Large Scale Visual Recognition Competition (ILSVR). Using this architecture, the research presents an AUC of 0.964, 0.88, and 0.93 on KIT, NIH and Shenzhen datasets respectively [13].

Spatial Pyramid Pooling (SPP) was developed to offset the requirement of feeding fixed length data into CNN, enforced by the fully connected layer to perform classification. SPP sit before the convolutional layers and fully connected layer where they compute aggregations of data and pool features into a singular size consequently relegating the need to transform into a fixed size. This equips CNNs with the ability to take input of variable sizes. Beyond this ability, SPP equips CNN architectures with the ability to access convolved features of different scales that in turn help in image classification [14]. Substantial experiments with SPP were done by Zhu et al. [15], in their research they propose a text descriptor for scene text detection CNN model which is equipped with SPP. Their experiments were conducted on ICDAR 2011 and 2013 datasets without any cropping and warping to allow training of the model using different image scales. Their proposed descriptor model saw about 2% increase in F-Measure than other relevant studies [15]. Multi-scale SPP with DCNN has also been used in vehicle detection from high-resolution images. The dataset that has been trained is neither cropped nor warped nor stretched but rather features from images of varying scales are extracted adding to the robustness of the DCNN model [16]. DCNN with SPP achieved better than a normal DCNN as well as other conventional algorithms. When RR is given as 95%, the detection accuracy was 92.9% compared to a general DCNN which only achieves 80.5% and the FAR is at 19.8% [16]. Han et al. [17] use SPP with a pre-trained AlexNet on high spatial resolution (HSR) remote sensing image dataset. Regardless of SPP's ability to train on variable image scales, all the images from USGS National Map Urban Area Imagery collection, Google image dataset of SIRI-WHU, WHU-RS dataset, were resized to 227x227. The results show that pre-trained AlexNet with SPP achieved 95.95 ± 1.01% accuracy higher than a general pre-trained AlexNet [17].

In this paper, we equip three DCNN architectures namely: AlexNet, GoogLeNet and ResNet50 with SPP. For simplicity, the suffix 'SPP' will serve as an identifier for architectures using SPP.

3. PROPOSED METHOD

In this paper, we propose using Spatial Pyramid Pooling (SPP) to increase the performance of CNN in accurately classifying Tuberculosis CXR. Spatial Pyramid Pooling (SPP) is a methodology proposed by He et al. [14] in 2015 to eliminate the requirement of a fixed input image size [14]. Commonly, CNN architecture input layers are designed to allow a fixed input image size mainly because of the fully

connected layers. The convolutional layers in CNNs have the ability to generate feature maps of any image of any size. Fully connected layers require finite definition of the resulting feature sizes. It is with this in mind that CNN architectures are set with an input size constraint to allow correct calculation of features for classification. Currently, this approach has produced remarkable results all the same on different data sets like ImageNet, COCO dataset, CIFAR and many more. The limitation of this approach is that, when dealing with datasets that have varying sizes may result in loss of key features for classification. Ideally, when using SPP, the convolutional layers are not changed, they still retain their configure kernel sizes but they setup to accept input of different sizes. Since convolutional layers will still extract features regardless of the size, the features will be extracted from the images.

To satisfy the need of definite features size in fully connected layers, the last pooling layer or convolutional layer before fully connected layer is replaced with spatial pyramid pooling layer. Figure 1, shows an example implementation of AlexNet with SPP. Similarly, other models we have are designed as such. This layer, pools all extracted features of each filter in spatial bins of different sizes proportional to the input image size [18]. This process is synonymous with bag of words model, where features are grouped based on their filters from finer to coarser levels of the image. The ability to train models without pre-setting a fixed length, multi-level pooling windows allow for more robust features of the image to be pooled for object perception and finally, the ability to pool features at different scales are some of the advantages of equipping CNN with an SPP layer. For instance, the final layer of a CNN has 256 feature maps, the subsequent SPP layer of (1 x 2 x 4) spatial bins will pool each feature map based on the different bins. All models that have an implementation of SPP are suffixed with SPP, i.e., AlexNet-SPP. By leveraging on existing models that have been successful on different classification problems we used three main CNN architectures. AlexNet, GoogLeNet and ResNet50.

As previously, discussed SPP is applied between the final layer of the CNN architecture and the beginning of the fully connected layer. Figure 2 below is a representation of how SPP was applied in this paper.

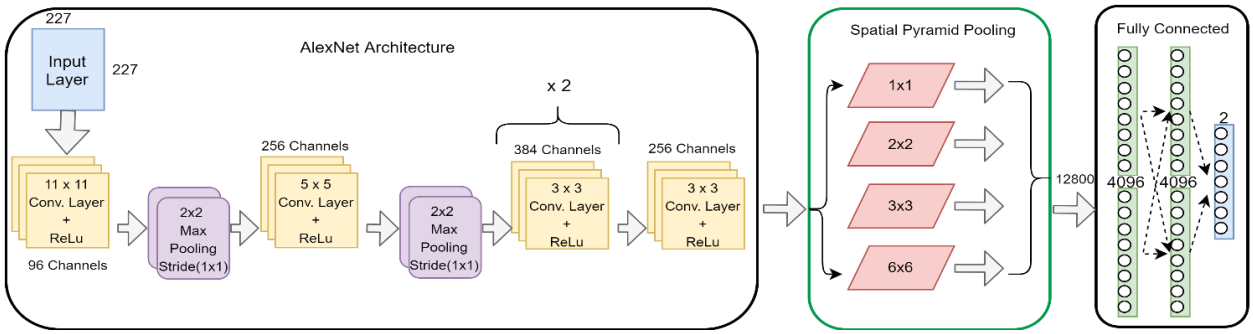


Figure 1. Implementation of SPP on CNN architecture (AlexNet)

