Comparing deep learning models for population screening using chest radiography

R. Sivaramakrishnan, Sameer Antani, Sema Candemir, Zhiyun Xue, Joseph Abuya, et al.

R. Sivaramakrishnan, Sameer Antani, Sema Candemir, Zhiyun Xue, Joseph Abuya, Marc Kohli, Philip Alderson, George Thoma, "Comparing deep learning models for population screening using chest radiography," Proc. SPIE 10575, Medical Imaging 2018: Computer-Aided Diagnosis, 105751E (27 February 2018); doi: 10.1117/12.2293140

Event: SPIE Medical Imaging, 2018, Houston, Texas, United States

Comparing Deep Learning Models for Population Screening using Chest Radiography

R. Sivaramakrishnan[*a], Sameer Antani[a], Sema Candemir[a], Zhiyun Xue[a], Joseph Abuya[b], Marc Kohli[c], Philip Alderson[a, d], George Thoma[a]

[a]Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD, USA 20894-0001; [b]Department of Radiology and Imaging, School of Medicine, Moi University, 3900 Eldoret, Kenya 30100; [c]Department of Radiology and Biomedical Imaging, University of California San Francisco, San Francisco, CA, USA 94143; [d]School of Medicine, Saint Louis University, 1402 South Grand Blvd., St. Louis, MO, USA 63104

ABSTRACT

According to the World Health Organization (WHO), tuberculosis (TB) remains the most deadly infectious disease in the world. In a 2015 global annual TB report, 1.5 million TB related deaths were reported. The conditions worsened in 2016 with 1.7 million reported deaths and more than 10 million people infected with the disease. Analysis of frontal chest X-rays (CXR) is one of the most popular methods for initial TB screening, however, the method is impacted by the lack of experts for screening chest radiographs. Computer-aided diagnosis (CADx) tools have gained significance because they reduce the human burden in screening and diagnosis, particularly in countries that lack substantial radiology services. State-of-the-art CADx software typically is based on machine learning (ML) approaches that use hand-engineered features, demanding expertise in analyzing the input variances and accounting for the changes in size, background, angle, and position of the region of interest (ROI) on the underlying medical imagery. More automatic Deep Learning (DL) tools have demonstrated promising results in a wide range of ML applications. Convolutional Neural Networks (CNN), a class of DL models, have gained research prominence in image classification, detection, and localization tasks because they are highly scalable and deliver superior results with end-to-end feature extraction and classification. In this study, we evaluated the performance of CNN based DL models for population screening using frontal CXRs. The results demonstrate that pre-trained CNNs are a promising feature extracting tool for medical imagery including the automated diagnosis of TB from chest radiographs but emphasize the importance of large data sets for the most accurate classification.

Keywords: Tuberculosis, deep learning, machine learning, convolutional neural network, chest radiograph, classification, customization, screening

1. INTRODUCTION

Tuberculosis (TB) resulted in 1.7 million deaths worldwide in 2016. The World Health Organization (WHO) recommends chest X-ray (CXR) screening as a part of the routine protocol for high-risk groups[1] such as those with HIV/AIDS. TB is endemic in under-resourced regions such as parts of sub-Saharan Africa. Lack of expertise in interpreting radiology reports has been reported, especially in TB endemic regions, severely impairing screening efficacy[2]. Thus, the current research is focused on developing cost-effective, computer-aided diagnosis (CADx) tools to assist medical providers in interpreting CXRs and improving the quality of diagnostic imaging. Appropriate use and development of these systems could help greatly improve the detection accuracy and reduce the human burden in screening and diagnosis of conditions like TB, particularly in endemic regions and third world countries that lack substantial radiology services.

State-of-the-art CADx software uses machine learning (ML) techniques that utilize global and local feature descriptors to extract features from the underlying data. ML tools have been previously applied to detect abnormal texture in chest radiographs and to demonstrate extraction of texture and shape features and classification with a binary classifier in the process of TB screening from CXRs[3–5]. Morphology-based algorithms have been proposed to extract features including circularity, size, contrast and local curvature of the lung nodules for classification of abnormal and normal CXRs[6]. A

Medical Imaging 2018: Computer-Aided Diagnosis, edited by Nicholas Petrick, Kensaku Mori, Proc. of SPIE Vol. 10575, 105751E · © 2018 SPIE · CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2293140

study focused on image level labeling using Local Binary Patterns (LBP) for detecting and classifying chest pathology was proposed[7]. Bag of Visual words (BOVW) was used in discriminating normal and pathological chest radiographs[8]. There are a few commercially available CADx tools based on ML approaches that use a combination of textural and morphological features. This includes CAD4TB, a CADx software from the Image Analysis Group, Netherlands that has an area under the curve (AUC) ranging from 0.71 to 0.84 in a range of studies in the process of detecting pulmonary abnormalities[9]. Another study achieved AUC of 0.87 to 0.90 while using a support vector machine (SVM) classifier to detect pulmonary TB from CXR images using texture and shape features[10]. However, these CADx tools used hand-engineered features that change with size, background, angle, and position of the region of interest (ROI).

