Assessment of an ensemble of machine learning models toward abnormality detection in chest radiographs

S. Rajaraman, S. Sornapudi, *Student Member, IEEE*, M. Kohli, and S. Antani, *Senior Member, IEEE*

Abstract— **Respiratory diseases account for a significant proportion of deaths and disabilities across the world. Chest X-ray (CXR) analysis remains a common diagnostic imaging modality for confirming intra-thoracic cardiopulmonary abnormalities. However, there remains an acute shortage of expert radiologists, particularly in under-resourced settings, resulting in severe interpretation delays. These issues can be mitigated by a computer-aided diagnostic (CADx) system to supplement decision-making and improve throughput while preserving and possibly improving the standard-of-care. Systems reported in the literature or popular media use handcrafted features and/or data-driven algorithms like deep learning (DL) to learn underlying data distributions. The remarkable success of convolutional neural networks (CNN) toward image recognition tasks has made them a promising choice for automated medical image analyses. However, CNNs suffer from high variance and may overfit due to their sensitivity to training data fluctuations. Ensemble learning helps to reduce this variance by combining predictions of multiple learning algorithms to construct complex, non-linear functions and improve robustness and generalization. This study aims to construct and assess the performance of an ensemble of machine learning (ML) models applied to the challenge of classifying normal and abnormal CXRs and significantly reducing the diagnostic load of radiologists and primary-care physicians.**

I. Introduction

Respiratory diseases are reported to be the leading cause of mortality and disability globally [1]. Chest X-ray (CXR) evaluation by expert radiologists remains a routine protocol to diagnose intra-thoracic cardiopulmonary disorders. In resource-constrained settings, however, acute radiologist shortage leads to delayed interpretation and severe backlogs in patient care. There is a high research interest in developing automated computer-aided diagnostic (CADx) tools to supplement radiological interpretation [2]. A study of the literature reveals several works pertaining to the use of handcrafted features toward classifying abnormalities in CXRs [3]–[5]. These handcrafted features are built and optimized to improve performance with individual datasets. The performance suffers from variability in source machinery, morphology, background, orientation, and position of the region of interest (ROI) [6]. Unlike handcrafted feature descriptors, data-driven methods including deep learning (DL) learn hierarchical feature representations directly from the underlying data without the need for manual feature extraction [7]. These methods deliver promising results in classifying highly heterogeneous images in the presence of a large amount of annotated data [8]. DL models have shown promise in detecting lung nodules [9], TB [10], image retrieval [11] and other applications. In classifying CXRs, these models outperformed conventional feature extraction and classification methods [12]. While the existing literature serves as a substantial proof of concepts, the methods discussed are not generalizable to different kinds of abnormalities. This is because new training data and associated labels need to be obtained to train the models toward learning each specific abnormality. As well, there exists a high inter-class variability and intra-class similarity across the thoracic abnormalities that do not have a defined shape and boundary. This makes it difficult for the models to predict the disease labels only from the CXRs in the absence of associated radiological reports. In real-world applications, it is a challenging task to train models and achieve generalization on datasets with varying distributions. Under these circumstances, improving results from multiple ML approaches could benefit from ensemble learning (EL) [13]. The process helps in combining the predictions of multiple, high-performing, less-correlated models called base-learners to create a robust system with improved performance and generalization. An ensemble could be performed in various ways that include: a) averaging; b) majority voting; and c) weighted averaging [14]. Averaging the models' predictions deliver promising results on a wide range of metrics and problems [15]. Majority voting considers the predictions with maximum recommendation/vote from base-learners while predicting the final outcome [16]. Weighted averaging improves generalization by assigning higher weights to more accurate base-learners [17].

Goal: This work aims to simplify the analysis in a binary triage classification problem that classifies CXRs into normal and abnormal categories. We evaluate the performance of different ensemble strategies that combine predictions of ML

S. Rajaraman and S. Antani are with the National Library of Medicine, Bethesda, MD 20894, USA (e-mail: sivaramakrishnan.rajaraman@nih.gov).

