Multi-View Ensemble Convolutional Neural Network to Improve Classification of Pneumonia in Low Contrast Chest X-Ray Images

José Raniery Ferreira Junior[1,#], Diego Armando Cardona Cardenas[1,#], Ramon Alfredo Moreno[1], Marina de Fátima de Sá Rebelo[1], José Eduardo Krieger[1], Marco Antonio Gutierrez[1]

***Abstract*— Pneumonia is one of the leading causes of childhood mortality worldwide. Chest x-ray (CXR) can aid the diagnosis of pneumonia, but in the case of low contrast images, it is important to include computational tools to aid specialists. Deep learning is an alternative because it can identify patterns automatically, even in low-resolution images. We propose herein a convolutional neural network (CNN) architecture with different training strategies towards detecting pneumonia on CXRs and distinguishing its subforms of bacteria and virus. We also evaluated different image pre-processing methods to improve the classification. This study used CXRs from pediatric patients from a public pneumonia CXR dataset. The pre-processing methods evaluated were image cropping and histogram equalization. To classify the images, we adopted the VGG16 CNN and replaced its fully-connected layers with a customized multilayer perceptron. With this architecture, we proposed and evaluated four different training strategies: original CXR image (baseline), chest-cavity-cropped image (A), and histogram-equalized segmented image (B). The last strategy method (C) implemented is based on ensemble between strategies A and B. The performance was assessed by the area under the ROC curve (AUC) with 95% confidence interval (CI), accuracy, sensitivity, specificity, and F1-score. The ensemble model C yielded the highest performances: AUC of 0.97 (CI: 0.96–0.99) to classify pneumonia *vs*. normal, and AUC of 0.91 (CI: 0.88–0.94) to classify bacterial *vs*. viral cases. All models that used pre-processed images showed higher AUC than baseline, which used the original CXR image. Image cropping and histogram equalization reduced irrelevant information from the exam, enhanced contrast, and was able to identify fine CXR texture details. The proposed ensemble model increased the representation of inflammatory patterns from bacteria and viruses with few epochs to train the deep CNNs.**

***Clinical relevance*— Deep learning can identify complex radiographic patterns in low contrast images due to pneumonia and distinguish its subforms of bacteria and virus. The correlation of imaging with lab results could accelerate the adoption of complementary exams to confirm the disease's cause.**

I. INTRODUCTION

The World Health Organization (WHO) states that pneumonia is a major pediatric problem and one of the leading causes of childhood mortality worldwide, especially in Africa, South America, and Southeast Asia [1][2]. At least 90% of newly diagnosed cases occur in those developing regions where medical resources are limited, and every year about two million children under five years old die due to pneumonia [3]. Bacterial and viral microorganisms are the most common etiologic agents responsible for community-acquired pneumonia, but the identification of those pathogens remains challenging [4]. Moreover, bacteria and virus need different treatments, but the former may require more urgent referral due to antibiotic intervention [2].

Chest x-ray (CXRs) is a low-risk and accessible exam that represents an essential component to evaluate patients with a suspicion of pneumonia. Radiography, along with computed tomography, can aid the diagnosis of pneumonia in conjunction with clinical and laboratorial data, according to the American Thoracic Society (ATS) [5]. The ATS also recommends CXRs to assess the extent of the disease and to detect complications (e.g., abscess formation) [4]. However, the detection of pneumonia in CXRs is still largely dependent on the skills of physicians, and it is not always possible to produce the image reports quickly as it relies on the availability of expert radiologists [3].

Furthermore, radiographic patterns (like opacities) of pneumonia are often related to the causative agent. Bacterial pneumonia typically exhibits a focal nonsegmental, homogenous lobar inflammation consolidation (focal opacities), whereas viral pneumonia generally manifests as more diffuse bilateral interstitial or interstitial-alveolar patterns in both lungs [2][4]. The radiographic appearance of it can overlap with other diseases, and it can mimic other lung consolidations and opacities due to low contrast of CXR, especially from children because of the dose of radiation received by the patient is relatively low, under normal circumstances [3][6]. Therefore, it is vital to include computer-based tools to aid physicians in detecting diseases early and potentially provide further information such as the type of the infection (i.e., bacterial or viral), as they can improve the accuracy and consistency of medical image diagnosis through computational support used as a reference [7].

