A novel stacked generalization of models for improved TB detection in chest radiographs*

S. Rajaraman, S. Candemir, Z. Xue, P. O. Alderson, M. Kohli, J. Abuya, G. R. Thoma, and S. Antani, *Senior Member, IEEE*

***Abstract*— Chest x-ray (CXR) analysis is a common part of the protocol for confirming active pulmonary Tuberculosis (TB). However, many TB endemic regions are severely resource constrained in radiological services impairing timely detection and treatment. Computer-aided diagnosis (CADx) tools can supplement decision-making while simultaneously addressing the gap in expert radiological interpretation during mobile field screening. These tools use hand-engineered and/or convolutional neural networks (CNN) computed image features. CNN, a class of deep learning (DL) models, has gained research prominence in visual recognition. It has been shown that Ensemble learning has an inherent advantage of constructing non-linear decision making functions and improve visual recognition. We create a stacking of classifiers with hand-engineered and CNN features toward improving TB detection in CXRs. The results obtained are highly promising and superior to the state-of-the-art.**

I. Introduction

Tuberculosis (TB) is an infectious disease caused by the rod-shaped bacterium called Mycobacterium tuberculosis. According to the 2015 World Health Organization (WHO) report, TB is the most common cause of infectious disease-related mortality across the world with more than 10 million people infected and 1.8 million reported deaths that year [1]. Postero-anterior (PA) or lateral chest X-ray (CXR) [2] is the most common imaging modality used to diagnose conditions affecting the chest [3], particularly pulmonary TB. Fig. 1 shows two examples of abnormal and a normal CXR. Lack of expertise in interpreting radiology reports adversely impacts TB endemic regions, severely impairing screening efficacy [4]. Thus, there is current research interest in developing cost-effective computer-aided diagnosis (CADx) systems that could aid in that effort [5], [6]. Advancing use of CADx systems could help greatly improve the detection accuracy and alleviate human burden in screening and diagnosis, particularly in disease-endemic regions that lack sufficient radiology resources. Most CADx methods use hand-engineered features combined with a supervised learning algorithm for decision-making. Of interest is work by [5] where the authors proposed a combination of standard computer vision algorithms for extracting features from CXR images. The study segmented the lung region of interest (ROI) [7], extracted the features using a combination of algorithms that included histogram of oriented gradients (HOG), local binary patterns (LBP), Tamura feature descriptors among others. A binary support vector machine (SVM) classifier was trained on these features to classify normal and TB-positive cases. HOG descriptors together with other commonly known image descriptors, viz., GIST, Pyramidal HOG (PHOG), are also used in another study on automated TB detection [8]. Unlike algorithms using hand-engineered features, deep learning (DL) offers a hierarchical analysis of the image using a cascade of layers of non-linear processing units for "end-to-end" feature extraction and classification [9]. There are three commonly used approaches to applying DL for a visual recognition task: (i) training a model from scratch [10], [11] (ii) fine-tuning a pre-trained model (also known as, transfer learning (TL)) [12], and, (iii) using pre-trained models as feature extractors followed by training a supervised machine learning algorithm of choice [13]. Improving results from these approaches could take advantage of ensemble learning (EL) [14].

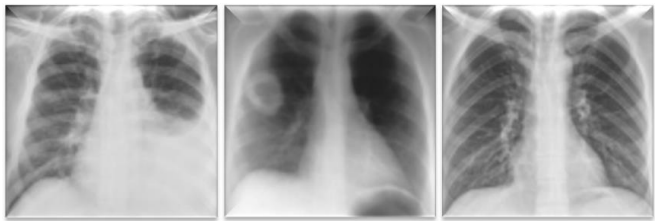

Figure 1. CXR images showing 2 examples of pulmonary abnormalities (left: pleural effusion, middle: cavitary lung lesion right lung), and normal lung image (right).

