Original Article

Automated detection of *Mycobacterium tuberculosi*s using transfer learning

Abdullahi Umar Ibrahim[1], Emrah Guler[2], Meryem Guvenir[2], Kaya Suer[3], Sertan Serte[4], Mehmet Ozsoz[1]

[1] *Department of Biomedical Engineering, Near East University, Nicosia, Turkey*
[2] *Department of Medical Microbiology and Clinical Microbiology, Near East University, Nicosia, Turkey*
[3] *Department of Infectious Disease and Clinical Microbiology, Near East University, Nicosia, Turkey*
[4] *Department of Electrical Electronics Engineering, Near East University, Nicosia, Turkey*

Abstract

Introduction: Quantitative analysis of *Mycobacterium tuberculosis* using microscope is very critical for diagnosing tuberculosis diseases. Microbiologist encounter several challenges which can lead to misdiagnosis. However, there are 3 main challenges: (1) The size of *Mycobacterium tuberculosis* is very small and difficult to identify as a result of low contrast background, heterogenous shape, irregular appearance and faint boundaries (2) *Mycobacterium tuberculosis* overlapped with each other making it difficult to conduct accurate diagnosis (3) Large amount of slide can be time consuming and tedious to microbiologist and which can lead to misinterpretations.
Methodology: To solve these challenges and limitations, we proposed an automated-based detection method using pretrained AlexNet to trained the model in 3 sets of experiments A, B and C and adjust the protocols accordingly. We compared the detection of tuberculosis using AlexNet Models with the ground truth result provided by microbiologist and analyzed inconsistencies between network models and human.
Results: 98.15 % accuracy, 96.77% sensitivity and 100% specificity for experiment A, 98.09% accuracy, 98.59% sensitivity and 97.67% specificity for experiment B and 98.73% testing accuracy, 98.59 sensitivity, 98.84% specificity ofr experiment C which sound robust and promising.
Conclusions: The results indicated that network performance was successful with high accuracies, sensitivities and specificities and it can be used to support microbiologist for diagnosis of tuberculosis.

Introduction

Tuberculosis is an airborne disease caused by a bacterium known as *Mycobacterium tuberculosis* which are rod shape microbes with length ranging from 1-10 μm. The disease can be transmitted from infected person to another through sneezing, coughing, spitting and speaking. The most common type of tuberculosis is pulmonary tuberculosis which attacks the lungs. The other type is known as extrapulmonary tuberculosis which cause damage to other organs such as Kidney, brain and spine. However, the disease can lead to mortality if not treated. Another classification of tuberculosis is based on active and dormant (latent). Active type is contagious and can be easily transmitted while latent type is not contagious but can become active if not treated with proper medications [1]

According to World Health Organization (WHO) [2], deaths as a result of tuberculosis-related disease decrease from 1.6 million in 2017 to 1.5 million in 2018. It was estimated that 10 million people fall ill as a result of tuberculosis in 2018 with the majority of patients coming from India, Pakistan and china. There are many approached adopted by pathologist for detection of tuberculosis, some of the techniques include Tuberculin Skin Test (TST), microscopy, chest X-ray, GeneXpert, culture test and interferon γ-release assay. However, among these techniques, microscopic sputum smear evaluation using microscope remain the most common approach worldwide especially in underdeveloped and middle-income countries due to its affordability, simplicity, speed and maintenance compare to other techniques [3].

Deep learning

Deep learning is a subfield of machine learning which is inspired by how human brain's function, due to connections or synopsis of nerve cells or neurons. Model learn as a result of data connection between neurons in the network. A simple neural network is called a perceptron which take input as data set and

produced an output as classification category or prediction outcome. Deep learning neural networks are made of multiple perceptron's with an input layer, and many hidden layers before output layer [4,5]. Since the emergence of deep learning in 2010, scientist have designed different models using Convolutional Neural Networks (CNN) that can classify and analyze medical images such as cancer, tuberculosis, radiological images for diagnosis of diseases [6,7].

There are various neural networks architectures developed. Some of the architectures have performed better than others in terms of regression, classification and denoising images. The current best models include AlexNet with 8 layers, VGGNet with 19 and 16 layers, Inception module also known as GoogleNet with 22 layers and 9 modules and Residual or ResNet with 152 layers [8]. To train a neural network, a backpropagation algorithm is used to adjust the weight according to the data pattern and optimize the error between predicted output and actual output [9].

The principle behind the application of CNN in classification or regression revolves around series of dot products of weight matrices and input matrix. These processes are categorized into two stages known as feature learning and classification. Future learning is based on the use of convolutional blocks with operations such as convolution which is a process of computing input matrix and feature matrix to obtain a convolve map or feature map, activation operation is the use of activation function such as tanh, sigmoid and Rectified Linear Unit (ReLu) to squash output into zero or within ranges of 0 and 1 or from -1 to 1, pooling operation on the other hand, is employ to reduce computation by taking the most important part of the convolve map by either max pooling or average (mean) pooling. The output is obtained after these operations in all the layers (including fully connected layers or global average pooling layers) which is pass to a classifier such as SoftMax (based on probabilities) to categorized output [10,11].