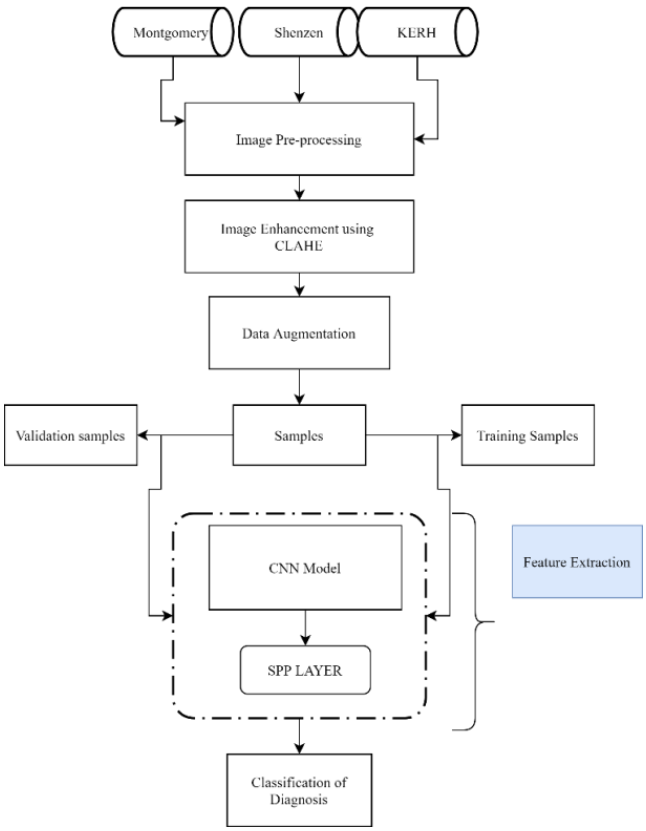


Figure 2. General overview of the methodology used

3.1 Image pre-processing

The same pre-processing pipeline was used on each separate dataset. For effective training, image pre-processing encompasses the methodology that involves shaping and transforming data into manageable patches for training. Our datasets comprise of medical volume data. All the datasets except KERH dataset were obtained in PNG format. For KERH dataset conversion from DICOM to PNG was required.

All the images go through several pre-processing techniques. The first is resizing to 256 x 256 dimensions. We randomly crop each of the 256 x 256 patches to 227 x 227 and 224 x 224. 227 x 227 patches were used on the AlexNet model and the other on the other models. The final step of the pre-processing step involves converting each of the 224 x 224 and 227 x 227 patches into greyscale. This is done deliberately to allow image enhancement.

3.2 Image enhancement

After pre-processing, all the images are enhanced using Contrast Limited Adaptive Histogram Equalisation Eq. (1). Contrast limited adaptive equalization is a modified part of adaptive histogram equalization. In this method enhancement function is applied over all neighbouring pixels and transformation function is derived. In this paper, the algorithm is applied on the foreground and limit the noise, enhance the contrast of the CXR images [19]. This process is the adjustment of intensity which is globally distributed across the

image. If we consider any greyscale image (x), n_i be the number of occurrences of grey level i and a probability function of occurrence of a pixel of level i in image (x) is:

$$p_x(i) = p\,(x = i) = \frac{n_i}{n}, 0 \leq i < L \qquad (1)$$

After enhancement of all the images, we finally convert each of the images back to RGB colour.

3.3 Data augmentation

Deep learning algorithms try to solve the problem of generalization given data. To be able to effectively generalize, they require a plethora of data presenting features related to different classes adequately [20]. Even if enough data is present, a model can easily overfit if it is given too many easy examples. In most case augmentation is performed for two reasons. The first being to compliment the model so it does not overfit. The second to multiply the dataset in cases where the data is not enough this is the result of the idea that adding more variant data to a deep learning model will improve the performance. Data augmentation involves going through geometric transformations like rotations, random cropping, random resizing, mirroring, changing colour, contrast, and sometimes even adding noise.

Data augmentation can be categorised into three main techniques; traditional image transformations (basic image manipulations), generative adversarial networks (GAN) and learning augmentations. Image manipulation augmentation, is the most basic type of augmentation that primarily involves changing the size, orientation or shape of the image. Flipping (horizontal or vertical), changing the colour channel, cropping which involves selecting a patch of a specific dimension from an image, rotations, translations are just some of the techniques under this augmentation techniques [21, 22]. GAN based augmentation, involve a model creating an artificial instance from the data whilst retain key characteristics (features). Implementation of GAN models have resulted in great success in classification models because of adversarial training that assist in accurately getting important features from the dataset. Finally, learning techniques for data augmentation involve feeding a network with two images from the same where a layer is returned with same size as a single image. The resulting layer is considered as separate image. Together with the original image are then used in another network for classification. In this study we only apply image manipulation augmentation techniques.