A pioneering work on EL proved that multiple, diverse and accurate *base-learners* could asymptotically fuse results to build a *strong-learner* [15]. Different fusing strategies were used to combine the decision made by the base-learners, e.g., in majority voting the base-learners vote to predict the outcome. Stacking, otherwise called *stacked generalization*, is an optimal fusing technique that highlights each base-learner that performs best and discredits others [14]. A second-level *meta-learner* optimizes the combination of individual base-learners. In [12], the authors evaluated the efficacy of an ensemble of various untrained and pre-trained deep CNNs toward TB detection in chest radiographs. As

*This research is supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

S. Rajaraman, S. Antani, S. Candemir, Z. Xue, and G. Thoma are with the Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA (e-mail: sivaramakrishnan.rajaraman@nih.gov).

P. Alderson is with Saint Louis University School of Medicine, Saint Louis, MO 63103 USA (e-mail: philip.alderson@health.slu.edu).

M. Kohli is with the Department of Radiology and Biomedical Imaging at the University of California, San Francisco, CA 94117 USA (e-mail: marc.kohli@ucsf.edu).

J. M. Abuya is with Department of Radiology and Imaging, School of Medicine, Moi University, Eldoret, Kenya (email: josephabuya@gmail.com)

reported in [16], pre-trained CNNs obtained better results than custom models on the publicly available datasets [5] also used in our study. These pre-trained models learned a comprehensive feature set, making them capable to serve as feature extractors in an extensive range of visual recognition applications. Since both DL and EL have their inherent advantages in constructing non-linear decision-making functions, the combination of the two could efficiently handle visual recognition tasks.

Our study evaluated the performance of a stacked ensemble that optimally combined classifiers using hand-engineered features with those extracted from pre-trained CNNs through two different proposals for improving the accuracy of TB detection in PA CXR images. In the first proposal, we used commonly known GIST, HOG, and SURF features extracted from CXRs and trained an SVM classifier to classify them into normal and abnormal classes. In the second proposal, we used four different pre-trained CNN models, viz., AlexNet [17], VGG-16[10], GoogLeNet [18], and ResNet-50 [19], to extract features from the CXR images and similarly trained an SVM classifier on them to detect abnormal images that exhibit "TB-like" manifestations. Finally, we performed a stacked ensemble of models from these proposals to improve the accuracy of TB detection. For this study, we used a Windows® system with Intel® ® Xeon® CPU E5-2640v3 2.60-GHz processor, 1 TB of Hard Disk space, 16 GB RAM, a CUDA-enabled Nvidia GTX 1080 Ti 11GB graphical processing unit (GPU), Matlab® R2017b, Weka® 3, and CUDA 8.0/cuDNN 5.1 dependencies for GPU acceleration. The remainder of the paper is organized as follows: Section II discusses the materials and methods, Section III discusses the results, and Section IV concludes this paper.

II. Materials and Methods

A. Datasets

This study was evaluated on four CXR datasets including two publicly available datasets provided by the U.S. National Library of Medicine (NLM), National Institutes of Health (NIH) described in [5], viz, Shenzhen and Montgomery. 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), a Kenyan NGO, and made available de-identified CXRs from rural western Kenya as a part of the mobile truck-based screening. This dataset contained 238 abnormal CXRs and 729 healthy controls. Expert radiologist annotated the images and generated zone-based clinical readings. These CXRs had pixel resolutions of either 2004×2432 or 1932×2348. The fourth dataset (India) was made available in [8]. Table 1 presents the information pertaining to the datasets and their characteristics. The acquisition and sharing of these datasets and all experimental procedures described here were approved by the NIH Institutional Review Board (IRB) (#5357).

B. Preprocessing

To enable algorithms to train on task-specific image ROI, the lungs were automatically segmented from the CXRs using method described in [7]. As shown in Fig. 2, after lung segmentation, the resulting images were cropped to the size of a bounding box that contains all the lung pixels. The cropped images were contrast-enhanced by applying Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm. For *Proposal 1*, the images were down-sampled to 3072×3072, 4096×4096, 2048×2048, and 1024×1024 pixel resolutions to obtain images with identical dimensionality for Shenzhen, Montgomery, Kenya, and India collections respectively.