Some works have used computational tools to classify pneumonia patients and normal subjects. For instance, Sousa et al. [8] extracted wavelet features from CXRs and used as input to three different machine learning methods. Chandra et al. [9] extracted histogram features and used five different image classifiers. However, they used hand-crafted image features, which is a time-consuming and labor-intensive task. Recently, the use of deep learning has been gaining importance in solving this type of problem. Deep learning, in particular convolutional neural networks (CNNs), is a machine learning branch that uses raw data (i.e., image pixels) as the algorithm input and abstracts layer-wise the original imaging data into the final feature vector without

[1]Heart Institute, Clinics Hospital, University of Sao Paulo Medical School, Av. Dr. Enéas de Carvalho, 44 - 05403-900 - São Paulo-SP, Brazil.
[#]These authors contributed equally to this work.
Corresponding author: jose.raniery@incor.usp.br

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

requiring manual procedures [2]. The automated pattern recognition and classification based on deep learning could significantly mitigate problems caused by visual assessment (e.g., subjectivity and time) and improve the efficiency of specialists, reduce medical costs, and support the diagnosis and treatment decisions of pediatric pneumonia [3][7].

In this context, this work evaluated the performance of deep learning methods in the detection of pneumonia and to distinguish between viral and bacterial pneumonia. Due to low contrast between soft tissues and the nature of the CXR exam to show overlapped structures, we hypothesize that medical image pre-processing may enhance and highlight CXR features to improve the classification tasks. Therefore, we also aim to evaluate image processing methods to improve recognition of radiographic patterns of pneumonia.

II. MATERIAL AND METHODS

A. Pneumonia CXR Dataset

This study used anteroposterior images from patients of one to five years old from a public dataset (the Guangzhou Pneumonia Chest X-ray dataset) [2], and hence, no Institutional Review Board approval was needed. CXR images were used in experiments according to the original dataset split training set: 2,538 bacterial pneumonia, 1,345 viral, and 1,349 without findings; testing set: 242 bacterial pneumonia, 148 viral pneumonia, and 234 without findings [2].

B. Image Pre-Processing

One of our hypotheses is that removing anatomical regions from the CXR that are not relevant to pneumonia detection would improve the training ability of deep CNNs, and consequently, the classification performance. For this purpose, an algorithm based on the U-Net CNN [10] was developed to crop the chest cavity from the radiograph images. This algorithm first segments the lungs from the CXR using pre-trained U-Net weights and a transfer learning approach to create a binary mask [11]. It then creates a region of interest (ROI) from the extreme points on the lungs mask and generates a bounding box to finally crop the chest cavity (Figure 1-I), removing regions that are not important for pneumonia detection, such as head, neck, arms, and exam objects. In some cases, due to large opacities in lung regions, the U-Net cannot detect one of the two lungs, and the consequent extreme points on the binary mask do not represent the extreme points on the chest cavity (Figure 1-II). In order to identify these cases, a simple rule was created. When the width of the chest cavity in the binary mask (w_{ref} in Figure 1-II)) is not greater than the half-width of the image (w), the distance w_{ref} is considered miscalculated. To recalculate the width of the chest cavity (w_{ref}), we take as reference the minor distance between the image's vertical borders and the extreme points (a in Figure 1-II). This value (a) will be the distance between the vertical borders of the ROI and the image, and $w_{ref} = w - 2a$.

We also hypothesize that enhancing the ROIs by traditional image pre-processing would also improve radiographic feature representation and pattern recognition. To increase

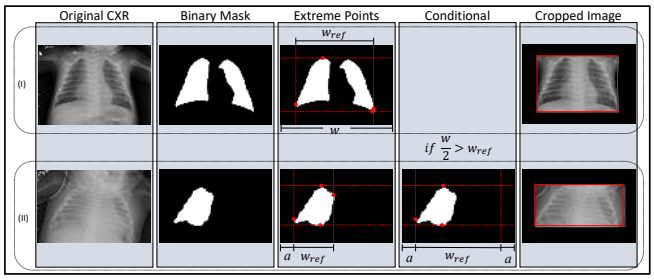

Fig. 1: Image cropping process.

the contrast of the overlapped projection of soft tissues, histogram equalization by contrast-limited adaptive histogram equalization (CLAHE) was then performed. CLAHE is a powerful procedure to locally enhance image patterns by limiting contrast amplification at a predefined value (in this work, 0.01) on the histogram before computing the cumulative distribution function [12].