Related work

Throughout last decade, scientist have been trying to integrate application of artificial intelligence, machine learning and deep learning in healthcare system. Researchers have utilized CNN to solve challenges in medicine such as detection using classification and segmentation approach for skin disease, brain cancer, breast cancer, diabetes (retinopathy) [12,13]. In the field of microbiology, Microbiologist, Radiologist and computer scientist have been working together to detect microbial diseases such as tuberculosis, malaria and pneumonia using computer aided diagnosis [14].

Mycobacterium tuberculosis are bacterial specie that causes pulmonary tuberculosis, in order to identify different species of bacteria quickly and accurately, Smith *et al*. [15] employed automated microscope which is designed to collect high resolution image dataset from microscopic slides. The study also collected blood samples from different patients and incubate the samples in order to amplified the bacterial colonies stained using dye. 25,000 images were obtained and further augmented to generate more dataset. For categorization a CNN is trained and tested to classify round shape, round chains and rod shape bacteria with an accuracy of 95%.

Xiong *et al*. [16] collected samples from 246 patients and carried out acid fast staining with 45 cases as training data set and 201 as testing dataset. The image is scanned with KF-PRO-005 digital machine and undergoes data augmentation using rotation, mirroring, position shifting to generate 96,530 small patches with 40X magnification. CNN pretrained on CIFAR-10 is used to classify the 32x32 images with 16 strides to achieved 97.94% sensitivity and 83.65% specificity. Khan *et al*. [17] developed an approached to predict tuberculosis using artificial neural network. 12,636 datasets are divided into training (70%) and testing (30%) where the network is trained using feedforward backpropagation neural network resulting in learning rate of 0.70 with 94% overall accuracy.

Costa *et al*. [18] developed an automatic detection of *Mycobacterium tuberculosis* using Support Vector Machine (SVM) classifier. The research utilized 12 cases to obtained 120 smear microscopic slide images which undergoes bacillus segmentation and conventional smear microscopy. The study reported an overall sensitivity of 96.80% and error rate of 3.38%. El-melegy *et al*. [19] employed deep learning approach to identified tuberculosis in ZN-stained sputum smear images. The research utilized Faster Region-based convolutional neural network F-R-CNN and Faster Region-based CNN plus CNN (F-R-CNN+CNN) with 500 images (80% for training and 20% for testing). F-R-CNN achieved 98.3% accuracy and 82.6% sensitivity while F-R-CNN+CNN achieved 98.4% and 85.1% sensitivity.

Aguiar *et al*. [20] developed an artificial neural network model to classify pulmonary tuberculosis, the network is made up of multi-perceptron fed up with clinical data (which includes demographics and respiratory symptoms), radiological data and pulmonary tuberculosis samples obtained in the form of

sputum and bronchoalveolar lavage. The total samples contained 315 presumed tuberculosis cases with 80% training and 20% for testing. The model achieved 96% sensitivity and 89% specificity. Yahiaoui *et al.* [24] utilized SVM to diagnose pulmonary tuberculosis. 150 chest X-ray images were collected with 50 patients diagnosed with tuberculosis (positive) and 100 (negative). The model achieved 96.68% accuracy.

Aims

Tuberculosis infection have been a major challenge for medical and healthcare sectors in many underdeveloped countries and remote communities with limited diagnosis tools and treatment approach. Interpretation of microscopic slide image of tuberculosis obtained from patients is sometimes tedious for qualified microbiologist. However, the development of fast, cheap, simple and accurate detection approach for diagnosis and predictions of these diseases are highly required.

Our contributions are enumerated below:
1. We utilized pretrained AlexNet and trained the model in 3 sets of experiments A, B and C and adjust the protocols accordingly;
2. We carried out 10k fold cross validation;
3. We compare the detection of tuberculosis using AlexNet Models with the ground truth result provided by microbiologist and analyzed inconsistencies between network models and human.

Methodology

Data collection of tuberculosis and non-tuberculosis patients

One hundred positive acid-fast stained (Ziehl-Nielsen) tuberculosis cases were obtained from Istanbul

Figure 1. Microscopic slide images. A) with purple background; B) with pink background.
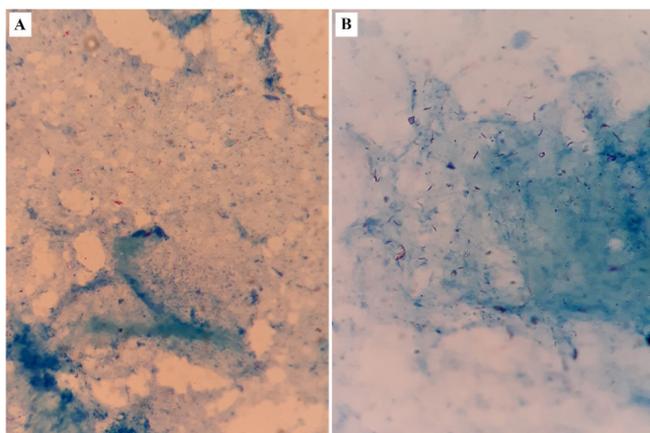
Tuberculosis Control Association (TCA). In addition, positive cases prepared using acid-fast stained method from Near East University hospital microbiology department were also included. Negative samples were prepared by specialist using the same method obtained from people without suspicion of TB infection. However, multiple images were obtained from the cases.