TABLE I. DATASETS AND THEIR CHARACTERISTICS. DATASETS CODE: S: SHENZEN, M: MONTGOMERY, K: KENYA, I: INDIA.

	# TB positive	*# Normal*	*File Type*	*Bit Depth*	*Resolution*
S	336	326	PNG	8-bit	948-3001×1130-3001
M	58	80	PNG	8-bit	4020-4892×4020-4892
K	238	729	PNG	8-bit	1312-1852×1094-1838
I	153	153	JPG	8-bit	1024-2480×1024-2480

For *Proposal 2*, the images were down-sampled to 224×224 and 227×227 pixel resolutions to suit the input requirements for the different pre-trained CNNs, across all the datasets.

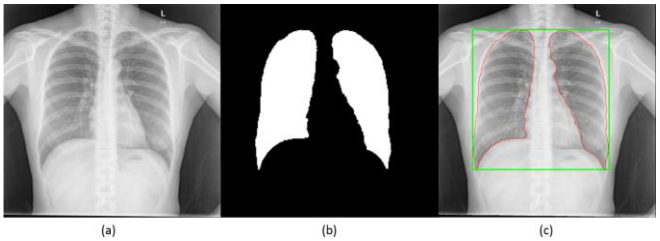

Figure 2. PA CXR lung ROI segmentation process: (a) original image, (b) computed lung mask, (c) segmented lung ROI with the bounding box.

C. Proposal 1 – Feature extraction using local and global feature descriptors and classification using SVM

Since pre-trained CNNs demand down-sampling of the underlying data to fit the specific requirements of the input layer, a lot of potentially viable information pertaining to the signs of TB infection may be lost. The best way to overcome this issue is to use the local and global feature descriptors that extract discriminative information from the entire CXR image without the need for down-sampling, a serious constraint with the usage of DL models. For this reason, local and global feature descriptors were used in this proposal (P1). We evaluated the performance of global image descriptors including GIST, HOG and local descriptors including SURF toward identifying TB manifestations. We performed nested cross-validation where in the outer loop, we performed 5-fold cross-validation for all the datasets. In the inner loop, we performed Bayesian optimization [20] to minimize 5-fold cross-validation error by varying the parameters for the SVM classifier that included the box constraint, kernel scale, kernel function, and order of the polynomial. The chosen ranges included [1e-3 1e3], [1e-3 1e3], and [2 4] for box constraint, kernel scale and order of the polynomial respectively. For the kernel function, the optimization process searched among linear, Gaussian, RBF and Polynomial kernels.

D. Proposal 2 – Feature extraction using pre-trained CNNs and classification using SVM

In the second proposal (P2), we evaluated the performance of CNN models pre-trained on the large-scale ImageNet dataset (listed earlier) in extracting the features from the CXR images across the normal and TB-positive categories. The segmented ROI constituting the lungs were down-sampled to match the input dimensions of the pre-trained CNNs. Literature study revealed that the features were typically extracted from the one, right before the classification layer [21] to train a classifier. The validation was conducted as in P1.

E. Stacked generalization of models

Finally, we created a stacked generalization of models from the proposals above to improve the accuracy of TB detection. Stacked ensembles were created for the models in P1 labeled (E [P1]), and P1 and P2, labeled (E [P1P2]) in the result tables below. The meta-learner used was a Logistic regression (LR) classifier that estimated the probability for a binary response. Fig. 3 shows the block diagram of the proposed stacked ensemble.