C. Multi-View Ensemble Convolutional Neural Network

For classification purposes, we adopted the VGG16 CNN [13] and replaced its fully-connected layers with a customized multilayer perceptron (MLP) with three layers of 1024, 512, and 2 hidden neurons. VGG16 is a reasonably simple, widely known neural network with only 19 convolutional and pooling layers that was formerly used on CXR classification [3][14]. The VGG16 convolutional layers were initialized with weights trained on the ImageNet dataset [15] and the MLP with random weights. Between each MLP-layer, a Dropout Regularization layer was added with a rate of 30%. The MLP used the ReLU activation function, except the last layer, where the activation function was a softmax.

To evaluate the influence of pre-processing techniques on classification performance, we proposed three different training strategies, besides the baseline using original images:

- A: uses cropped images, as described in II.B;
- B: uses cropped images followed by histogram equalization using the CLAHE method;
- C: the ensemble of the strategies A and B (Figure 2). This strategy has an extra 3-layer MLP with 4, 10, and 2 hidden neurons. Here, the strategies A and B, composed by different gray-level intensities of the CXR (i.e., normal and equalized, forming the multi-view approach), are frozen, and their softmax probabilities are concatenated in the first MLP-layer.

In this work, we did not use data augmentation to balance the dataset. The training of the CNN and extra MLP of the ensemble strategy were performed with thirty and ten epochs, respectively. All training strategies were performed by stochastic gradient descent in batches of 16 images per step using an RMSprop Optimizer with a learning rate of 0.0001. As input, all images were resized to 224×224 pixels.

The metrics of the area under the ROC curve (AUC), sensitivity, specificity, accuracy, and F1 score assessed the performance of the four strategies. The Keras framework v2.2.5 with TensorFlow backend v1.14.0 was used for deep learning. Statistical analysis was performed by the DeLong's

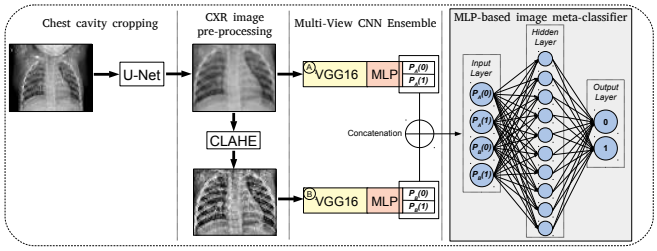

Fig. 2: Multi-view ensemble deep learning model for pneumonia classification. Different input images compose the proposed strategy of multi-view ensemble learning of CNNs.

test using R v3.4.4. The experiments were performed on Foxconn HPC M100-NHI with an 8-GPU cluster of NVIDIA Tesla V100 16GB cards. This computing infrastructure allowed the improvement in processing time of at least 7x in comparison with a workstation of 12 cores of 3.2 GHz (16 Gb RAM) and NVIDIA GeForce GTX 1050 with 4GB.

III. RESULTS

Table I presents the results for two classification tasks (normal *vs*. pneumonia and bacteria *vs*. virus). The ensemble strategy C yielded the highest performances for both binary tasks. This strategy combines the best properties of strategies A and B, improving classification efficiency, according to the statistical difference from the baseline. Due to the CLAHE local enhancement mechanism, opacities can get more homogenous for bacteria and virus, making the classification task more difficult between those patterns.

TABLE I: Performance obtained in the classification. AUC is presented with 95% confidence interval (CI). The asterisk indicates statistical difference from the baseline.

Strategy		*Baseline*	*A*	*B*	*C*
Pneumonia	AUC	0.87	0.95*	0.96*	0.97*
***vs*. Normal**	CI	0.85-0.90	0.93-0.97	0.95-0.98	0.96-0.99
Bacteria	AUC	0.85	0.88	0.83	0.91*
***vs*. Virus**	CI	0.81-0.88	0.85-0.92	0.80-0.87	0.88-0.94

Figure 3 shows models' class activation maps (CAMs) to corroborate that image cropping could leverage the detection of diseases on CXR as the regions of interest are more representative due to less information to be recognized. Image cropping also reduced irrelevant information from the exam and improved the representations of the region of interest (i.e., chest cavity and lungs). Figure 4 presents an example in which strategies A and B misclassified a patient with pneumonia as normal. However, strategy C correctly classified the patient, which could indicate multi-view ensemble deep learning may produce more reliable classification results, as it would not consider the information of only one "observer". Tables II and III present the performance comparison of the proposed method and the methods in the literature for the same dataset to classify pneumonia vs. normal patients and bacterial vs. viral pneumonia cases, respectively.