To overcome challenges of devising high-performing hand-engineered features that capture the variation in the underlying data, Deep Learning (DL), also known as hierarchical machine learning, has been used with significant success[11]. DL models are constructed using a cascade of layers of non-linear processing units for end-to-end feature extraction and classification[12]. The models excel with high-dimensional datasets, especially images having multiple levels of representations. Convolutional neural networks (CNN), a class of DL models, have gained immense research prominence as they promise to deliver high-quality classification without the need for manual feature selection. Unlike kernel-based algorithms like SVMs, DL models exhibit improved performance with an increasing number of training samples and computational resources, making them highly scalable[13]. However, high performance using DL comes at the cost of huge amounts of labeled data which are difficult to obtain, particularly in biomedical applications. Transfer learning methods are commonly used to relieve issues with data inadequacy where DL models are pre-trained on a very large dataset like ImageNet, containing 15 million annotated natural/stock photography images from over 21,000 categories[14]. These models can either be used as an initialization for a wide range of computer vision problems or a feature extractor for the task of interest[15]. In 2012, AlexNet was proposed[16] that used a sequential stacking of convolutional layers and rectified linear units (ReLU). The model used dropout layers to combat the problem of data overfitting and was trained using a stochastic gradient descent (SGD) algorithm. A VGG model was proposed in 2014 that used only 3×3 sized filters all through its length. The model won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) benchmark in object localization task in 2014. Several variants of these networks including VGG-16 and VGG-19 were developed, where “16” and “19” indicated the number of weight layers in the network[17]. Another model named Xception was proposed that used depth-wise separable convolutions[18] to outperform the Inception-V3 DL model[19] on the ImageNet data classification task. In 2015, a model based on deep residual connections (ResNet) was proposed that achieved superhuman performance and won the ILSVRC classification task[20].

Medical images containing visual representations of the internal structures of the body have little in common with natural images[21]. Under these circumstances, in contrast to pre-trained DL models, a customized DL model, trained on the underlying medical imagery could learn task-specific features to aid in improved accuracy. A customized model would be highly compact with less trainable parameters and reduced computation cost. Literature studies reveal the use of customized CNNs toward medical image understanding/analysis. In[22], a customized CNN based DL model was trained to perform automated classification of parasitized and uninfected red blood cells (RBCs) in thin blood smear images to aid in malaria diagnosis. The customized model achieved 97.37% accuracy in comparison to pre-trained CNNs that achieved 91.99%. The proposed model demonstrated superiority in performance metrics including sensitivity, specificity, precision, and F1-score. In[23], a 12-layer customized CNN model was used to classify the parasitized and uninfected RBCs to aid in diagnosing malaria. The study also investigated the learned features and salient network activations in the customized model to aid in understanding the learning strategy. The study revealed that the customized CNN outperformed the pre-trained DL models in terms of classification accuracy, model complexity and computation time.

A survey of literature revealed the use of pre-trained CNNs as feature extractors toward automated disease prediction from biomedical imagery. A study reported to use pre-trained models toward identifying pleural effusion and cardiomegaly from frontal CXRs[24]. The performance of the pre-trained CNN was compared to the classifiers trained on features extracted using LBP, GIST, and Pico-Descriptors[25]. The study achieved promising results with AUC of 0.93 and 0.89 for right pleural effusion and cardiomegaly, respectively through a combination of the features extracted from the pre-trained CNN and Pico-Descriptors. The application of CNN toward TB detection was demonstrated in another study that used a customized CNN model with AlexNet framework, trained on a private CXR dataset of approximately 10K images[26]. The customized model gave poor results when trained with random weight initializations. However, with pre-trained CNNs, the authors obtained competitive results on the publicly available Montgomery and Shenzhen CXR datasets[27], achieving AUC of 0.884 and 0.926 respectively.

Our study aims to evaluate the performance of one customized model and five pre-trained CNNs towards improving the accuracy of TB screening using frontal CXRs. We evaluate the performance of a customized CNN based DL model that learns task-specific features from the posterior-anterior (PA) CXRs that could aid in improving the accuracy of TB detection. The proposed model is optimized for its hyper-parameters in the process of minimizing the classification error. We employ five different pre-trained CNN models to extract features from the PA CXRs to aid in detecting TB manifestations.

The most important contributions of this work are as follows: The proposal of a customized CNN based DL model, optimized for its hyper-parameters toward the process of learning task-specific features from the underlying biomedical imagery, a comparative analysis of the performance of different pre-trained DL models as feature extractors for the task of TB detection, identification of optimal layers in the pre-trained CNNs for extracting the features from the underlying data, and a statistical analysis to test for the presence or absence of a statistically significant difference in performance across the models under study. The following paper is organized as follows: Section 2 elaborates on the materials and methods, section 3 discusses the results, and section 4 concludes the paper.