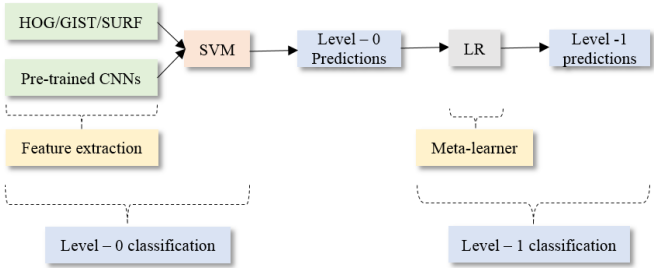

Figure 3. The stacked ensemble of models from the proposals.

III. Results and Discussion

Table 2 shows the results obtained with different feature descriptors used in P1 in terms of accuracy and area under ROC curve (AUC). For the Shenzhen dataset ("S"), best results were obtained with the GIST features and SVM/RBF classifier with an accuracy of 0.845 and AUC of 0.921. For the Montgomery dataset ("M"), BOW model using SURF and SVM/RBF classifier demonstrated superior performance with an accuracy of 0.775 and AUC of 0.845. For Kenya dataset ("K"), HOG features and SVM/Gaussian classifier showed better performance in terms of accuracy but GIST features and SVM/RBF classifier gave the best AUC of 0.748. For India dataset ("I"), GIST features and SVM/RBF classifier demonstrated superior performance with an accuracy of 0.882 and AUC of 0.961. We observed that no feature descriptor performed equally well. The CXR images varied across the datasets in source machinery and pixel resolutions. The local and global feature descriptors were rule-based feature extraction mechanisms that were built and optimized to improve performance with the individual datasets. It may be for this reason that they did not perform equally well across the datasets. It can be noted here that the results obtained with the India dataset were superior to those obtained on the other datasets. A similar pattern was observed in the result tables for different proposals. Though the India collection is sparse, TB manifestations were obvious and distributed throughout the lungs that gave the feature descriptors the opportunity to capture highly discriminative features across the normal and abnormal categories. Least performance was observed with the Kenya dataset, the reason being the dataset had a highly imbalanced distribution of instances across the classes with 238 abnormal CXRs in comparison to 729 healthy controls. The patients were all HIV+ with a low-immune response. The expression of the disease even for severe cases was significantly weaker than ordinary TB. They were cassette-based radiographic images obtained as a result of mobile truck-based screening and hence the image resolution was not commendable that further impaired the performance of feature extraction and classification.

TABLE II. SVM-BASED CLASSIFICATION OF HAND-ENGINEERED FEATURES (ACC = ACCURACY, AUC = AREA UNDER ROC CURVE). CODES SAME AS TABLE I.

	HOG		GIST		SURF	
	Acc	*AUC*	*Acc*	*AUC*	*Acc*	*AUC*
S	0.841	0.917	**0.845**	**0.921**	0.816	0.890
M	0.708	0.772	0.750	0.817	**0.775**	**0.845**
K	**0.683**	0.741	0.667	**0.748**	0.672	0.747
I	0.880	0.947	**0.882**	**0.961**	0.864	0.938

With Montgomery dataset, performance limitation may be attributed to the limited size of the dataset and the degree of imbalance across the classes where 40% of the samples were TB-positive as compared to 60% of healthy controls. Table 3 presents the results of the second proposal using pre-trained CNNs for feature extraction and SVM for the classification task. For Shenzhen dataset, AlexNet obtained the best accuracy of 0.859 and AUC of 0.924. The same pattern was observed across Montgomery, Kenya and India datasets. For Montgomery dataset, AlexNet obtained the best accuracy of 0.725 and AUC of 0.817. For India dataset, AlexNet outperformed the other pre-trained CNNs with an accuracy of 0.872 and AUC of 0.950. For the Kenya dataset, we observed that the AUC of VGG-16 was slightly better than that of AlexNet, however, the accuracy of AlexNet was higher than that of the other pre-trained CNNs. The results obtained with the India dataset were superior to the results obtained with the other datasets for the reasons discussed earlier.

TABLE III. PRE-TRAINED CNNS BASED FEATURE EXTRACTION AND SVM BASED CLASSIFICATION. CODES SAME AS TABLE I.