S. Sornapudi is with the Department of Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA (e-mail: ssbw5@mst.edu).

M. Kohli is with the Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA 94143, USA (e-mail: Marc.Kohli@ucsf.edu).

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

classifiers trained with handcrafted/CNN-extracted features toward the current task. We extract HOG and LBP features and train a binary SVM classifier on the extracted features. We use DL models including a custom CNN, pretrained VGG16, and VGG19 [18] to learn hierarchical feature representations from the CXRs. Finally, we combine the predictions of individual base-learners through different ensemble strategies including majority voting, simple averaging, and weighted averaging to observe for a possible improvement in performance.

We note that the Kaggle Pneumonia detection challenge, organized by the joint effort of radiologists from the Radiological Society of North America (RSNA) and Society of Thoracic Radiology (STR) aims to predict the presence/absence of pneumonia in a given CXR. This is done by categorizing the data into pneumonia and non-pneumonia classes, the latter includes both normal, and abnormal images with lung opacities that are not related to pneumonia. This is distinct from our goal that aims to classify CXRs as normal versus abnormal, with the intent of serving as a triage for global health applications with a special interest in applicability in resource-challenged settings.

We use a Linux Ubuntu System with Nvidia GTX 1080 Ti GPU and CUDA/cuDNN dependencies for GPU acceleration. The remainder of this study is organized as follows: Section II elaborates on the materials and methods, Section III discusses the results, and Section IV concludes this report of the study.

II. Materials and Methods

A. Datasets

The Kaggle pneumonia detection challenge (https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data) dataset was used in this study. The dataset includes images with pulmonary opacities that may represent pneumonia and other images that are normal and those without a pulmonary opacity suspicious for pneumonia. The distribution of data across the classes is tabulated in Table 1. All images were of 1024×1024 pixel dimensions with 8-bit depth. Institutional Review Board (IRB) do not apply since the dataset has been de-identified and made publicly available.

TABLE I. DATASET AND ITS CHARACTERISTICS.

# Abnormal	*# Normal*	*File type*	*Bit-depth*
17833	8851	DICOM	8

B. Preprocessing

The lung ROI is segmented using the all-dropout UNET (AD-UNET) [19] to help the base-learners learn relevant information toward arriving at the predictions. After lung segmentation, the resulting images are cropped to the size of a bounding box containing all the lung pixels and resized to 224×224 pixel dimensions for further study.

C. Feature extraction using LBP/HOG and classification using SVM

In LBP based feature extraction, the local textural representation is obtained by the comparison of each pixel with its surrounding neighbors [20]. While extracting HOG features, the input images are divided into cells and features are computed for the pixels within each cell [21]. The histograms across multiple cells are accumulated to form the final feature vector. The extracted LBP and HOG feature descriptors are used to train an SVM classifier with a radial basis kernel (RBF) and the predictions are recorded. We used the original images of 1024×1024 pixel dimensions for extracting HOG and LBP features.

D. Sequential CNN model

We designed a sequential CNN as the baseline for the current task. The model has a linear stack of convolutional, ReLU, max-pooling, and dense layers. The global average pooling (GAP) layer computes the average of each feature map in the deepest convolutional layer. A Softmax [7] probabilistic classifier regularizes the outputs to the interval [0, 1] and assigns a probability to each image category. Fig. 1 shows the architecture of the custom CNN used in this study.

E. Feature extraction and classification using pretrained VGG16 and VGG19 models

We used the pretrained VGG16 and VGG19 models and customized them for the task under study as shown in Fig. 2. The models are truncated at the deepest convolutional layer, a GAP and dense layer are added to predict on the outcome. The VGG16 model is trained end to end to learn CXR-specific feature representations and categorize them to their respective classes. The VGG19 model is instantiated with the convolutional base and loaded with the pretrained weights. The activation maps before the dense, fully-connected layers are extracted and a dense model is trained on top of the stored features.