2. INTRODUCTION

2.1 Data collection and preprocessing

This study uses four datasets that include two publicly available datasets from Montgomery County, Maryland, and Shenzhen, China, maintained by the National Library of Medicine (NLM), National Institutes of Health (NIH). Table 1 presents the details pertaining to the origin and characteristics of the datasets.

Table 1. Datasets and their characteristics.

Origin	# TB positive	# Normal	File type	Bit depth	Resolution
Shenzhen	58	80	PNG	8-bit	4020-4892 × 4020-4892
Montgomery	336	326	PNG	8-bit	948-3001 × 1130-3001
Kenya	238	729	PNG	8-bit	1312-1852 × 1094-1838
India	153	153	JPG	8-bit	1024-2480 × 1024-2480

Ground truth information is available in the form of clinical readings, annotating the abnormal locations in the CXRs. The India dataset was created by the National Institute of Tuberculosis and Respiratory Diseases, New Delhi, India and made available by the authors[4]. For the Kenya dataset, Indiana University School of Medicine and Academic Model Providing Access to Healthcare (AMPATH), a Kenyan NGO, collaborated with NLM to make available de-identified CXRs from rural western Kenya as a part of the mobile truck-based screening.

The datasets used in this study contain regions other than the lungs that are irrelevant toward lung TB detection. To alleviate issues due to models that learn features that are irrelevant to detecting lung TB and demonstrating sub-optimal performance, the lung region constituting the ROI is segmented by a method that uses anatomical atlases with non-rigid registration[10]. An instance of a CXR with the detected lung region and cropped lung area using the proposed method is shown in Fig. 1. The method follows a content-based image retrieval (CBIR) approach to identify the training samples that bear resemblance to the patient CXR. The patient-specific anatomical lung shape model is created using SIFT-flow (SIFT: scale-invariant feature transform)[28] for registering the training masks to the patient CXRs. The refined lung boundaries are extracted using an approach based on graph-cut optimization[29] using a customized energy function. The approach is highly robust, resulting in a segmentation accuracy of 94.1% and 91.7% on the Montgomery and India datasets respectively. After lung segmentation, the resulting image is cropped to the size of a bounding box that contains all the lung pixels. The resultant images are enhanced for contrast by applying Contrast Limited Adaptive Histogram Equalization (CLAHE). The images are down-sampled to 224×224 and 299×299 pixel resolutions to suit the input requirements for the customized and pre-trained CNNs.

In this study, we used a Windows® system with Intel® Xeon® CPU E5-2640v3 2.60-GHz processor, 1 TB HDD, 16 GB RAM, a CUDA-enabled Nvidia GTX 1080 Ti 11GB graphical processing unit (GPU), Matlab® R2017b, Keras® with Tensorflow® backend, and CUDA 8.0/cuDNN 5.1 dependencies for GPU acceleration.

2.2 Customized model configuration

The performance of the customized CNN model relies on the architecture shown in Fig. 2. We propose a sequential, 12-layered CNN for the binary task of classifying normal and TB-positive CXRs. The individual datasets were randomly divided into 80% for training and 20% for testing. Within training, we performed 5-fold cross-validation to aid in optimal model selection. Images are down-sampled to 224×224 pixel resolutions. Each CNN block has a convolutional layer, followed by a batch normalization[30] and ReLU layer[16].

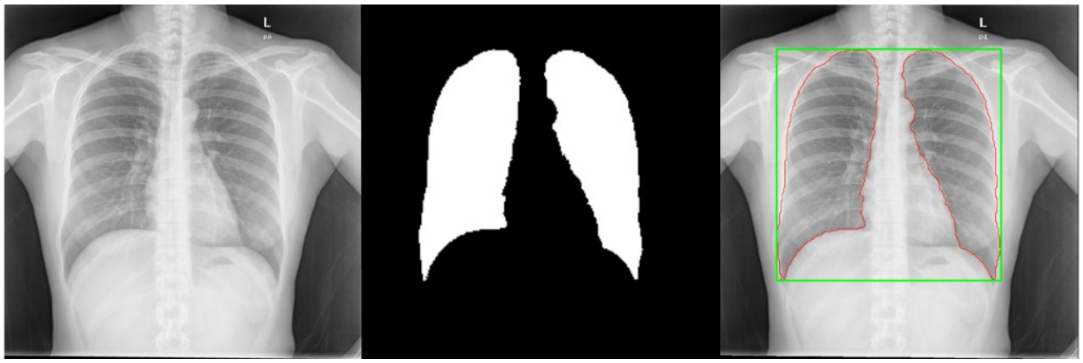

Figure 1. Lung ROI segmentation. a) Original image, b) Lung mask, c) Segmented ROI.