Image acquisition and Model training using pretrained AlexNet

In all the experiments (A, B, and C), datasets are divided into 70% for training and 30% for testing. A total of 178 images validated and confirmed by microbiologists were used for experiment A of which 124 images were selected as training dataset and 54 images were selected as testing dataset. For experiment B, 524 images were validated and confirmed by microbiologists, of which 371 images were selected as training dataset and 157 images were selected as testing dataset. For experiment C, 530 images were validated and confirmed by microbiologists, of which 367 images were selected as training dataset and 159 images were selected as testing dataset. The images acquired from the microscopes are shown in Figure 1.

Data augmentation

Experiment A

The training images (124) were augmented by rotation and cropping to obtained a total dataset of 1050 images of which 526 are negative and 524 are positive. Originally, the slides images are below 4416×3312 pixels but are further reduced to 227×227×3 pixels to fit into pretrained AlexNet model. Since the images are 3 channels (RGB) there is no need for conversion.

Experiment B

The training images (367) were augmented by rotation and cropping to obtained a total dataset of 2444 images of which 1132 are negative and 1312 are positive. Originally, the slides images are below 3456×4608 pixels but are further reduced to 227×227×3 pixels to fit into pretrained AlexNet model.

Experiment C

The training images (371) were augmented by rotation and cropping to obtained a total dataset of 2444 images of which 1144 are negative and 1320 are positive. Originally, the slides images are above 3456×4608 but are further reduced to 227×227×3 pixels to fit into pretrained AlexNet model. The description of images for experiment A, B and C are shown in Table 1.

Model Training

MATLAB is employed to train the Neural Network. Windows 64-bit personal computer is used with an intel ® Core i7-3537U and 8GB random access memory with graphical Processing unit (GPU) present in the RAM. The 30% saved slides test images are used to evaluate the performance of the network.

This research utilized AlexNet as the pretrained model due to it high performance in carrying out feature extraction. The architecture of the deep neural network use to classify Mycobacterium tuberculosis into positive (presence of Mycobacterium tuberculosis) and negative (absence of Mycobacterium tuberculosis) has 8 layers with series of operations such as CONV, FM and BN. The term "CONV" stands for convolution, Conv 1-5 for convolutional layers 1-5. "FM" stand for feature map, "BN" for Batch Normalization and "FC" for Fully Connected layers (FC1-FC3), the output layer utilized SoftMax activation function for classification. The network employed convolutional filters (matrix) of size 3×3 without any padding. Max pooling is carried out using a 2×2 window size. Minibatch optimization is used to optimize the model using gradient descent. 0.0001 is used for training learning rate with 20 epochs. The code used for this study is presented in the Annex to this article.

Cross Validation

Cross validation is a vital method used in machine learning for parameter selection and evaluation of learning performance and prediction. In this study, we utilized K-fold cross validation approach where the datasets are split into K sets of equal size (i.e. K = 10). In each K sets K-1 is used as training dataset and 1 set is used as validation dataset. Training of the dataset is repeated for K number of times (i.e. n = K) [21]. The average performance of the training and testing dataset is computed as the evaluation index for the models. This approach is very efficient especially when there are limited number of samples as it takes advantage of the whole dataset [22].

Evaluation and Confusion Matrix

To evaluate the performance of the trained models, three parameters are employed; accuracy, sensitivity and specificity. Accuracy is termed as the ratio of correctly classified images over total number of images; it is also termed as the sum of sensitivity and specificity. For evaluating the accuracy and loss of a model the following formulas are utilized:

$$Loss = \frac{-1}{n}\sum_{i=1}^{n} logPC$$

$$Accuracy = \frac{-C}{N}$$

Where N is the overall number of images during training and testing, n is the number of images and PC is the probability of the correctly classified images.

Confusion matrix is the common approach used for evaluation of model performance based on true positive (TP), true negative (TN), false positive (FP) and false negative (FN). TPs is the number of samples that are correctly identified by the model as positive cases or number of cases who actually have tuberculosis according to each model. TNs is the number of samples that are correctly identified by the model as negative cases or number of cases who are actually healthy (normal) and classified as negative according to each model. FPs are the number of samples that are incorrectly classified as negative by the model or number of cases that are actually negative (normal or healthy) but classified as tuberculosis according to each model. FNs are the number of samples that are incorrectly classified as positive by the model or number of cases that are actually positive (tuberculosis) but classified as normal or healthy according to each model as shown in Table 2.

Table 1. Experiment A, B and C Dataset for training image after augmentation.

Label	Training		
	Experiment A	Experiment B	Experiment C
Positive	526	1,312	1,320
Negative	524	1,132	1,144
Total	1,050	2,444	2,464

Table 2. Confusion matrix.

Predicted	Actual	
	True Positive (+)	False Negative (-)
True Positive	True +	False +
False Negative	False -	True -

Table 3. Model Learning and Testing Parameters for Experiment A, B and C.

Learning parameters	Experiment A	Experiment B	Experiment C
	Values	Values	Values
Training ratio (%)	70	70	70
Learning rates	0.0001	0.0001	0.0001
N of epochs	20	20	20
Training accuracy (%)	99.68	100	99.19
Testing accuracy (%)	98.15	98.09	98.73
Achieved error rate	0.0185	0.0185	0.0127
Sensitivity (%)	96.77	98.59	98.59
Specificity (%)	100	97.67	98.84

Table 4. Cross validation result for Tuberculosis.