IV. DISCUSSION

Some works have already evaluated deep learning models in the detection of pediatric pneumonia on CXRs. Kermany et al. [2] adapted an Inception V3 architecture pre-trained on the ImageNet dataset. A hundred epochs were used, and it is not clear if they used or not data augmentation techniques. Liang et al. [3] proposed an architecture based on CNN and residual network. They trained for 100 epochs and used data augmentation in images of 150x150 pixels to balance the dataset. Moreover, they used the ChestXray14 dataset for transfer learning [16]. The proposed methods by Kermany et al. and Liang et al. yielded AUCs of 0.968 and 0.953, respectively, in the classification of normal and abnormal children's exams over the same dataset. Kermany et al. also subclassified pneumonia cases (bacterial vs. viral) and yielded AUC of 0.940. However, all of those models were trained with single architectures, which can lead to limited prediction accuracy, even with optimum parameters. To decrease this limitation, we proposed in this work a multi-view ensemble CNN model to classify pneumonia.

TABLE II: Comparison with the literature using the same dataset of normal *vs*. pneumonia classification.

Article	*AUC*	*Sensitivity*	*Specificity*	*Accuracy*	*F1 Score*
Kermany [2]	0.968	0.932	0.901	0.928	-
Liang [3]	0.953	0.967	0.803	0.905	0.927
This work	**0.973**	**0.979**	**0.966**	**0.974**	**0.979**

TABLE III: Comparison with the literature using the same dataset of viral *vs*. bacterial pneumonia classification.

Article	*AUC*	*Sensitivity*	*Specificity*	*Accuracy*	*F1 Score*
Kermany [2]	**0.940**	0.886	**0.909**	0.907	-
This work	0.907	**0.963**	0.851	**0.921**	**0.890**

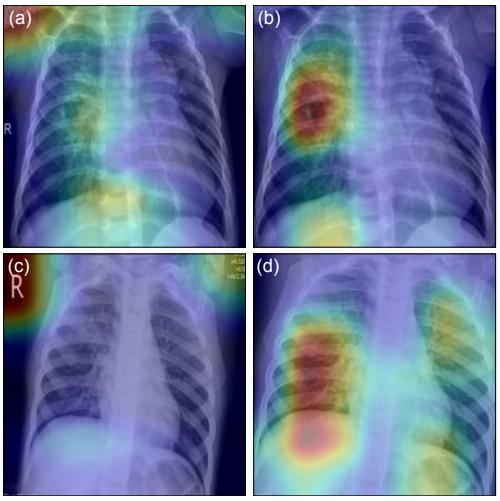

Fig. 3: CAMs show the most informative regions for classification: (a-b) CAMs from original and cropped images, respectively, of a patient with bacterial pneumonia; (c-d) CAMs from original and cropped images, respectively, of a patient with viral pneumonia.

Our model achieved a performance of AUC of 0.973, sensitivity of 0.979, specificity of 0.966, accuracy of 0.974, and F1 score of 0.979. This approach also can distinguish its subforms of bacteria and virus with an AUC of 0.907. For the best of our knowledge, the proposed method yielded higher AUC for the classification of normal and patients with

pneumonia when compared with the literature for the same dataset. Unlike the literature [3], we adopted the VGG16 architecture with its default input size [13], which potentially extracted more efficiently the network standard information. A different VGG16 input size could limit network training and reduce image characterization potential. ImageNet weight initialization was an easy method to initialize the network, potentially enabling rapid convergence [15]. Moreover, only 30 epochs were necessary to achieve this convergence.

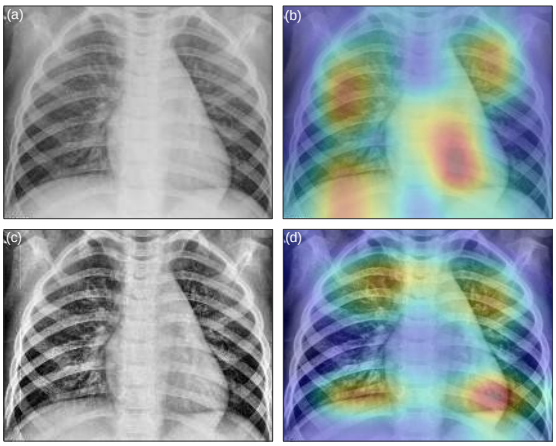

Fig. 4: Classification results of a patient with pneumonia: (a-b) cropped and CAM images, and (c-d) equalized and CAM images, respectively. Strategies A and B misclassified the patient as normal (false negative), while the proposed strategy correctly classified it as a true positive.