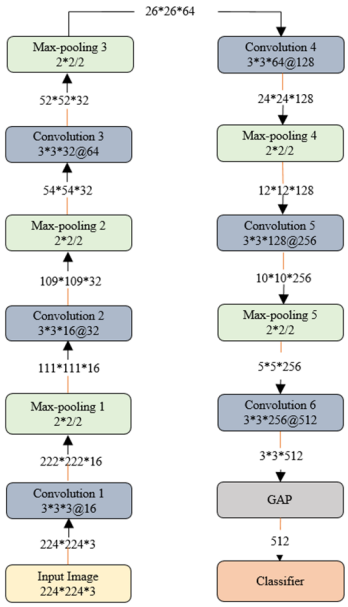

Figure 1. The architecture of the customized CNN.

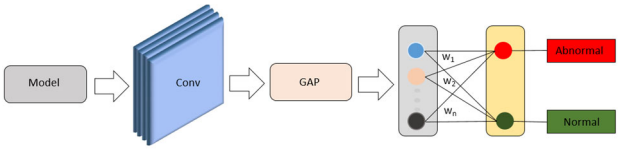

Figure 2. The customized architecture of VGG16 and VGG19 models.

We performed a randomized grid search [22] to obtain the optimal values for the hyperparameters including learning rate, momentum, and L2-regularization for the DL models under study. The search ranges are initialized to [1e-7, 1e-1], [0.8, 0.99], and [1e-9, 1e-1] for the learning rate, momentum, and L2-penalty respectively. The models' performance is evaluated in terms of accuracy, the area under the ROC curve (AUC), F-score and Matthews Correlation Coefficient (MCC).

F. Ensemble learning

We performed multiple ensembles of the predictions of individual base-learners through averaging, majority voting, and weighted averaging strategies to classify the CXRs into normal and abnormal categories. Fig. 3 shows the block diagram of the proposed ensembles.

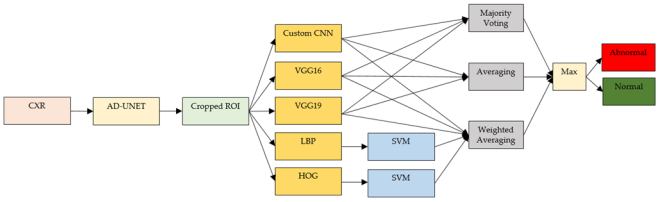

Figure 3. Ensemble learning for CXR classification.

III. Results and Discussion

The results obtained with LBP/SVM and HOG/SVM are shown in Table 2. We performed 5-fold cross-validation and presented the results in terms of mean and standard deviation (SD).

TABLE II. SVM-BASED CLASSIFICATION OF HANDCRAFTED FEATURES.

Fold	*LBP*				*HOG*			
	Acc	***AUC***	***F***	***MCC***	***Acc***	***AUC***	***F***	***MCC***
1	87.9	86.6	91.1	72.6	96.5	95.1	97.5	92.0
2	87.8	86.6	91.1	72.2	96.1	94.5	97.3	91.1
3	87.7	86.2	90.9	72.0	96.4	94.9	97.5	91.8
4	86.7	85.5	90.3	70.4	96.0	93.5	97.2	90.7
5	88.7	87.2	91.7	74.0	96.4	94.9	97.5	91.7
Mean	87.8	86.3	91.1	72.3	96.3	94.6	97.4	91.5
SD	0.65	0.63	0.50	1.29	0.22	0.65	0.14	0.54

As observed from Table 2, HOG descriptors gave promising results across the folds than LBP. The accuracy of the HOG/SVM model is 96.3±0.22 in comparison to 87.8±0.65 achieved with the LBP/SVM. The AUC, F-score, and MCC values for the HOG/SVM outperformed that of LBP/SVM by achieving 94.6±0.65, 97.4±0.14, and 91.5±0.54 respectively. The performance of custom CNN and pretrained VGG16 are tabulated in Table 3. The optimal values of hyperparameters including learning rate, momentum, and L2 penalty are found to be [1e-2, 0.95, 1e-7], [1e-4, 0.99, 1e-6], and [1e-4, 0.99, 1e-6] for the custom CNN, VGG16 and VGG19 models under study. We observed that the VGG16 demonstrated promising performance than the custom CNN under study. The generic feature representations learned from the ImageNet data served as a good initialization that assisted the model in faster convergence with reduced overfitting and improved generalization. The custom CNN is initialized with random weights. The custom model didn't optimally learn discriminative feature representations owing to the imbalanced distribution of samples across the normal and abnormal categories.