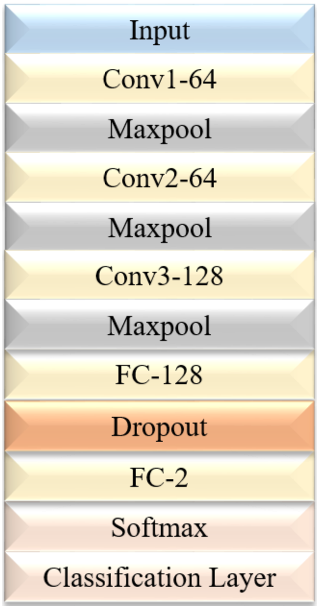

Figure 2. Custom model architecture.

Padding is added to the convolutional layers to ensure that the spatial output dimensions match the original input. The first and second convolutional layer has 64 filters each and the third has 128 filters. A filter size of 3×3 is used uniformly across the layers. A 2×2 max-pooling layer with 2-pixel strides follows the ReLU layers for dimensionality reduction. The model is optimized for its hyper-parameters by a randomized grid search method[31]. The search ranges include [1e-3 5e-2], [0.8 0.99] and [1e-10 1e-2] for the learning rate, SGD momentum and L2-regularization parameters respectively. A Softmax classifier follows the second fully-connected layer to output the prediction distribution between the normal and abnormal classes[32]. The convergence of the model is attributed to proper weight initializations. A method for weight initializations for networks with ReLU non-linear activations is used[20], a more sophisticated initialization than the regular method of weight initialization[11] using the Gaussian distribution, however, the Gaussian is rescaled in accordance with the number of neurons connected to the input of a given layer.

2.3 Feature extraction using pre-trained models

We evaluate the performance of pre-trained DL models that include AlexNet (winner of ILSVRC 2012), VGG-16 and VGG-19 (winner of ILSVRC's localization task in 2014), Xception and ResNet-50 (winner of ILSVRC 2015) in the process of extracting the features from the CXRs across the normal and TB-positive categories. Table 2 shows the number of parameters and depth for the customized and pre-trained CNNs used in this study.

Table 2. Models and their parameters.

Model	# Parameters	Depth
Customized model	2,797,730	9
AlexNet	60,000,000	25
VGG-16	138,357,544	23
VGG-19	143,667,240	26
Xception	22,910,480	126
ResNet-50	25,636,712	168

The segmented ROI constituting the lungs are down-sampled to match the input dimensions of the pre-trained models. Each layer of the pre-trained CNNs produces an activation for the given image. Earlier layers capture primitive features that include blobs, edges, and colors that are abstracted by the deeper layers to form higher level features to present a more affluent image representation. Studies from the literature reveal that features are conventionally extracted from the layer, right before the classification layer[33]. The convolutional part of the pre-trained models including AlexNet, VGG-16, VGG-19, Xception, and ResNet-50 is instantiated, everything up to the fully-connected layers and the models are run on the data to record the activation maps. A small fully-connected model is trained on top of the stored features. The performance of these models is compared with the customized DL model for the datasets under study. For the pre-trained models, the architecture and weights are downloaded from the Keras GitHub repository[34]. We also experimentally determined the optimal layer in these pre-trained DL models for extracting the features to aid in improved TB detection. As with the customized model, we performed 5-fold cross-validation with the individual datasets and evaluated the performance.

3. RESULTS AND DISCUSSIONS

3.1 Comparing the results of feature extraction from different layers

Table 3 shows the results for the customized CNN, trained to learn task-specific features from the datasets under study. The performance of the custom model, with respect to the India dataset, is the most promising, with an accuracy of 0.824 and AUC of 0.900. The reason may be that TB manifestations in this dataset are obvious and are distributed throughout the lungs that give the customized model the opportunity to capture highly discriminative features across healthy controls and TB-positive cases.

Table 3. Performance of the customized model.

Datasets	Accuracy	AUC
Shenzhen	0.820	0.894
Montgomery	0.658	0.744
Kenya	0.572	0.642
India	**0.824**	**0.900**

With the Kenya dataset, the custom model demonstrated sub-optimal performance. The principal reason may be because the dataset has a highly imbalanced distribution of instances across the classes, with 238 abnormal CXRs in comparison to 729 healthy controls. Also, the resolution of CXRs, even after ROI segmentation and CLAHE enhancement is not ideal that further impaired the performance of feature extraction and classification. With the Montgomery dataset, performance limitation may be attributed to the limited size of the dataset and also the degree of imbalance across the classes where 40% of the samples are TB-positive as compared to 60% healthy controls. Table 4 and Table 5 shows the results of using pre-trained CNNs as feature extractors in the process of classifying healthy and TB-positive cases. The second fully connected layer has been selected for feature extraction in AlexNet, VGG-16 and VGG-19 models. The last layer, before the final classification layer, is selected for extracting the features from Xception and ResNet-50 models. The models are trained using SGD with momentum and learning rate of 1e-4.