K Fold	Training accuracy	Validation	Testing accuracy	Sensitivity	Specificity
1	98.15	0.9815	98.09	0.9677	1.000
2	99.07	0.9907	97.45	0.9852	0.9963
3	99.81	0.9981	97.45	1.0000	0.9963
4	99.07	0.9907	98.73	1.0000	0.9815
5	98.89	0.9889	98.09	0.9889	0.9889
6	100.00	1.0000	100	1.0000	1.0000
7	99.44	0.9944	98.64	1.0000	0.9868
8	99.44	0.9944	98.73	0. 9963	0.9926
9	99.63	0.9963	97.13	0.9963	0.9963
10	99.26	0.9926	98.54	0.9859	1.0000
Average	9.9276/10	9.99276/10	9.8285/10	9.92303/10	9.9387/10
	99.28	0.9928	98.29	0.9920	0.9939

Table 5. Diagnosis accuracy: CNN and microbiologist.

Participants (N = 50)	Sensitivity (%)	Specificity (%)	Accuracy (%)
Machine (**Model in Ex B**)	93.94	86.36	90.15
Beginner 1	88.57	100	94.29
Beginner 2	86.11	100	93.06
Certified 1	96.88	100	98.44
Certified 2	100	100	100
Participants (N = 30)	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**
Machine (**Model in Ex C**)	100	100	100
Beginner 1	100	85.71	92.86
Beginner 2	94.74	92.31	96.15
Certified 1	94.74	100	97.37
Certified 2	100	92.31	96.15

Ex: Experiment.

Table 6. Results of Proposed approach Vs other approaches.

Reference	Model	Accuracy (%)	Sensitivity (%)
[17]	CNN	94.0	-
[16]	CNN	97.0	83.6
[15]	CNN	95.0	-
[18]	SVM+CNN	96.8	-
[19]	FRCNN	98.3	82.6
	FRCNN+CNN	98.4	85.1
Experiment A	Pre-trained AlexNet	98.15	96.77
Experiment B	Pre-trained AlexNet	98.09	98.59
Experiment C	Pre-trained AlexNet	98.73	98.59

True positive rate (sensitivity) is the proportion of positive image samples that are correctly identified as positive sample (i.e. it shows the percentage of positive samples that are correctly identified as positives).

$$Sensitivity \frac{TPs}{TPs + FNs}$$

False positive rate (FPR) also known as specificity is the proportion of positive samples that are incorrectly identified as positive samples (i.e. it shows the percentage of negative samples that are incorrectly identified as positives).

$$Specificity \frac{TNs}{TNs + FPs}$$

Machine vs Real and Human

To check the application of the models on new samples (microscopic stained slides of new patients), unknown images are tested using the models and compared with the ground truth (real) result provided by Microbiologist. The same result (machine) is also use to compared between microbiologist with less than one year working experience (i.e. beginners) and certified microbiologist.

Results
AlexNet model

The network was able to learned different features from the convolutional layers. In convolutional layer 1, the neurons are very active in spotting features during training while the final convolutional layer (Conv 5) of the AlexNet consists of neurons that are very active in detecting different levels and descriptive features. Both first and fifth convolutional layers learn features differently which contribute to the learning efficiency of the model. The final layer in the network (fully connected layer 8) employed SVM classifier to

Figure 2. Classification of microscopic slide images using AlexNet model.

distinguish between negative and positive images. The parameters of the models are shown in Table 3.

Iteration is defined as the number of batches needed to complete a single epoch. For experiment A, the model reached 1460 maximum iterations. The training of our network resulted in 20 epochs which is obtained when number of iterations is equal to total number of images. Each of the completed epoch represents number of iterations multiplied by batch size divided by total number of images during training. Iterations by epoch is calculated by dividing 1460 by 20 which result into 73 as total iterations per epoch. After undergoing training, 30% of the dataset are used to test the network performance and the ability to discriminate between positive and negative slides as shown in Figure 2. The model achieved a learning rate of 0.0001 with 20 number of epochs. 99.68% training accuracy is achieved with 98.15% testing accuracy, 96.77% sensitivity, 100 specificity and 0.0185 error rate.

For Experiment B, the model reached 3420 maximum iterations with 20 epochs resulting to 171 total iterations per epoch. The model achieved a learning rates of 0.0001 with 20 number of epochs. 99.68% training accuracy is achieved with 98.09% testing accuracy, 98.59 sensitivity, 97.67% specificity and 0.0191 error rate.

For Experiment C, the model reached 3440 maximum iterations with 20 epochs. resulting to 171 total iterations per epoch. The model achieved a learning rates of 0.0001 with 20 number of epochs. 99.19% training accuracy is achieved with 98.73% testing accuracy, 98.59 sensitivity, 98.84% specificity and 0.0127 error rate. The overall description of the results obtained in experiment C is illustrated in Figure 2.

Cross validation

For general dataset the model achieved 98.83% training accuracy, 97.64% testing accuracy, 98.67% sensitivity and 100% specificity, while for cross validation model achieved an average performance of 99.28% training accuracy, 98.29% testing accuracy, 99.20% sensitivity and 99.39% specificity as shown in Table 4. This shows that the average performance of cross validation achieved a higher training accuracy, testing accuracy and sensitivity than general dataset. General dataset (model for experiment C) only outperforms mean of cross validation in specificity.