Table 4 shows the performance of the VGG19 model, trained as a classifier toward the current task. As observed, the performance is not promising like the other methods. This may be attributed to several reasons: a) The architecture depth of VGG19 appears adverse to this binary classification task; b) The data variability is several orders of magnitude smaller in comparison to ImageNet and deeper networks do not appear to be a fitting tool; and c) The top layers of VGG19 are probably too specialized and progressively more complex and may not be the best candidate to be used for the current task.

TABLE III. PERFORMANCE METRICS OF CUSTOMIZED CNN AND PRETRAINED VGG16 MODELS.

Fold	*Custom CNN*				*VGG16*			
	Acc	***AUC***	***F***	***MCC***	***Acc***	***AUC***	***F***	***MCC***
1	93.4	97.8	95.2	85.2	97.5	99.8	98.2	94.2
2	93.1	97.7	95.1	84.0	96.8	99.8	97.8	92.6
3	94.3	98.3	95.9	87.0	96.8	99.7	97.8	92.7
4	92.3	98.1	94.5	82.2	97.2	99.8	98.0	93.5
5	92.5	97.8	94.6	82.5	97.1	99.8	98.0	93.3
Mean	93.2	98.0	95.1	84.2	97.1	99.8	98.0	93.3
SD	0.81	0.23	0.56	1.99	0.30	0.05	0.17	0.65

TABLE IV. PERFORMANCE METRICS OF THE PRETRAINED VGG19 MODEL.

Fold	*VGG19*			
	Acc	***AUC***	***F***	***MCC***
1	87.5	93.3	91.0	71.2
2	87.9	93.6	91.3	71.7
3	88.7	86.2	91.9	73.7
4	87.9	93.9	91.4	71.7
5	87.9	93.8	91.3	71.7
Mean	88.0	92.2	91.4	72.0
SD	0.43	3.36	0.33	0.97

The results of ensemble strategies are tabulated in Table 5 and Table 6. In weighted averaging, we awarded high/low importance to the predictions by assigning higher weights to more accurate base-learners. We found that the VGG16 outperformed the other methods toward the current task. Thus, we assigned weights of [0.1, 0.4, 0.1, 0.1, 0.3] to the predictions of the custom CNN, VGG16, VGG19, LBP/SVM, and HOG/SVM models respectively. We observed that weighted averaging outperformed majority voting and simple averaging ensembles by achieving an accuracy of 98.7±0.078, AUC of 100±0.02, F-score of 99.1±0.05, and MCC of 96.8±0.18.

TABLE V. PERFORMANCE METRICS OF MAJORITY VOTING AND SIMPLE AVERAGING ENSEMBLES.

Fold	*Majority voting*			*Averaging*			
	Acc	*F*	*MCC*	*Acc*	*AUC*	*F*	*MCC*
1	98.2	98.7	95.8	98.7	99.9	99.1	97.1
2	97.6	98.3	94.5	98.3	99.9	98.8	95.9
3	98.0	98.6	95.3	98.4	99.9	98.9	96.3
4	97.5	98.3	94.3	98.3	100.0	98.8	96.1
5	97.7	98.4	94.6	98.5	100.0	99.0	96.5
Mean	97.8	98.5	94.9	98.5	100.0	99.0	96.4
SD	0.27	0.18	0.63	0.18	0.02	0.13	0.46

TABLE VI. PERFORMANCE METRICS OF THE WEIGHTED AVERAGING ENSEMBLE.