Table 4. Performance of the pre-trained DL models – Accuracy.

Datasets	AlexNet	VGG-16	VGG-19	Xception	ResNet-50
Shenzhen	**0.842**	0.815	0.778	0.731	0.819
Montgomery	**0.725**	0.708	0.650	0.600	0.676
Kenya	0.657	0.666	**0.679**	0.653	0.678
India	**0.864**	0.748	0.840	0.828	0.812

Table 5. Performance of the pre-trained DL models – AUC.

Datasets	AlexNet	VGG-16	VGG-19	Xception	ResNet-50
Shenzhen	**0.912**	0.881	0.865	0.862	0.893
Montgomery	**0.800**	0.736	0.724	0.672	0.616
Kenya	0.743	0.739	0.722	0.702	**0.753**
India	**0.944**	0.852	0.905	0.890	0.902

For the Shenzhen dataset, AlexNet obtained the best accuracy of 0.842 and AUC of 0.912. The same pattern is observed across the Montgomery and India datasets. For Montgomery dataset, AlexNet obtained the best accuracy of 0.725 and AUC of 0.800. For the India dataset, AlexNet outperformed the other pre-trained CNNs with an accuracy of 0.864 and AUC of 0.944. Only for the Kenya dataset, we observed that the AUC of VGG-19 is slightly better than that of AlexNet, however, the accuracy of AlexNet is higher than that of the other pre-trained models. It can be noted that the results obtained with the India dataset are superior to the results obtained with the other datasets for the reasons discussed earlier. We empirically found that adding dropout improved the classification accuracy of shallow, sequential networks that include AlexNet, VGG-16, and VGG-19 but degraded the performance of deep CNNs that include Xception and ResNet-50.

Among the pre-trained CNNs evaluated in this study, AlexNet outperformed the other models for the datasets used in this study. It could be expected that ResNet-50 would beat all other architectures since their performances were reported in the literature to be clearly superior with the large-scale ImageNet data[20], but they didn't, in this study. The architecture of ResNet-50 is deep, and may be more than is needed for the underlying task of binary medical image classification. For ImageNet data, deeper networks outperform shallow counterparts for the reason that the data is diverse and the networks learn abstractions for a huge selection of classes. In our case, for the binary task of TB detection, the variability in data is several orders of magnitude smaller as compared to ImageNet collection. The top layers of pre-trained CNNs like Xception and ResNet-50 are probably too specialized, progressively more complex and not the best candidate to re-use for the task of our interest. This explains the difference in performance in our case.

Literature studies reveal that features extracted from shallow layers of deep CNNs are useful in detecting small objects in the underlying data[35]. We found that these results hold good for our TB detection task. We evaluated the performance of pre-trained CNNs by extracting features from different layers in the process of identifying the optimal layer for feature extraction, for the datasets under study. We chose the candidate convolutional layers from the 3rd, 4th, 5th, and final stage of the pre-trained models. The naming conventions for these layers are based on the models obtained from Keras® neural network library. We observed that for pre-trained models the performance of the layer before the classification layer was degraded compared to the other layers. The layers that gave the best classification accuracy and AUC for the different pre-trained CNNs are listed in Table 6. Table 7 and Table 8 presents the results obtained by extracting the features from these optimal layers in the pre-trained CNNs. In contrast to the results obtained in Table 4 and Table 5, for Shenzhen dataset, VGG-16 performed better in terms of accuracy, however, AlexNet gave the highest AUC of 0.926. For Montgomery dataset, Xception model was more accurate, however, VGG-19 gave the best AUC of 0.833. For Kenya dataset, both AlexNet and VGG-16 performed equally well with an accuracy of 0.695, however, AlexNet demonstrated a slightly better AUC of 0.775 as compared to other models. For India dataset, VGG-16 gave the highest accuracy of 0.876, however, VGG19 gave the best AUC of 0.956.

Table 6. Candidate layers giving the best performance for the datasets.

Model	Layer
AlexNet	fc6
VGG-16	Conv5_1
VGG-19	Conv5_1
Xception	Block11_Sepconv1
ResNet-50	Res4c_branch2a

Table 7. Performance of the pre-trained DL models with optimal features – Accuracy.

Datasets	AlexNet	VGG-16	VGG-19	Xception	ResNet-50
Shenzhen	0.853	**0.855**	0.852	0.815	0.802
Montgomery	0.725	0.742	0.733	**0.758**	0.717
Kenya	**0.695**	**0.695**	0.690	0.679	0.691
India	0.872	**0.876**	0.872	0.812	0.860

Table 8. Performance of the pre-trained DL models with optimal features – AUC.