Machine vs Human Result

To compare the model of experiment B with microbiologist (professionals and beginners), 50
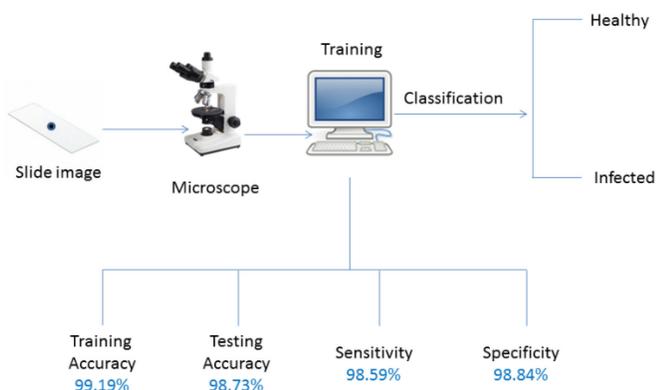
unknown images are utilized. However, microbiologist outperform the model in experiment B. The result has shown that machine identified 45 positive and negative images among 50 total images (45/50) achieving 90.15% accuracy. The 5 images that the model miss interpreted were investigated which shows some inconsistencies. After validation, the miss-interpreted images are retrained by using the model of experiment C. To compare the model (experiment C) with microbiologist (professionals and beginners), 30 unknown images are utilized and the model outperformed microbiologist as shown in Table 5.

Discussion

The aim of this study is to developed a neural network model that can accurately detect Mycobacterium tuberculosis and discriminate between a positive cases and negative cases. Computer aided diagnosis (CAD) was introduced in the 1990s and its application has been adopted in medicine and biological sciences for detection of diseases and biological components. One of CAD's application is detecting acid-fast stain bacilli. Even though trained microbiologist can identify bacilli due to their morphological features but it is highly challenging to be precise in identifying and parameterizing morphological features. The application of CNN to classify and characterize bacterial species base on shape such as cluster, rod and round shape and morphology has become an ideal solution to challenges faced by Microbiologist. These models utilized mathematical algorithms developed by imitating brain and neurons of the nervous system which utilize the same mechanism to learn features and classify them based on probability scores [23].

Comparison with other studies

Studies on detection of tuberculosis using artificial neural networks utilized either chest X-ray images [20, 24] and microscopic stained slide images [16,17,19]. To compare the result of the models with state of art, studies that utilize microscopic slide image instead of chest x-ray images are used. As shown in Table 4, the models performances using pre-trained AlexNet (for experiment A & B) achieved a better performance in terms of accuracy than CNN (94%, 97% AND 95%) and SVM+CNN (96.8%) and below F-R-CNN and F-R-CNN+CNN (98.3% and 98.4% respectively) while experiment C achieved higher accuracy than models listed in Table 6. In terms of sensitivity, the proposed network achieved a better sensitivity with 96.77% and 98.59% than other approaches listed in Table 6. This

shows the efficiency of using pretrained network which was already use to trained thousands or millions of datasets compare to building network from scratch.

Machine vs Human Result
Comparison of Experiment B model (using 50 unknown images) model with Real result and microbiologist

True positives (TPs) are the number of image samples that are correctly identified by the classifier (in this context the Actual real result) as positive samples. This also means the total number of real positive samples (31) among 50 images which remain constant. False positives (FPs) are the number of image samples that are incorrectly identified as negative by the classifier (in this context machine, beginners and certified Microbiologists) when in actual result they are positive. False negatives (FNs) are the number of image samples that are incorrectly classified as positive by the classifier (in this context machine, beginners and certified Microbiologists) when in actual result they are negative. True negatives (TNs) are the number of negative image samples that are correctly identified as negative samples by the classifier (in this context the Actual real result) as negative samples. This also means the total number of real negative samples (19) among 50 images which remain constant as shown in Supplementary Table 1 and 2.

Sensitivity is termed as "True positive rate (TPR)" which is the proportion of positive image samples that are correctly identified as positive sample (i.e. it shows the percentage of positive samples that are correctly identified as positives). These results show that the machine correctly identified 29 positive images among 31 positive images (29/31) achieving 93.94% sensitivity. Beginner 1 correctly identified 27 positive images among 31 positive images (27/31) achieving 88.57% sensitivity. Beginner 2 correctly identified 26 positive images among 31 positive images (26/31) achieving 86.11% sensitivity. Certified 1 correctly identified 30 positive images among 31 positive images (30/31) achieving 96.88% sensitivity. Certified 2 identified all the positive samples correctly achieving 100 sensitivity.

Specificity is termed as "False positive rate (FPR)" which is the proportion of positive samples that are incorrectly identified as positive samples (i.e. it shows the percentage of negative samples that are incorrectly identified as positives). These results have shown that the machine correctly identified 16 negative images among 19 negative images (16/19) achieving 86.36% specificity. Beginner 1 and 2, Certified 1 and 2

identified all the negative samples correctly achieving 100% specificity.

Accuracy is the sum of sensitivity and specificity over 2. The result has shown that machine identified 45 positive and negative images among 50 total images (45/50) achieving 90.15% accuracy. Beginner 1 identified 46 positive and negative images among 50 total images (46/50) achieving 94.29% accuracy. Beginner 2 identified 45 positive and negative images among 50 total images (45/50) achieving 93.06% accuracy. Certified 1 identified 49 positive and negative images among 50 total images (49/50) achieving 98.44% accuracy. Certified 2 identified all the positive and negative samples correctly achieving 100% accuracy. Certified pathologists have more experience that is why they achieved higher accuracy, sensitivity and specificity as shown in Supplementary Table 1.