Fold	*Weighted averaging*			
	Acc	*AUC*	*F*	*MCC*
1	98.7	100.0	99.1	97.0
2	98.6	100.0	99.1	96.8
3	98.6	100.0	99.1	96.7
4	98.5	99.9	99.0	96.5
5	98.6	100.0	99.0	96.8
Mean	98.7	100.0	99.1	96.8
SD	0.78	0.02	0.05	0.18

IV. CONCLUSION

We conclude that the weighted averaging of the predictions of individual base-learners significantly improves the performance toward the challenge of classifying normal and abnormal CXRs. We also observed that the winning solution in the Kaggle pneumonia detection challenge used an ensemble of pretrained CNN models toward detecting pneumonia in CXRs. Model ensembles provide a reliable solution by making a combined prediction where the final accuracy is promising than that of the individual learners. The ensemble promises to be a functional classification framework and would serve as a triage, particularly in resource-constrained settings to supplement diagnosis and improve patient treatment.

V. ACKNOWLEDGMENTS

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).

REFERENCES

[1] The Forum of International Respiratory Societies. The Global Impact of Respiratory Disease, 2017.

[2] S. Rajaraman et al., "Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs," Appl. Sci. vol. 8, no. 10, pp. 1715, Sep. 2018.

[3] S. Jaeger, S. Candemir, S. Antani, Y. X. J. Wang, P. X. Lu, and G. Thoma, "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.

[4] 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.

[5] A. Chauhan, D. Chauhan, and C. Rout, "Role of gist and PHOG features in computer-aided diagnosis of tuberculosis without segmentation," PLoS One, vol. 9, no. 11, pp. 1–12, Nov. 2014.

[6] M. I. Neuman et al., "Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children," J. Hosp. Med., vol. 7, no. 4, pp. 294–298, Apr. 2012.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017.

[8] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *Proc. 2017 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*., pp. 3462–3471.

[9] K. L. Hua, C. H. Hsu, S. C. Hidayati, W. H. Cheng, and Y. J. Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," Onco. Targets Ther., vol. 8, pp. 2015–2022, Aug. 2015.

[10] S. Rajaraman et al., "A novel stacked generalization of models for improved TB detection in chest radiographs," in *2018 Conf Proc IEEE Eng Med Biol Soc.,* pp. 718–721.

[11] A. Rajkomar, S. Lingam, A. G. Taylor, M. Blum, and J. Mongan, "High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks," J. Digit. Imaging, vol. 30, no. 1, pp. 95–101, Feb. 2017.

[12] P. Lakhani and 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, Apr. 2017.

[13] T. G. Dietterich, "Ensemble Methods in Machine Learning," in Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 1857, Heidelberg: Springer, 2000.

[14] F. Schwenker, "Ensemble Methods: Foundations and Algorithms [Book Review]," IEEE Comput. Intell. Mag., vol. 8, no. 1, pp. 77–79, Feb. 2013.

[15] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, "Using Bayesian Model Averaging to Calibrate Forecast Ensembles," Mon. Weather Rev., vol. 133, no. 5, pp. 1155–1174, May 2005.

[16] D. Ruta and B. Gabrys, "Classifier selection for majority voting," Inf. Fusion, vol.6, no. 1, pp. 63–81, Mar. 2005.

[17] O. G. Guleryuz, "Weighted averaging for denoising with overcomplete dictionaries," IEEE Trans. Image Process., vol. 16, no. 12, pp. 3020–3034, Dec. 2007.

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

[19] A. A. Novikov, D. Lenis, D. Major, J. Hladuvka, M. Wimmer, and K. Buhler, "Fully Convolutional Architectures for Multiclass Segmentation in Chest Radiographs," IEEE Trans. Med. Imaging, vol. 37, no. 8, pp. 1865–1876, Aug. 2018.

[20] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[21] N. Dalal and W. Triggs, "Histograms of Oriented Gradients for Human Detection," 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR05, vol. 1, no. 3, pp. 886–893, 2004.

[22] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," J. Mach. Learn. Res., vol. 13, no. 1, pp. 281–305, Jan. 2012.