3.4 Feature extraction

Features are extracted automatically by leveraging popular architectures like AlexNet architecture that won the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2012), proposed by Krizhevsky et al. [23]. It consists of 5 convolutional layers that connect to the pooling layers before finally joining to 2 fully connected layers both have 4096 neurons and an output layer with Softmax activation. Local response normalization (LRN) is used to assist in model generalization and reduce the top-1 and top-5 error rate. AlexNet uses rectified linear unit (ReLU) in its convolutional layers this alone increases the speed of training compared to other activations functions like tanh [24]. Similarly, we use GoogLeNet and ResNet50 in the same way. GoogLeNet's architecture consists of 22 layers which are way higher than AlexNet, making it a very deep architecture. The fix input size of this architecture is 224 x 224 and uses ReLU activation to create non-linearity similar to AlexNet [25]. The architecture design aims to replicate how a human neurological system functions by finding the optimal local sparse structure of convolutional networks. Because of this, this architecture does not follow standard CNN design where convolutional layers followed by normalization and max-pooling rather a string of inception modules contribute to filter learning and dimensionality reduction. One of the key features of GoogLeNet is the introduction of *Inception Layers*, which reinforces the concept of sparsely connected architecture. An Inception layer is an amalgamation of several convolutional layers with different kernel sizes.

Residual Network architecture, aptly acronymized as ResNet, won the ILSVRC 2015. It is an ultra-deep neural network designed by He et al. [14] which aims to solve the vanishing gradient problem [26]. With 152 layers, this architecture was able to achieve 3.57% error on ImageNet test set and subsequently achieved a 28% improvement on COCO object detection dataset. When training deep networks, the accuracy of the model saturates and is then followed by a quick degradation. Thus, adding mode layers to network consequently results in higher training layer. He et al. [14] propose a method for training an ultra-deep neural network by introducing residual learning.

3.5 Network architectures

Famous architectures like AlexNet, GoogLeNet and ResNet50 have demonstrated remarkable results hence we use them together with SPP in this paper. The following are the modifications we have performed on each of the networks. Figure 1 shows how we implemented SPP on AlexNet as such it will not be discussed in this section of the paper. In the case of GoogLeNet, it uses Inception modules that reduce dimensionality by efficiently computing convolutions in deep networks. As such we anchor to this ability provided by GoogLeNet and apply SPP layer before classification. The same layer was implemented on ResNet50 which consists of residual identity blocks as shown in Figure 4. Figure 1, Table 1 and Table 2 show the resulting feature map after apply SPP layer on AlexNet and GoogLeNet respectively. Details of Inception block are given in Figure 3.

Table 1. GoogleNet (InceptionV1) architecture

Layer / Stride	Repeat	Output Size
Input		224 x 224 x 3
Conv(7x7)/2	1	112 x 112 x 64
MaxPool(3x3)/2	1	56 x 56 x 64
Conv(3x3)/1	1	28 x 28 x 192
MaxPool(3x3)/2	1	28 x 28 x 192
Inception1	1	28 x 28 x 256
Inception2	1	28 x 28 x 480
MaxPool (3x3)/2	1	14 x 14 x 480
Inception3	3	14 x 14 x 512
Inception4	1	14 x 14 x 528
Inception5	1	14 x 14 x 832
MaxPool(3x3)/2	1	7 x 7 x 832
Inception6	1	7 x 7 x 832
Inception7	1	7 x 7 x 1024
AveragePool(7x7)/1	1	1 x 1 x 1024
Dropout (40%)	1	1 x 1 x 1024
Spatial Pyramid Pool(1x2x3x6)	1	1 x 51200
Softmax	1	2

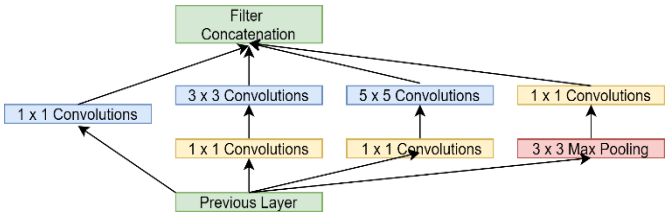


Figure 3. GoogLeNet's inception block representation

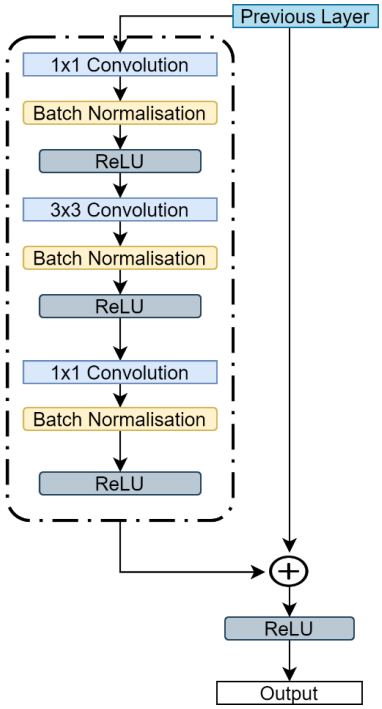


Figure 4. Residual identity block

Table 2. ResNet 50 architecture

Layer/Stride	Repeat	Output size
Input		224 x 224 x 3
Conv1(7x7)/2	1	112 x 112 x 64
IndentityBlock1	3	56 x 56 x 256
IndentityBlock2	4	28 x 28 x 512
IndentityBlock3	6	14 x 14 x 1024
IndentityBlock4	3	7 x 7 x 2048
Spatial Pyramid Pool(1x2x3x6)	1	1 x 102400
Softmax	1	2

4. EXPERIMENTATION

4.1 Datasets

Our experiments were done on three datasets one of which we collected ourselves from Konya Education and Research Hospital.