Datasets	AlexNet	VGG-16	VGG-19	Xception	ResNet-50
Shenzhen	**0.926**	0.917	0.916	0.900	0.892
Montgomery	0.818	0.829	**0.833**	0.810	0.820
Kenya	**0.775**	0.774	0.774	0.754	0.740
India	0.949	0.950	**0.956**	0.894	0.944

The results demonstrated that the final layer of pre-trained CNNs is not always optimal for the underlying data. Features from shallow layers performed better than deep features to aid in improved disease prediction.

3.2 Statistical analysis

The selection of the best performing model is validated using one-way analysis of variance (ANOVA) parametric test[36]. The test is performed to determine the presence or absence of a statistically significant difference between the means of three or more individual, unrelated groups. Specifically, one-way ANOVA tests the null hypothesis (H0), given by:

$$H0\colon \mu 1 = \mu 2 = \mu 3 = \mu k \tag{1}$$

where μ = mean of parameters for the individual groups and k = total number of groups. If a statistically significant result is returned by the test, the alternative hypothesis (HA) is accepted that infers that there is a statistically significant difference between the means of at least two groups involved in the study. One-way ANOVA is an omnibus test that couldn't identify the specific groups that have statistically significant differences in their mean values. A post-hoc analysis is needed to identify the specific groups that demonstrate statistically significant difference in their mean values. A Tukey's post-hoc test is performed to identify these groups[37]. The consolidated results of one-way ANOVA and Tukey's post-hoc tests for different performance metrics, for the different models are shown in Table 9. To conduct one-way ANOVA, the data has to satisfy the assumptions of normality of data, homogeneity of variances and independence of observations. Normality of data is tested with Shapiro-Wilk test for normality[38] and Levene's statistic is used to test the homogeneity of variances for the data under study[39].

Table 9. Consolidated results of one-way ANOVA and Tukey post-hoc test.

Datasets	**Parameter**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**ANOVA summary**		**Tukey**
								F	***p***	
Shenzhen	Accuracy	0.853	0.855	0.852	0.815	0.802	0.709	6.674	0.001	M1, M2, M3 & M6, *p* = 0.001 M4 & M6, *p* = 0.022 M5 & M6, *p* = 0.005
	AUC	0.925	0.917	0.916	0.900	0.892	0.786	8.153	0.000	M1, M2, M3 & M6, *p* = 0.000 M4 & M6, *p* = 0.002 M5 & M6, *p* = 0.005
Kenya	Accuracy	0.695	0.695	0.690	0.679	0.691	0.572	1.963	0.121	-
	AUC	0.775	0.774	0.774	0.754	0.740	0.642	1.428	0.250	-
Montgomery	Accuracy	0.725	0.742	0.733	0.758	0.717	0.658	0.761	0.587	-
	AUC	0.818	0.829	0.833	0.810	0.818	0.744	0.585	0.711	-
India	Accuracy	0.872	0.876	0.872	0.812	0.860	0.824	1.923	0.128	-
	AUC	0.949	0.950	0.956	0.894	0.944	0.890	4.411	0.005	M3 & M6, *p* = 0.036

When the results of these test statistics are not statistically significant, we could ensure that the data is normally distributed and have equal or homogeneous variances. We have used the notations, M1, M2, M3, M4, M5 and M6 to denote the mean values of accuracy and AUC metrics for AlexNet, VGG-16, VGG-19, Xception, ResNet-50 and customized model respectively. The values of the parameters across different folds for the individual models are tested for normality and homogeneity of variances with the Shapiro-Wilk and Levene's test respectively. Since no statistical significance is observed, we justify the use of one-way ANOVA test statistic in this study. We performed these analyses using IBM® SPSS® statistical package (IBM Corp. Released 2015. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.).

One-way ANOVA test reveals that, for Shenzhen dataset, there is a statistically significant difference ($p < 0.05$) between the models for the accuracy metric (Accuracy: F (5, 24) = 6.674, p = 0.001 and AUC: F (5, 24) = 8.153, p = 0.000). Tukey post-hoc test further reveals that the customized model has a statistically significant difference in its mean value for accuracy, as compared to the other pre-trained models, except ResNet-50. With AUC, the customized model differs statistically significantly from all the pre-trained models. For Kenya and Montgomery datasets, there are no statistically significant differences observed in the mean values for the performance metrics. For India dataset, a statistically significant difference is observed for the mean value for AUC (F (5, 24) = 4.411, p = 0.005). Tukey post-hoc test reveals that the customized model has a statistically significant difference for AUC as compared to that of VGG-19. These findings indicate that there is no statistically significant difference in the performances among the pre-trained models across the datasets, however, the customized model shows a statistically significant difference in performance when compared to the pre-trained models, for Shenzhen and India datasets.