	AlexNet		VGG-16		GoogLeNet		ResNet-50	
	Acc	*AUC*	*Acc*	*AUC*	*Acc*	*AUC*	*Acc*	*AUC*
S	**0.859**	**0.924**	0.829	0.901	0.768	0.870	0.819	0.893
M	**0.725**	**0.817**	0.717	0.757	0.678	0.648	0.676	0.616
K	**0.693**	0.776	0.691	**0.777**	0.674	0.750	0.678	0.753
I	**0.872**	**0.950**	0.812	0.892	0.796	0.888	0.812	0.902

Among pre-trained CNNs evaluated in this study, a shallow model like AlexNet outperformed other models across the datasets. It was surprising that deep models like ResNet-50 and GoogLeNet did not perform better than shallow models. The architecture depth of ResNet-50 and GoogLeNet seemed adverse to this task of binary medical image classification. For ImageNet data, deeper networks outperformed shallow counterparts for the reason that the data was diverse and the models learned abstractions for a

huge selection of classes [22]. In our case, for the binary task of TB detection from CXR datasets, the variability in data was several orders of magnitude smaller where deeper networks did not seem to be a fitting tool. Also, the top layers of the pre-trained CNNs like GoogLeNet and ResNet-50 were probably too specialized, progressively more complex and not the best candidate to re-use for the task of our interest. Table 4 demonstrates the advantage of using a stacked ensemble of models where the ensemble of all proposals (E [P1P2]) had the highest AUC across all datasets. Table 5 compares the results obtained across the ensembles of different proposals presented in this study and top ranking relevant literature on TB detection [5], [8], [13], and [16]. In terms of accuracy, the stacked ensemble E [P1P2] outperformed the results presented in the literature. The proposed ensemble demonstrated the highest accuracy of 0.960 for India, followed by 0.934 for Shenzhen, 0.875 for Montgomery, and 0.776 for the Kenya dataset. Accuracy across different ensembles remained the same with the value of 0.875 and 0.960 for Montgomery and India datasets respectively. We observed similar patterns with AUC values where E [P1P2] demonstrated high AUC values than the state-of-the-art. The results for Shenzhen dataset were superior with an AUC of 0.991, followed by 0.965 for India, 0.962 for Montgomery and 0.826 for Kenya dataset. As observed from these results, E [P1P2] achieved promising results across the datasets than the state-of-the-art.

TABLE IV. THE ENSEMBLE OF MODELS FROM DIFFERENT PROPOSALS. CODES SAME AS TABLE I.

	E[P1]		**E[P1P2]**	
	Acc	***AUC***	***Acc***	***AUC***
S	0.934	0.955	0.934	**0.991**
M	**0.**875	0.875	**0.**875	**0.962**
K	0.733	0.825	**0.776**	**0.826**
I	0.960	0.960	0.960	**0.965**

TABLE V. COMPARISON OF THE PROPOSED ENSEMBLES WITH RESULTS FROM THE LITERATURE. CODES SAME AS TABLE I.

		Literature				**Proposed approaches**	
		[5]	***[16]***	***[13]***	***[8]***	***E[P1]***	***E[P1P2]***
S	**Acc**	0.840	0.837	0.847	-	**0.934**	**0.934**
	AUC	0.900	0.926	0.926	-	0.955	**0.991**
M	**Acc**	0.783	0.674	0.826	-	**0.875**	**0.875**
	AUC	0.869	0.884	0.926	-	0.875	**0.962**
K	**Acc**	-	-	-	-	0.733	**0.776**
	AUC	-	-	-	-	0.825	**0.826**
I	**Acc**	-	-	-	0.943	**0.960**	**0.960**
	AUC	-	-	-	0.960	0.960	**0.965**

IV. CONCLUSION

We have discussed different proposals for improving the performance of TB detection. The goal of the proposed method is not to develop the most computationally efficient system. However, one can estimate the computation time as a combination of individual proposals. Based on the results obtained in the present work, we believe that the stacked ensemble of models using local and global feature descriptors and pre-trained CNNs could be a promising option for improving the detection accuracy. An appealing use case is to apply this method in applications with sparse data, particularly in biomedical imagery.