4.1.1 Public datasets

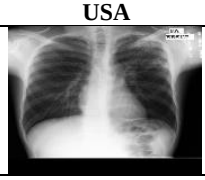
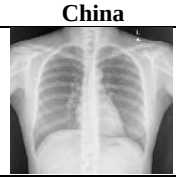
We evaluate the performance of using SPP on DCNN using three datasets. It is important to mention that the lack of large public datasets is a reason for stagnation on research of this kind. Two datasets can be accessed publicly; one being created by the National Library of Medicine, Maryland, USA. This dataset is a result of collaboration by the Department of Health and Human Services, Montgomery, County, Maryland, USA. It consists of a total of 138 images depicting the front chest area. 80 of the images are classified as normal cases and the remaining 58 cases are abnormal (in this case, CXRs with TB). Each CXR is exported into the PNG (Portable Network Graphics) as 12-bit grayscale images. The resolution of each image borders between 4,020×4,892 or 4,892×4,020 pixels [8, 9].

The second dataset is from People's Hospital, Guangdong Medical College, Shenzhen, China. The data from Shenzhen, China consists of 336 abnormal (with tuberculosis) and 326 normal CXRs. A Phillips DR Digital Diagnost system was the primary tool used to collect all the images over period of a month. This dataset also in the same PNG format, 12-bit grayscale as well as having a 3000 X 3000 pixels [8].

4.1.2 Konya Education Research Hospital (KERH) dataset

In additional to these two-public datasets, we use a dataset obtained from Konya Education Research Hospital, Konya, Turkey in collaboration with The Radiography Department. A Samsung DR Digital Diagnost system was used to collect all the images, each with a 4000X4000 pixels resolution. Samples were collected using an Image Archiving and Communication System called PACS (Picture Archiving Communication System) that enabled us to collect data in different formats including DICOM, JPG, BITMAP and PNG, we extracted all images in PNG format as well. Chest radiographs of new diagnosed ARB (Acid- Resistant Bacilli) positive (detected in sputum sample) tuberculosis patients who have not received any treatment before were retrospectively scanned. Collection of data was under the supervision of a trained Radiologist professional. This set has 206 normal and 159 abnormal CXRs. Samples of CXRs from all the datasets can be seen in Table 3. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was approved by the Clinical Research Ethics Committee of the Faculty of Medicine, Selcuk University, Konya, Turkey (No: 2019/240).

Table 3. Samples of CXR images from each dataset

Montgomery County, USA	Shenzhen, China	Konya, Turkey
		

4.2 Experiment

Three DCNN models were trained using three datasets described separately. Each architecture was fitted with a SPP layer between the final layer and fully connected layer. Thus, models trained using SPP will be annotated with a 'SPP' for easy identification in the results section. Naturally, given such lean datasets, data augmentation techniques can be used to reduce overfitting and make the model more robust against subtle changes in real world data. Consequently, automatic feature extraction of DCNN leaves very little data pre-processing to be done. For all the datasets, we first resized them to 256 x 256 then randomly cropped them to 227 x 227 and 224 x 224 dimensions for AlexNet, AlexNet-SPP and

GoogLeNet, GoogLeNet-SPP, ResNet50, ResNet50-SPP respectively. We also horizontally flip and randomly rotate image by 90, 180 270 degree angles. This is all done in-training to converse memory. Contrast Limited Adaptative Histogram Equalization (CLAHE) was used to amplify contrast of the CXRs.

The models were trained on an NVdia GeForce RTX 2070 with a dedicated memory of 8GB. Validation accuracy was used as an evaluation metric. All samples were split as 75%-25% for training and testing. The largest fraction goes to training, amounting to 103, 496, 274, and the remaining 35, 166, 91 were used for validation of the model for dataset 1, dataset 2 and dataset 3, respectively. We execute all models for 120 epochs with varying learning rate from 0.001 to 0.00001. We use a batch size of 8, with an SGD for global optimisation during networks back-propagation process.

Multi-level pooling windows were used in all of our models, we use about 50 bins which can be denoted as (1 x 1, 2 x 2, 3 x 3, 6 x 6) [14]. As observed by He et al. [14] multi-level enables robust pooling of different features from input images [14].

5. RESULTS

In this study, the primary focus is on the classification of Tuberculosis from CXRs using Spatial Pyramid Pooling (SPP). To achieve this, we fit three DCCN with an SPP layer instead of the commonly used Pooling layer in between the last convolutional layer and the start of the fully connected layer. In presenting our results we also explore and compare results from other methodologies that have contributed to similar research. Table 4 presents a concise summary of authors, year and accuracy achieved of relevant researches.

One of the earliest researches on CADx was in 2014 by Jaeger et al. [5] where a Graph-cut image segmentation method was first applied on the CXR. The resulting image was classified by Support Vector Machine (SVM) classifier. 0.74, 0.84 accuracy was achieved on Montgomery and Shenzhen [5]. Training Peruvian partners at "Socios en Salud", Partners in Health in Lima, Peru, a largely unbalanced dataset, with 453 normal CXR from 4701. Liu et al. [27] use fine-tune AlexNet and GoogLeNet (using ImageNet weights). AlexNet performed better than GoogLeNet with an accuracy of 0.85 [27]. To overcome, an unbalanced data problem, they perform shuffle sampling which significantly increased results on both models. Whilst on the same dataset, Cao et al. [28] present results with the aim of building a TB diagnostic for m-Heath (mobile health, with embodies solutions to assist health workers). They use a pre-trained GoogLeNet model, in their results the post accuracy of 0.89 after 100,000 iterations [12, 28]. Becker et al. [29] whilst utilising a commercially available deep learning software performed classification on 138 patient CXR. Cavity, consolidation, effusion, interstitial changes and normal examination are the classes in which the CXRs were grouped [29]. They achieved an accuracy of 0.82 on a lean dataset.