Tables 10 and Table 11 compare the results obtained in this study with the studies available in the literature on TB detection[4,3,26]. In terms of accuracy, with Shenzhen dataset, the features extracted from the optimal layer of VGG-16 model outperforms the state-of-the-art, however, the AUC values remains similar to that reported by Hwang *et al.* With Montgomery dataset, literature studies yield better results than the proposed study, the reason may be attributed to the fact that, unlike rule-based local and global feature descriptors, the performance of DL models suffer due to scarcity of data that prevent learning highly discriminative features and a highly imbalanced distribution of instances across the positive and negative classes. The same holds good to India dataset since DL models don't have enough data to learn the discriminative features to arrive at promising results.

Table 10. Comparison with literature – Accuracy.

Datasets	**Proposed**	**Stefan *et al.*[3]**	**Hwang *et al.*[26]**	**Chauhan *et al.*[4]**
Shenzhen	**0.855**	0.840	0.837	-
Montgomery	0.758	**0.783**	0.674	-
Kenya	**0.695**	-	-	-
India	0.876	-	**-**	**0.943**

Table 11. Comparison with literature – AUC.

Datasets	**Proposed**	**Stefan *et al.*[3]**	**Hwang *et al.*[26]**	**Chauhan *et al.*[4]**
Shenzhen	**0.926**	0.900	**0.926**	-
Montgomery	0.833	0.869	**0.884**	-
Kenya	**0.775**	-	-	-
India	0.956	-	**-**	**0.960**

4. CONCLUSIONS

In this study, we compared the performance of one customized DL model and five pre-trained DL models toward improving the accuracy of TB screening from frontal CXRs. We observed that the performance of pre-trained DL models is statistically significantly better than the customized model. We also identified that features from shallow layers of pre-trained CNNs gave better results in comparison to features from the deeper layers. The performance of the models used in this study is impacted by the scarcity of data and highly imbalanced distribution across the classes. With regard to advancements in TB detection, the current study and other studies[40] demonstrate that future analysis demands large-scale biomedical datasets. The performance of customized and pre-trained DL models could be better with such a large selection of data. Under the circumstances in the current study, which we believe simulate current real-world conditions, the pre-trained DL models provided the best performance.

5. ACKNOWLEDGMENT

This work is supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications. We acknowledge Kelvin Ogot and Pamela Wamboye of AMPATH for their assistance with data collection.

6. CONFLICT OF INTEREST

The authors have no conflict of interest to report.

REFERENCES

[1] "World Health Organization. Global tuberculosis report 2016," (2016).

[2] Melendez, J., Sanchez, C. I., Philipsen, R. H. H. M., Maduskar, P., Dawson, R., Theron, G., Dheda, K. and van Ginneken, B., "An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information.," Sci. Rep. 6, 25265 (2016).

[3] 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. X., Lu, P. X. and McDonald, C. J., "Automatic tuberculosis screening using chest radiographs," IEEE Trans. Med. Imag. 33(2), 233–245 (2014).

[4] Chauhan, A., Chauhan, D. and Rout, C., "Role of gist and PHOG features in computer-aided diagnosis of tuberculosis without segmentation," PLoS One 9(11), 1–12 (2014).

[5] Ding, M., Antani, S., Jaeger, S., Xue, Z., Candemir, S., Kohli, M. and Thoma, G., "Local-global classifier fusion for screening chest radiographs," Proc. SPIE 10138, 101380A (2017).

[6] Katsuragawa, S. and Doi, K., "Computer-aided diagnosis in chest radiography," Comput. Med. Imag. Graph. 31(4–5), 212–223 (2007).

[7] Carrillo-de-Gea, J. M. and García-Mateos, G., "Detection of Normality / Pathology on Chest Radiographs Using Lbp," Proc. ACM-BCB, 167–172 (2010).

[8] Avni, U., Greenspan, H., Konen, E., Sharon, M. and Goldberger, J., "X-ray categorization and retrieval on the organ and pathology level using patch-based visual words," IEEE Trans. Med. Imag. 30(3), 733–746 (2011).

[9] Pande, T., Cohen, C., Pai, M. and Ahmad Khan, F., "Computer-aided detection of pulmonary tuberculosis on digital chest radiographs: A systematic review," Int. J. Tuberc. Lung Dis. 20(9), 1226–1230 (2016).

[10] Candemir, S., Jaeger, S., Palaniappan, K., Musco, J. P., Singh, R. K., Xue, Z., Karargyris, A., Antani, S., Thoma, G. and McDonald, C. J., "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," IEEE Trans. Med. Imag. 33(2), 577–590 (2014).

[11] LeCun, Y., Bengio, Y. and Hinton, G., "Deep learning," Nature 521(7553), 436–444 (2015).

[12] Schmidhuber, J., "Deep Learning in neural networks: An overview," Neural Networks 61, 85–117 (2015).