Comparison of Experiment C model (using 30 unknown images) model with Real result and microbiologist

For experiment C, true positive (TP) are the number of image samples that are correctly identified by the classifier (in this context the Actual real result) as positive samples. This also means the total number of real positive samples (18) among 30 images which remain constant. False positive is the number of image samples that are incorrectly identified as negative by the classifier (in this context machine, beginners and certified Microbiologist) when in actual result they are Positive. False negatives are the number of image samples that are incorrectly classified as positive by the classifier (in this context Machine, beginners and certified microbiologist) when in actual result they are negative. True negatives (TNs) are the number of negative image samples that are correctly identified as negative samples by the classifier (in this context the Actual real result) as negative samples. This also means the total number of real negative samples (12) among 30 images which remain constant as shown in Supplementary Table 3.

For sensitivity, the results show that machine, beginner 1 and certified 2 correctly identified 18 positive images among 18 positive images (18/18) achieving 100% sensitivity. Beginner 2 and certified 1 correctly identified 17 positive images among 18 positive images (17/18) achieving 94.74% sensitivity.

For specificity, the results show that machine and certified 1 correctly identified 12 negative images among 12 negative images (12/12) achieving 100% specificity. Beginner 1 correctly identified 10 negative images among 12 negative images (10/12) achieving 92.86 % specificity. Beginner 2 and certified 2 correctly

identified 11 negative images among 12 negative images (11/12) achieving 92.31% specificity.

For accuracy, the result show that machine identified 30 positive and negative images among 30 total images (30/30) achieving 100% accuracy. Beginner 1 identified 28 positive and negative images among 30 total images (28/30) achieving 92.86% accuracy. Beginner 2 identified 28 positive and negative images among 30 total images (28/30) achieving 96.15% accuracy. Certified 1 identified 29 positive and negative images among 30 total images (29/30) achieving 97.37% accuracy. Certified 2 identified 1 identified 29 positive and negative images among 30 total images (29/30) achieving 96.15% accuracy as shown in Supplementary Table 3 and 4.

There is increase application of deep learning models in healthcare system ranging from diagnosis of diseases, therapy, prediction and medical records. However, one of the major drawbacks is lack of sufficient data. Even though there is rise in data storage but still more is needed. Another challenge is related to variability of data which include resolution, noise and contrast which limit application of deep learning in healthcare system. Deep learning models function less efficiently due to generalization as a result of noise and generation of data from different modalities. Deep learning networks are hungry of data, the more data you feed the network the better accuracy and efficiency. Even though transfer learning-based models require less data compare to other models but still researchers encountered challenges of obtaining data due to privacy and confidentiality of medical data.

Conclusions and limitations

In these studies, the detection of tuberculosis bacilli in stain-slide images is achieved by utilizing deep neural network in the form of AlexNet architecture. For Experiment A, the model achieved 98.15 % accuracy, 96.77% sensitivity and 100% specificity, For Experiment B, the model achieved 98.09% accuracy, 98.59% sensitivity and 97.67% specificity. For Experiment C, the model achieved 98.73% testing accuracy, 98.59 sensitivity, 98.84% specificity which sound robust and promising. The result in experiment C outperformed employed microbiologist in terms of accuracy, sensitivity and specificity. Pretrained AlexNet Models can be a promising support system for microbiologist to detect microscopic stain images that contain *Mycobacterium tuberculosis* bacilli and help microbiologist make accurate clinical decisions. These models have shown potential to relieve the tedious and

heavy workload experiencing by pathologist and maximize or prevent miss diagnosis.

There are several limitations in this study. Dataset were collected from two different centers which use different microscopic magnification range and background. Validation by using microscopic images obtained at other laboratories, and other microscope devices and techniques may enhance the generalizability the results achieved. Therefore, further studies should be done with microscopic slides images obtain from multi-center laboratories and training using deeper models such as ResNet and GoogleNet.