Further, a novel stacked generalisation CNN model is used by Rajaraman et al. [30] on the National Library of Medicine (NLM), National Institutes of Health (NIH) Dataset 1 and Dataset 2. The third data set was a private collection of CXRs, obtained with the assistance of Indiana University School of Medicine and Academic Model Providing Access to Healthcare (AMPATH) and a Kenyan NGO, and made available CXRs from rural western Kenya as a part of the mobile truck-based screening. They used segmentation to extract the Region of Interest (ROI). Their methodology yielded an accuracy of 0.875, 0.934, 0.733, 0.960 on Montgomery, Shenzhen, Kenyan, and Indian dataset respectively [30]. In Table 5, we explore the AUC results obtained from our experimentations. We perform comparisons between architectures with SPP and those not fitted with SPP. We also present results obtained from KERH.

Table 4. A summary of the different results from different research papers

Author	Method	Year	Dataset	Accuracy
Jaeger et al. [5]	SVM Classifier	2014	Montgomery	0.84
Lui et al. [27]	Pre-trained GoogLeNet	2017	Patners In Health Lima, Peru	0.85
Cao et al. [28]	Pre-trained GoogLeNet	2016	Patners In Health Lima, Peru	0.89
Hwang et al. [31]	Pre-trained AlexNet	2016	Montgomery, Shenzhen	0.88
Lopes et al. [32]	Pre-trained ResNet	2017	Montgomery, Shenzhen	0.83
Hooda et al. [33]	Custom 7-layer CNN	2017	Montgomery, Shenzhen	0.82

Table 5. Experimental results of models with SPP and without SPP on public datasets

Model	Dataset	Without SPP	With SPP
AlexNet		0.94	0.97
GoogLeNet	Montgomery	0.97	0.97
ResNet50		0.99	0.97
AlexNet		0.95	0.95
GoogLeNet	Shenzhen	0.97	0.98
ResNet50		0.95	0.96

Table 6. Results with SPP and without SPP on KERH dataset

Model	Dataset	Without SPP	With SPP
AlexNet		0.99	0.99
GoogLeNet	KERH	1.0	1.0
ResNet50		0.98	0.99

Table 7. Confusion matrix for three models with SPP and without SPP on Montgomery dataset

Montgomery		Predicated	
		Negative	Positive
AlexNet	Actual Negative	57	3
	Actual Positive	3	41
AlexNet-SPP	Actual Negative	58	2
	Actual Positive	1	43
GoogleNet	Actual Negative	58	2
	Actual Positive	3	43
GoogleNet-SPP	Actual Negative	57	3
	Actual Positive	0	44
ResNet50	Actual Negative	60	0
	Actual Positive	1	43
ResNet50-SPP	Actual Negative	59	1
	Actual Positive	2	42

Table 8. Confusion matrix for three models with SPP and without SPP on Shenzhen dataset

	Shenzhen		Predicated	
			Negative	Positive
Actual	AlexNet	Negative	234	8
		Positive	15	240
	AlexNet-SPP	Negative	226	16
		Positive	8	247
	GoogleNet	Negative	228	14
		Positive	2	253
	GoogleNet-SPP	Negative	238	4
		Positive	7	248
	ResNet50	Negative	229	13
		Positive	10	245
	ResNet50-SPP	Negative	232	10
		Positive	10	245

Table 9. Confusion Matrix for three models with SPP and without SPP on Shenzhen dataset

	Shenzhen		Predicated	
			Negative	Positive
Actual	AlexNet	Negative	234	8
		Positive	15	240
	AlexNet-SPP	Negative	226	16
		Positive	8	247
	GoogleNet	Negative	228	14
		Positive	2	253
	GoogleNet-SPP	Negative	238	4
		Positive	7	248
	ResNet50	Negative	229	13
		Positive	10	245
	ResNet50-SPP	Negative	232	10
		Positive	10	245

The validation accuracy results in Table 5 shows the performance of the models on publicly available datasets. Whilst Table 6. Shows validation accuracy results obtained from the KERH dataset. Further, Tables 7-10 are confusion matrices for untrained AlexNet, AlexNet-SPP GoogLeNet, GoogLeNet-SPP, ResNet50, and ResNet50-SPP on all of the datasets.

Further, Table 11-12, present the performance of models without SPP and those with SPP on the KERH dataset in terms of recall, precision, specificity and F measure. Given the following classification results in form of a confusion matrix, the recall, precision specificity and F score can be calculated using the Eq. (2) through Eq. (5) respectively.

Table 10. Confusion Matrix for three models with SPP and without SPP on KERH dataset

	KERH		Predicated	
			Negative	Positive
Actual	AlexNet	Negative	163	1
		Positive	3	106
	AlexNet-SPP	Negative	163	1
		Positive	3	106
	GoogleNet	Negative	164	0
		Positive	0	109
	GoogleNet-SPP	Negative	164	0
		Positive	0	109
	ResNet50	Negative	164	0
		Positive	6	103
	ResNet50-SPP	Negative	163	1
		Positive	3	106

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$F\ Score = \frac{(Precision * Recall)}{(Precision + Recall)} \tag{5}$$

where, TP, FP, TN, FN represent the true positive, false positive, true negative and false negative of a model's prediction respectively.