[13] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," J. Mach. Learn. Res. 15, 1929–1958 (2014).

[14] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K. and Fei, L. F., "ImageNet: a Large-Scale Hierarchical Image Database," Proc. CVPR, 248–255 (2009).

[15] Chen, Y., Lin, Z., Zhao, X., Wang, G. and Gu, Y., "Deep Learning-Based Classification of Hyperspectral Data," IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens. 7(6), 2094–2107 (2014).

[16] Krizhevsky, A., Sutskever, I. and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," Proc. NIPS 1, 1097–1105 (2012).

[17] Simonyan, K. and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv Prepr. arXiv:1409.1556 (2014).

[18] Chollet, F., "Xception: Deep Learning with Separable Convolutions," arXiv Prepr. arXiv1610.02357 (2016).

[19] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., "Rethinking the Inception Architecture for Computer Vision," arXiv Prepr. arXiv:1512.00567 (2015).

[20] He, K., Zhang, X., Ren, S. and Sun, J., "Deep Residual Learning for Image Recognition," Proc. CVPR, 770–778 (2016).

[21] Larobina, M. and Murino, L., "Medical image file formats," J. Digit. Imaging 27(2), 200–206 (2014).

[22] Liang, Z., Powell, A., Ersoy, I., Poostchi, M., Silamut, K., Palaniappan, K., Guo, P., Hossain, M. A., Sameer, A., Maude, R. J., Huang, J. X., Jaeger, S. and Thoma, G., "CNN-based image analysis for malaria diagnosis," Proc. BIBM, 493–496 (2017).

[23] Sivaramakrishnan, R., Antani, S. and Jaeger, S., "Visualizing Deep Learning Activations for Improved Malaria Cell Classification," Proc. PMLR, 40–47 (2017).

[24] Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E. and Greenspan, H., "Chest pathology detection using deep learning with non-medical training," Proc. ISBI, 294–297 (2015).

[25] Bergamo, A. and Torresani, L., "PiCoDes: Learning a Compact Code for Novel-Category Recognition," Adv. Neural Inf. Process. Syst. 24, 2088–2096 (2011).

[26] Hwang, S., Kim, H.-E., Jeong, J. and Kim, H.-J., "A novel approach for tuberculosis screening based on deep convolutional neural networks," Proc. SPIE 9785, 97852W (2016).

[27] Jaeger, S., Candemir, S., Antani, S., Wang, Y. X., Lu, P. X. and Thoma, G., "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases.," Quant. Imaging Med. Surg. 4(6), 475–477 (2014).

[28] Liu, C., Yuen, J. and Torralba, A., "Sift flow: Dense correspondence across scenes and its applications," IEEE Trans. Pattern Anal. Mach. Intell. 33(5), 978-994 (2015).

[29] Boykov, Y. and Funka-Lea, G., "Graph cuts and efficient N-D image segmentation," Int. J. Comput. Vis. 70(2), 109–131 (2006).

[30] Choromanska, A. et al., "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," arXiv Prepr. arXiv:1502.03167 (2015).

[31] Bergstra, J. and Bengio, Y., "Random Search for Hyper-Parameter Optimization," J. Mach. Learn. Res. 13, 281–305 (2012).

[32] Boser, B. E., Guyon, I. M. and Vapnik, V. N., "A training algorithm for optimal margin classifiers," Proc. COLT, 144–152 (1992).

[33] Razavian, A. S., Azizpour, H., Sullivan, J. and Carlsson, S., "CNN features off-the-shelf: An astounding baseline for recognition," arXiv Prepr. arXiv:1403.6382 (2014).

[34] Chollet, F., "Deep Learning Models," GitHub, 27 March 2015, https://github.com/fchollet/deep-learning-models (2 February 2017).

[35] Ashraf, K., Wu, B., Iandola, F. N., Moskewicz, M. W. and Keutzer, K., "Shallow Networks for High-Accuracy Road Object-Detection," arXiv Prepr. arXiv:1606.01561 (2016).

[36] Rossi, J. S., "One-Way Anova from Summary Statistics," Educ. Psychol. Meas. 47(1), 37–38 (1987).

[37] Trawiński, B., Smętek, M., Telec, Z. and Lasota, T., "Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms," Int. J. Appl. Math. Comput. Sci. 22(4), 867–881 (2012).

[38] Shapiro, S. S. and Wilk, M. B., "An Analysis of Variance Test for Normality (Complete Samples)," Biometrika 52(3–4), 591–611 (1965).

[39] Gastwirth, J. L., Gel, Y. R. and Miao, W., "The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice," Stat. Sci. 24(3), 343–360 (2009).

[40] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M. and Thrun, S., "Dermatologist-level classification of skin cancer with deep neural networks," Nature 542(7639), 115–118 (2017).