Overall, pre-processing techniques yielded the highest performances in binary classifications. Specifically, image cropping reduced irrelevant information from the exam and improved the representations of the region of interest (i.e., chest cavity and lungs). Histogram equalization enhanced contrast of soft tissues and was able to identify fine textures from CXR. This multi-view approach with both preprocessing methods (ROI cropping and CLAHE) is simple to implement, and each method offers a different advantage to the training, yielding better results on the classification when compared with the use of the original CRX (baseline).

The external generalization using multi-view ensemble CNN has potential due to the extra MLP that takes advantage of the different features extracted from the original intensity and histogram-equalized gray levels, and it learns to weigh and associate the probabilities of different training methods on the classification task. For these reasons, we highlight the need for pre-processing of input medical images for performance improvement of the VGG16 network.

Our main limitation was the lack of an independent external dataset for generalization purposes. There are other publicly available CXR datasets on the literature, i.e., ChestX-ray8 and CheXpert [16], [17]. However, neither includes exams from pediatric patients. Moreover, they are known to be inconsistent as the image labeling procedure was performed by natural language processing on radiology reports, which could lead to text-mining errors as labels may not accurately reflect the visual content of the images [14][18].

We propose for future works to evaluate the generalization of the multi-view ensemble deep learning model with other cohorts and expand the investigation to include adult exams from patients with pneumonia.

ACKNOWLEDGMENT

This work was supported by Foxconn Brazil and Zerbini Foundation as part of the research project "Machine Learning in Cardiovascular Medicine".

REFERENCES

[1] I. Rudan, C. Boschi-Pinto, Z. Biloglav, K. Mulholland, and H. Campbell, "Epidemiology and etiology of childhood pneumonia," *Bulletin of the world health organization*, vol. 86, pp. 408–416B, 2008.

[2] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[3] G. Liang and L. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," *Comput Methods Programs Biomed*, 2019, DOI:10.1016/j.cmpb.2019.06.023.

[4] T. Franquet, "Imaging of community-acquired pneumonia," *Journal of Thoracic Imaging*, vol. 33, no. 5, pp. 282–294, 2018.

[5] L. A. Mandell, R. G. Wunderink, A. Anzueto, J. G. Bartlett, G. D. Campbell, N. C. Dean, S. F. Dowell *et al.*, "Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults," *Clin Infect Dis*, vol. 44, pp. S27–S72, 2007.

[6] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *preprint arXiv:1711.05225*, 2017.

[7] J. Ferreira, M. Santos, A. Tenório, M. Faleiros, F. Cipriano, A. Fabro, J. Näppi, H. Yoshida, and P. Marques, "CT-based radiomics for prediction of histologic subtype and metastatic disease in primary malignant lung neoplasms," *Int J Comput Assist Radiol Surg*, vol. 15, pp. 163–172, 2020.

[8] R. Sousa, O. Marques, F. Soares, I. Sene, L. Oliveira, and E. Spoto, "Comparative performance analysis of machine learning classifiers in detection of childhood pneumonia using chest radiographs," *Procedia Computer Science*, vol. 18, pp. 2579–2582, 2013.

[9] T. B. Chandra and K. Verma, "Pneumonia detection on chest x-ray using machine learning paradigm," in *3rd International Conference on Computer Vision and Image Processing*, 2020, pp. 21–33.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[11] I. Pazhitnykh and V. Petsiuk, "Lung segmentation (2D)," https://github.com/imlab-uiip/lung-segmentation-2d, 2017.

[12] K. Zuiderveld, *Graphic Gems IV.* Academic Press Professional, Inc., Cambridge, Massachusetts, USA, 1994.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *preprint arXiv:1409.1556*, 2014.

[14] S. Candemir, S. Rajaraman, G. Thoma, and S. Antani, "Deep learning for grading cardiomegaly severity in chest x-rays: an investigation," in *2018 IEEE Life Sciences Conference (LSC)*, 2018, pp. 109–113.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[16] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2097–2106.

[17] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," *preprint arXiv:1901.07031*, 2019.

[18] L. Oakden-Rayner, "Exploring large-scale public medical image datasets," *Academic Radiology*, vol. 27, no. 1, pp. 106–112, 2020.