Receiver operating characteristic (ROC) curves are used to qualify the predictive ability of a binary classifier. Figure 5 illustrates the roc curves for all the models on all the given datasets. The table also displays area under the roc curve score. For further insight, the area under the curve of each model on each dataset was calculated presented in Figure 5. The performance in terms of AUC varies across the datasets, however, untrained GoogLeNet and GoogLeNet-SPP on average performs very well on all the datasets, GoogLeNet achieving 0.98, 0.93, 1.0 on Montgomery, Shenzhen and KERH respectively. The SPP version scores and AUC of 0.93, 0.98 and 1.0 almost similarly save for the Montgomery dataset.

Table 11. Recall, precision, specificity and F measure score of without SPP models on all datasets

Model	Dataset	Recall	Precision	Specificity	F-Score
AlexNet	Montgomery	0.93	0.93	0.95	0.93
GoogLeNet		0.97	0.95	0.96	0.96
ResNet50		0.97	1.0	1.0	0.98
AlexNet	Shenzhen	0.94	0.96	0.96	0.95
GoogLeNet		0.99	0.94	0.94	0.96
ResNet50		0.96	0.94	0.94	0.95
AlexNet	KERH	0.97	0.99	0.99	0.98
GoogLeNet		1.0	1.0	1.0	1.0
ResNet50		0.94	1.0	1.0	0.97

Table 12. Recall, precision, specificity and F measure score of SPP models on all datasets

Model	Dataset	Recall	Precision	Specificity	F-Score
AlexNet-SPP		0.97	0.95	0.96	0.96
GoogLeNet-SPP	Montgomery	1.0	0.93	0.95	0.96
ResNet50-SPP		0.95	0.97	0.98	0.96
AlexNet-SPP		0.96	0.93	0.93	0.95
GoogLeNet-SPP	Shenzhen	0.97	0.98	0.98	0.97
ResNet50-SPP		0.96	0.96	0.95	0.96
AlexNet-SPP	KERH	0.97	0.99	0.99	0.98
GoogLeNet-SPP		1.0	1.0	1.0	1.0
ResNet50-SPP		0.97	0.99	0.99	0.98