References

1. Panicker RO, Kalmady KS, Rajan J, Sabu MK (2018) Automatic detection of tuberculosis bacilli from microscopic sputum smear images using deep learning methods. Biocybern Biomed Eng 38: 691-699.
2. World Health Organization (WHO) (2019) Global tuberculosis report 2019. World Health Organization, Geneve (ISBN 978-92-4-156571-4).
3. Priya E, Srinivasan S (2015) Separation of overlapping bacilli in microscopic digital TB images. Biocybern Biomed Eng 35: 87-99.
4. Abiyev RH, Ma'aitah MK (2018) Deep convolutional neural networks for chest diseases detection. J Healthc Eng 2018: 4168538.
5. Helwan A, Ozsahin DU (2017) Sliding window based machine learning system for the left ventricle localization in MR cardiac images. Appl Comput Intell Soft Comput 2017: 3048181.
6. Mnih A, Hinton GE (2009) A scalable hierarchical distributed language model. In D Koller, D Schuurmans, Y Bengio, L Bottou (Eds.) Advances in neural information processing systems 21: 22nd annual conference on neural information processing systems 2008 (pp. 1081-1088). San Diego, CA, USA: Neural Information Processing Systems Foundation.
7. Helwan A, Abiyev R (2016) Shape and texture features for the identification of breast cancer. In Ao SI, Douglas C, Grundfest WS (Eds.) Proceedings of the world congress on engineering and computer science 2016, 19-21 October. Volume 2 (pp. 542-547). San Francisco, USA: Newswood Limited.
8. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115: 211-252.
9. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint: 1409.1556.
10. Acharya UR, Oh SL, Hagiwara Y, Tan JH, Adeli H (2018) Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. Comput Biol Med 100: 270-278.
11. Smith LN, Topin N (2016) Deep convolutional neural network design patterns. arXiv preprint: 1611.00847.
12. Al-Turjman FM (2016) Towards smart ehealth in the ultra large-scale internet of things era. In S. Barati (Ed.) 2016 23rd Iranian conference on biomedical engineering and 2016 1st international Iranian conference on biomedical engineering (ICBME) (pp. 102-105). Tehran Iran: Institute of electrical and electronics engineers (IEEE).
13. Serte S, Demirel H (2019) Gabor wavelet-based deep learning for skin lesion classification. Comp Bio Med 113: 103423.
14. Kallianos K, Mongan J, Antani S, Henry T, Taylor A, Abuya J, Kohli M (2019) How far have we come? Artificial intelligence for chest radiograph interpretation. Clin Radiol 74: 338-345.
15. Smith KP, Kang AD, Kirby JE (2018) Automated interpretation of blood culture gram stains by use of a deep convolutional neural network. J Clin Microbiol 56: e01521-17.
16. Xiong Y, Ba X, Hou A, Zhang K, Chen L, Li T (2018) Automatic detection of *Mycobacterium tuberculosis* using artificial intelligence. J Thorac Dis 10: 1936-1940.
17. Khan MT, Kaushik AC, Ji L, Malik SI, Ali S, Wei DQ (2109) Artificial neural networks for prediction of tuberculosis disease. Front Microbiol 10: 395.
18. Costa Filho CF, Levy PC, Xavier CD, Fujimoto LB, Costa MG (2015) Automatic identification of tuberculosis mycobacterium. Res Biomed Eng 31: 33-43.
19. El-Melegy M, Mohamed D, ElMelegy T, Abdelrahman M (2019) Identification of tuberculosis bacilli in ZN-stained sputum smear images: a deep learning approach. In 2019 IEEE/CVF conference on computer vision and pattern recognition workshops CVPRW (pp. 1131-1137). Long Beach, CA, USA: Institute of electrical and electronics engineers (IEEE).
20. Aguiar FS, Torres RC, Pinto JV, Kritski AL, Seixas JM, Mello FC (2016) Development of two artificial neural network models to support the diagnosis of pulmonary tuberculosis in hospitalized patients in Rio de Janeiro, Brazil. Med Biol Eng Comput 54: 1751-1759.
21. Fan C, Hauser H (2018) Fast and accurate cnn-based brushing in scatterplots. Comput Graph Forum 37: 111-120.
22. Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S (2012) The'K'in K-fold cross validation. In 20th european symposium on artificial neural networks, ESANN 2012, Bruges, Belgium, 25-27 April 2012 (pp. 441-446) Bruges Belgium.
23. Matsugu M, Mori K, Mitari Y, Kaneda Y (2003) Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Netw. 16: 555-559.
24. Yahiaoui A, Er O, Yumusak N (2017) A new method of automatic recognition for tuberculosis disease diagnosis using support vector machines. Biomed Res 28: 4208-4213.

Corresponding author

Abdullahi Umar Ibrahim, PhD
Department of Biomedical Engineering, Near East University Near East Boulevard, ZIP: 99138 Nicosia TRNC Mersin 10 – Turkey,
Phone: +905488314346
Email: Abdullahi.umaribrahim@neu.edu.tr

Conflict of interests: No conflict of interests is declared.

Annex – Supplementary Items

Supplementary Table 1. Results of participants *vs.* actual results using 50 images.

S No	Machine	Beginner 1	Certified 1	Beginner 2	Certified 2	Actual (Real) results
1	+	- (MI)	+	+	+	+
2	+	+	+	+	+	+
3	+	+	+	+	+	+
4	- (MI)	+	+	+	+	+
5	+ (MI)	-	-	-	-	-
6	-	-	-	-	-	-
7	-	-	-	-	-	-
8	+	+	+	+	+	+
9	-	-	-	-	-	-
10	+	+	+	- (MI)	+	+
11	+	+	+	+	+	+
12	+	+	+	+	+	+
13	+	+	+	+	+	+
14	+	- (MI)	+	+	+	+
15	-	-	-	-	-	-
16	-	-	-	-	-	-
17	+ (MI)	-	-	-	-	-
18	+	+	+	+	+	+
19	+	+	+	+	+	+
20	+	+	+	+	+	+
21	-	-	-	-	-	-
22	-	-	-	-	-	-
23	+	+	+	+	+	+
24	+	- (MI)	+	+	+	+
25	+	+	+	+	+	+
26	+	+	+	+	+	+
27	+	+	+	+	+	+
28	-	-	-	-	-	-
29	-	-	-	-	-	-
30	+	+	+	+	+	+
31	-	-	-	-	-	-
32	-	-	-	-	-	-
33	+	+	+	+	+	+
34	+ (MI)	-	-	-	-	-
35	-	-	-	-	-	-
36	+	+	+	+	+	+
37	+	+	+	+	+	+
38	+	+	+	- (MI)	+	+
39	-	-	-	-	-	-
40	-	-	-	-	-	-
41	+	+	+	+	+	+
42	+	+	+	- (MI)	+	+
43	+	+	+	+	+	+
44	-	-	-	-	-	-
45	+	+	+	- (MI)	+	+
46	+	+	+	+	+	+
47	+	+	+	+	+	+
48	+	- (MI)	- (MI)	- (MI)	+	+
49	-	-	-	-	-	-
50	- (MI)	+	+	+	+	+
	45 CC	46 CC	49 CC	45 CC	50 CC	
	5 MI	4 MI	1 MI	5 MI	0 MI	