ACKNOWLEDGMENT

This work is supported by the Intramural Research Program of National Library of Medicine (NLM), National Institutes of Health (NIH) and Lister Hill National Center for Biomedical Communications (LHNCBC).

REFERENCES

[1] World Health Organization, Global tuberculosis report, 2015.

[2] M. F. Iademarco, J. O'Grady, K. Lönnroth, "Chest radiography for tuberculosis screening is back on the agenda," *Int J Tuberc Lung Dis.*, vol. 16, no. 11, pp.1421-1422, Nov. 2012.

[3] F. A. Mettler et al., "Effective doses in radiology and diagnostic nuclear medicine: a catalog," *Radiology*, vol. 248, no. 1, pp. 254-263, Jul. 2008.

[4] J. Melendez et al., "An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information," *Sci. Rep.*, vol. 6, no. 1, pp. 25265, Apr. 2016.

[5] S. Jaeger et al., "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant Imaging Med Surg*., vol. 4, no. 6, pp. 475–477, Dec. 2014.

[6] P. Maduskar et al., "Detection of tuberculosis using digital chest radiography: automated reading vs. interpretation by clinical officers," *Int J Tuberc Lung Dis*., vol. 17, no. 12, pp. 1613-1620, Dec. 2013.

[7] S. Candemir et al., "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Trans. Med. Imaging*, vol. 33, no. 2, pp. 577-590, Feb. 2014.

[8] A. Chauhan, D. Chauhan, C. Rout, "Role of gist and PHOG features in computer-aided diagnosis of tuberculosis without segmentation," *PLoS One*, vol. 9, no. 11, pp. e112980, Nov. 2014.

[9] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, no. 1, pp. 85-117, Jan. 2015.

[10] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv*: 1409.1556 [cs.CV], Apr. 2015.

[11] R. Sivaramakrishnan et al., "Visualizing abnormalities in chest radiographs through salient network activations in Deep Learning", in *Proc. 1st IEEE International Life Sciences Conference*, Australia, 2017, pp. 71-74.

[12] P. Lakhani, B. Sundaram, "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks," *Radiology*, vol. 284, no. 2, pp. 574-582, Aug. 2017.

[13] U. K. Lopes, J. F.Valiati, "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection," *Comput. Biol. Med.*, vol. 89, no. 1, pp. 135-143, Oct. 2017.

[14] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241-259, Dec. 1992.

[15] L. K. Hansen, P. Salamon, "Neural Network Ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993-1001, Oct. 1990.

[16] S. Hwang et al., "A novel approach for tuberculosis screening based on deep convolutional neural networks," in *Proc. SPIE Medical Imaging*, San Diego, 2016, pp. 97852W.

[17] A. Krizhevsky, I. Sutskever, H. Geoffrey, "ImageNet Classification with Deep Convolutional Neural Networks," *26th Ann. Conf. NIPS*, Nevada, 2012, pp. 1097-1105.

[18] C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," *arXiv*: 1512.00567 [cs.CV], Dec. 2015.

[19] K. He et al., "Deep Residual Learning for Image Recognition," *arXiv*: 1512.03385 [cs.CV], Dec. 2015.

[20] J. Močkus, *On bayesian methods for seeking the extremum*. Heidelberg, Berlin: Springer, 1974, pp. 401-404.

[21] A. S. Razavian et al., "CNN Features Off-the-Shelf: An astounding baseline for recognition," *arXiv*: 1403.6382 [cs.CV], May 2014.

[22] K. Ashraf et al., "Shallow Networks for High-Accuracy Road Object-Detection," *arXiv*: 1606.01561 [cs.CV], Jun. 2016.