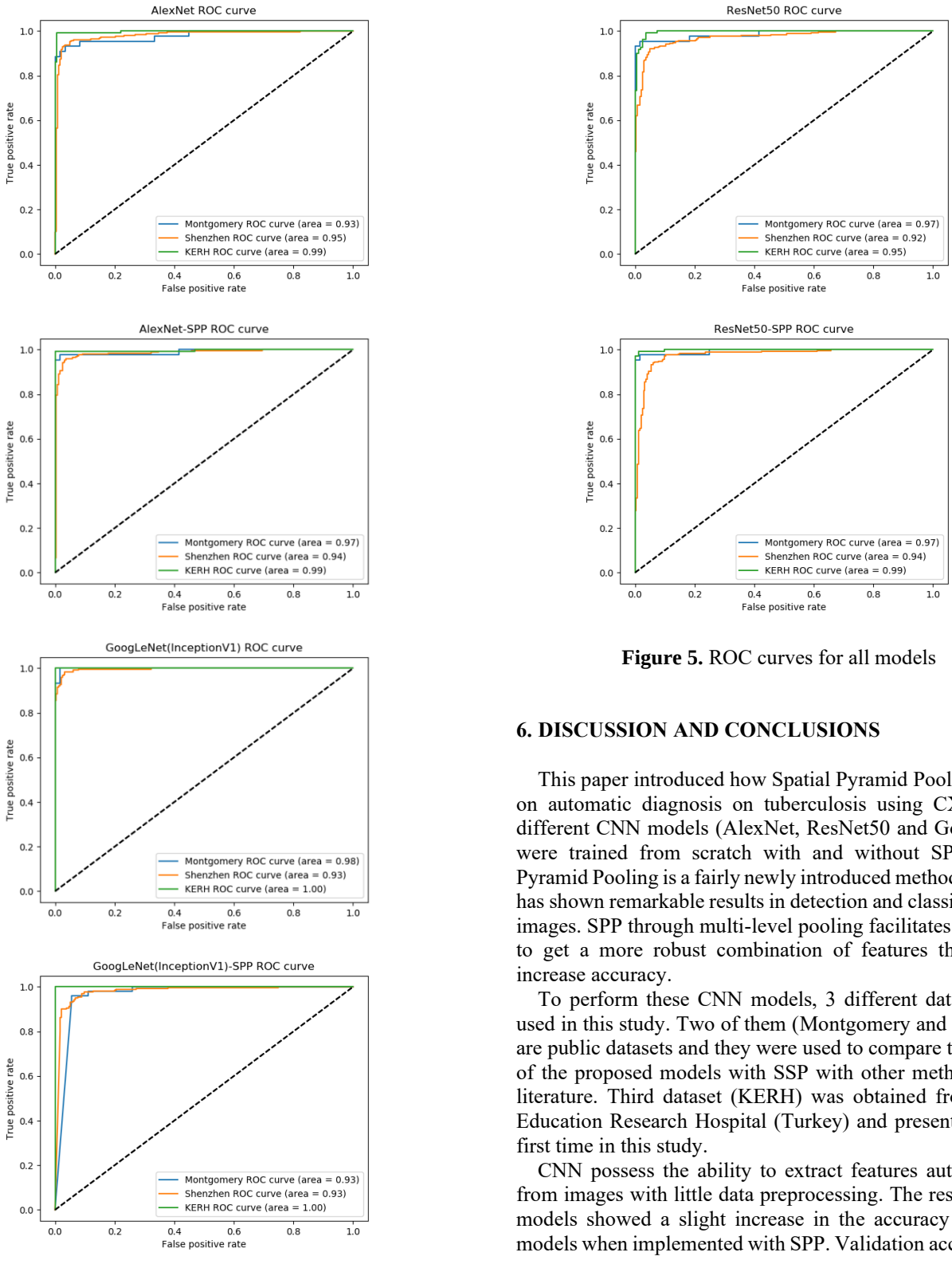


Figure 5. ROC curves for all models

6. DISCUSSION AND CONCLUSIONS

This paper introduced how Spatial Pyramid Pooling effects on automatic diagnosis on tuberculosis using CXR. Three different CNN models (AlexNet, ResNet50 and GoogLeNet) were trained from scratch with and without SPP. Spatial Pyramid Pooling is a fairly newly introduced methodology and has shown remarkable results in detection and classification of images. SPP through multi-level pooling facilitates the ability to get a more robust combination of features that in turn increase accuracy.

To perform these CNN models, 3 different datasets were used in this study. Two of them (Montgomery and Shenzhen) are public datasets and they were used to compare the success of the proposed models with SSP with other methods in the literature. Third dataset (KERH) was obtained from Konya Education Research Hospital (Turkey) and presented for the first time in this study.

CNN possess the ability to extract features automatically from images with little data preprocessing. The results of our models showed a slight increase in the accuracy across all models when implemented with SPP. Validation accuracy was

used as a metric to determine this improvement. Both the validation and training images set artificially increased the data by randomly cropping the all-original images 3 times. Untrained AlexNet-SPP achieved a significant increase or achieved the same validation accuracy on Montgomery, Shenzhen, and KERH dataset. Compared to other methodologies we have examined in this paper, AlexNet-SPP has either matched or slightly outperformed the other models. Untrained GoogLeNet and ResNet50 trained on Montgomery confirming the hypothesis that even on a dataset trained with same size SPP increases the accuracy. Due to a small dataset, we can also note the model significantly overfit. To reduce this, we performed several augmentation methods as regularisation as well dropout except on ResNet50 was used. Training results of all the models on KERH's dataset performed better in contrast to the results of models on our two other public datasets. AlexNet achieves a remarkable 0.94 without SPP and 0.95 with SPP. ResNet50 performs similarly to AlexNet with 0.93 without SPP and 0.94 with SPP. The most outstanding result is achieved on untrained GoogLeNet and GoogLeNet-SPP with 0.97 and 0.98 validation accuracy respectively.

As observed in this paper, all the models have seen an accuracy improvement from the base untrained models. As a result, we conclude that CNNs can be equipped with SPP to train better thus creating robust models to assist in the diagnosis of tuberculosis. For future improvements, we plan on using fine-tuning with SPP on a bigger dataset as it has been observed that DCNN train better on large datasets. We aim to also incorporate different activation methods i.e., using Mish and perform comparisons with ReLU.

ACKNOWLEDGMENT

This work was financially supported by the Coordinators of Scientific Research Projects of Konya Technical University (P.N.: 191013018).

REFERENCES

[1] World Health, Global tuberculosis report 2018. Geneva: World Health Organization (in en), 2018.

[2] McBee, M.P., Awan, O.A., Colucci, A.T., Ghobadi, C.W., Kadom, N., Kansagra, A.P., Tridandapani, S., Auffermann, W.F. (2018). Deep learning in radiology. Academic Radiology, 25(11): 1472-1480. https://doi.org/10.1016/j.acra.2018.02.018

[3] McCoppin, R., Rizki, M. (2014). Deep learning for image classification. Proc. SPIE 9079, Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR V, p. 90790T. https://doi.org/10.1117/12.2054045

[4] Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E. (2017). A survey of deep neural network architectures and their applications. Neurocomputing, 234: 11-26. https://doi.org/10.1016/j.neucom.2016.12.038

[5] Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R.K., Antani, S., Thoma, G., Wang, Y., Lu, P., McDonald, C.J. (2014). Automatic tuberculosis screening using chest radiographs. IEEE Transactions on Medical Imaging, 33(2): 233-245. https://doi.org/10.1109/TMI.2013.2284099

[6] Melendez, J., Sánchez, C.I., Philipsen, R.H.H.M., Maduskar, P., Dawson, R., Theron, G., Dheda, K., van Ginneken, B. (2016). An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information. Scientific Reports, 6(1): 25265. https://doi.org/10.1038/srep25265

[7] Maduskar, P., Muyoyeta, M., Ayles, H., Hogeweg, L., Peters-Bax, L., van Ginneken, B. (2013). Detection of tuberculosis using digital chest radiography: Automated reading vs. interpretation by clinical officers. The International Journal of Tuberculosis and Lung Disease, 17(12): 1613-1620. https://doi.org/10.5588/ijtld.13.0325

[8] Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G. (2014). Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quantitative Imaging in Medicine and Surgery, 4(6): 475-477. https://doi.org/10.3978/j.issn.2223-4292.2014.11.20

[9] Candemir, S., Jaeger, S., Palaniappan, K., Musco, J.P., Singh, R.K., Xue, Z., Karargyris, A., Antani, S., Thoma, G., McDonald, C.J. (2014). Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Transactions on Medical Imaging, 33(2): 577-590. https://doi.org/10.1109/TMI.2013.2290491