CC: Correctly classified or identified; MI: Misinterpreted.

Supplementary Table 2. Distribution of tuberculosis positive and negative in confusion matrix for 50 images.

Participants	True Positive	False Positive	False Negative	True Negative
Machine	31	3	2	19
Beginner 1	31	0	4	19
Beginner 2	31	0	5	19
Certified 1	31	0	1	19
Certified 2	31	0	0	19

Supplementary Table 3. Results of participants *vs.* actual results using 30 images.

s/No	Machine	Beginner 1	Certified 1	Beginner 2	Certified 2	Actual (Real results)
1	-	+ (MI)	-	-	-	-
2	-	-	-	-	-	-
3	+	+	+	+	+	+
4	+	+	+	+	+	+
5	-	-	-	-	-	-
6	+	+	+	+	+	+
7	-	-	-	-	-	-
8	-	-	-	-	-	-
9	+	+	+	+	+	+
10	+	+	+	+	+	+
11	+	+	- (MI)	- (MI)	+	+
12	+	+	+	+	+	+
13	-	-	-	-	-	-
14	+	+	+	+	+	+
15	+	+	+	+	+	+
16	-	-	-	-	-	-
17	+	+	+	+	+	+
18	-	-	-	-	-	-
19	-	+ (MI)	-	+ (MI)	+ (MI)	-
20	-	-	-	-	-	-
21	+	+	+	+	+	+
22	+	+	+	+	+	+
23	+	+	+	+	+	+
24	+	+	+	+	+	+
25	+	+	+	+	+	+
26	-	-	-	-	-	-
27	+	+	+	+	+	+
28	-	-	-	-	-	-
29	+	+	+	+	+	+
30	+	+	+	+	+	+
		28 CC	29 CC	28 CC	29 CC	
		2 MI	1 MI	2 MI	1 MI	

CC: Correctly classified or identified; MI: Misinterpreted.

Supplementary Table 4. Distribution of tuberculosis positive and negative in confusion matrix for 30 images.

Participants	True Positive	False Positive	False Negative	True Negative
Machine	18	0	0	12
Beginner 1	18	2	0	12
Beginner 2	18	1	1	12
Certified 1	18	0	1	12
Certified 2	18	1	0	12

Code

```
TRAINING
>> imds = imageDatastore('training', ...
      'IncludeSubfolders',true, ...
      'LabelSource','foldernames');
>> [imdsTrain,imdsValidation] = splitEachLabel(imds,0.7,'randomized');
>> numTrainImages = numel(imdsTrain.Labels);
>> net = alexnet;
>> net.Layers
>> inputSize = net.Layers(1).InputSize
>> layersTransfer = net.Layers(1:end-3);
>> numClasses = numel(categories(imdsTrain.Labels))
>> layers = [
      layersTransfer
      fullyConnectedLayer(numClasses,'WeightLearnRateFactor',20,'BiasLearnRateFactor'
,20)
      softmaxLayer
      classificationLayer];
>> pixelRange = [-30 30];
imageAugmenter = imageDataAugmenter( ...
      'RandXReflection',true, ...
      'RandXTranslation',pixelRange, ...
      'RandYTranslation',pixelRange);
augimdsTrain = augmentedImageDatastore(inputSize(1:2),imdsTrain, ...
      'DataAugmentation',imageAugmenter);
>> augimdsValidation = augmentedImageDatastore(inputSize(1:2),imdsValidation);
>> options = trainingOptions('sgdm', ...
      'MiniBatchSize',10, ...
      'MaxEpochs',20, ...
      'InitialLearnRate',1e-4, ...
      'ValidationData',augimdsValidation, ...
      'ValidationFrequency',3, ...
      'ValidationPatience',Inf, ...
      'Verbose',true, ...
      'Plots','training-progress');
>> netTransfer = trainNetwork(augimdsTrain,layers,options);

VALIDATION
>> [YPred,scores] = classify(netTransfer,augimdsValidation);
>> YValidation = imdsValidation.Labels;
>> accuracy = mean(YPred == YValidation)

Specificity and Sensitivity
>> y = grp2idx(YValidation);
>> test = grp2idx(YPred);
>> classperf(y,test)

TESTING
>> imdsTest = imageDatastore('Testing', ...
'IncludeSubfolders',true, ...
'LabelSource','foldernames');
>> [YPred,scores] = classify(netTransfer,imdsTest);
>> YValidation = imdsTest.Labels;
>> accuracy = mean(YPred == YValidation)
```