[10] Panicker, R.O., Kalmady, K.S., Rajan, J., Sabu, M.K. (2018). Automatic detection of tuberculosis bacilli from microscopic sputum smear images using deep learning methods. Biocybernetics and Biomedical Engineering, 38(3): 691-699. https://doi.org/10.1016/j.bbe.2018.05.007

[11] Lakhani, P., Sundaram, B. (2017). Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology, 284(2): 574-582. https://doi.org/10.1148/radiol.2017162326

[12] Alcantara, M.F., Cao, Y., Liu, C., Liu, B., Brunette, M., Zhang, N., Sun, T., Zhang, P., Chen, Q., Li, Y., Albarracin, C.M., Peinado, J., Garavito, E.S., Garcia, L.L., Curioso, W.H. (2017). Improving tuberculosis diagnostics using deep learning and mobile health technologies among resource-poor communities in Perú. Smart Health, 1-2: 66-76. https://doi.org/10.1016/j.smhl.2017.04.003

[13] Hwang, S., Kim, H.E., Jeong, J., Kim, H.J. (2016). A novel approach for tuberculosis screening based on deep convolutional neural networks. Proc. SPIE 9785, Medical Imaging 2016: Computer-Aided Diagnosis. https://doi.org/10.1117/12.2216198

[14] He, K., Zhang, X., Ren, S., Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

[15] Zhu, R., Mao, X., Zhu, Q., Li, N., Yang, Y. (2016). Text detection based on convolutional neural networks with spatial pyramid pooling. 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, pp. 1032-1036. https://doi.org/10.1109/ICIP.2016.7532514

[16] Qu, T., Zhang, Q., Sun, S. (2017). Vehicle detection from high-resolution aerial images using spatial pyramid pooling-based deep convolutional neural networks.

Multimedia Tools and Applications, 76(20): 21651-21663. https://doi.org/10.1007/s11042-016-4043-5

[17] Han, X., Zhong, Y., Cao, L., Zhang, L. (2017). Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. Remote Sensing, 9(8): 848. https://doi.org/10.3390/rs9080848

[18] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T. (2017). Recent advances in convolutional neural networks. Pattern Recognition, 77: 354-377. https://doi.org/10.1016/j.patcog.2017.10.013

[19] Yadav, G., Maheshwari, S., Agarwal, A. (2014). Contrast limited adaptive histogram equalization based enhancement for real time video system. 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), New Delhi, pp. 2392-2397. https://doi.org/10.1109/ICACCI.2014.6968381

[20] Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D. (2016). Understanding data augmentation for classification: when to warp? arXiv preprint arXiv:1609.08764.

[21] Perez, L.,Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

[22] Shorten, C., Khoshgoftaar, T.M. (2019). A survey on Image data augmentation for deep learning. Journal of Big Data, 6(1): 60. https://doi.org/10.1186/s40537-019-0197-0

[23] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, pp. 1097-1105. https://doi.org/10.1145/3065386

[24] Nair, V., Hinton, G.E. (2010). Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807-814.

[25] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, pp. 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[26] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[27] Liu, C., Cao, Y., Alcantara, M., Liu, B., Brunette, M., Peinado, J., Curioso, W. (2017). TX-CNN: Detecting tuberculosis in chest X-ray images using convolutional neural network. 2017 IEEE International Conference on Image Processing (ICIP), Beijing, pp. 2314-2318. https://doi.org/10.1109/ICIP.2017.8296695

[28] Cao, Y., Liu, C., Liu, B., Brunette, M.J., Zhang, N., Sun, T., Zhang, P., Peinado, J., Garavito, E.S., Garcia, L.L., Curioso, W.H. (2016). Improving tuberculosis diagnostics using deep learning and mobile health technologies among resource-poor and marginalized communities. 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), Washington, DC, pp. 274-281. https://doi.org/10.1109/CHASE.2016.18

[29] Becker, A.S., Blüthgen, C., Phi van, V.D., Sekaggya-Wiltshire, C., Castelnuovo, B., Kambugu, A., Fehr, J., Frauenfelder, T. (2018). Detection of tuberculosis patterns in digital photographs of chest X-ray images using deep learning: Feasibility study. The international journal of tuberculosis and lung disease: The official journal of the International Union against Tuberculosis and Lung Disease, 22(3): 328-335. https://doi.org/10.5588/ijtld.17.0520

[30] Rajaraman, S., Candemir, S., Xue, Z., Alderson, P.O., Kohli, M., Abuya, J., Thoma, G.R., Antani, S. (2018). A novel stacked generalization of models for improved TB detection in chest radiographs. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, pp. 718-721. https://doi.org/10.1109/EMBC.2018.8512337

[31] Hwang, S., Kim, H.E., Jeong, J., Kim, H.J. (2016). A novel approach for tuberculosis screening based on deep convolutional neural networks. Medical Imaging 2016: Computer-Aided Diagnosis, 9785: International Society for Optics and Photonics, p. 97852W. https://doi.org/10.1117/12.2216198

[32] Lopes, U.K., Valiati, J.F. (2017). Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. Computers in Biology and Medicine, 89: 135-143. https://doi.org/10.1016/j.compbiomed.2017.08.001

[33] Hooda, R., Sofat, S., Kaur, S., Mittal, A., Meriaudeau, F. (2017). Deep-learning: A potential method for tuberculosis detection using chest radiography. 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuching, pp. 497-502. https://doi.org/10.1109/ICSIPA.2017.8120663