Multicenter Validation of Convolutional Neural Networks for Automated Detection of Cardiomegaly on Chest Radiographs

Diego A. Cardona Cardenas, José R. Ferreira Jr., Ramon A. Moreno, Marina F. S. Rebelo, José E. Krieger, Marco A. Gutierrez

Heart Institute, Clinics Hospital, University of Sao Paulo Medical School
Av. Dr. Enéas de Carvalho Aguiar 44, 05403-900, São Paulo – SP – Brazil

{diego.cardenas, jose.raniery, ramon.moreno}@incor.usp.br

{marina.rebelo, krieger, marco.gutierrez}@incor.usp.br

Abstract. *This work focused on validating five convolutional neural network models to detect automatically cardiomegaly, a health complication that causes heart enlargement, which may lead to cardiac arrest. To do that, we trained the models with a customized multilayer perceptron. Radiographs from two public datasets were used in experiments, one of them only for external validation. Images were pre-processed to contain just the chest cavity. The EfficientNet model yielded the highest area under the curve (AUC) of 0.91 on the test set. However, the Inception-based model obtained the best generalization performance with AUC of 0.88 on the independent multicentric dataset. Therefore, this work accurately validated radiographic models to identify patients with cardiomegaly.*

1. Introduction

Cardiomegaly is a medical condition in which a patient has an enlargement of the heart temporarily or permanently, depending on the condition. This increase in size is usually a manifestation of another pathologic process, and it may result in heart failure, cardiac arrest, and in some cases, even sudden death [Amin and Siddiqui 2019]. Some of the causes for a heart enlargement are weakening of the cardiac muscle, coronary artery disease, high blood pressure, cardiomyopathy, and heart valve disease. However, the heart may enlarge for unknown reasons, a condition known as idiopathic cardiomegaly, which may increase the risks of complications for the patient [Mayo Foundation for Medical Education and Research 2020]. Moreover, the heart enlargement in the form of either dilatation or hypertrophic cardiomyopathy leads to a spectrum of clinical heart failure syndrome with a very poor prognosis and a five-year survival rate of only 50% [Amin and Siddiqui 2019, Bui et al. 2011].

An abnormal heart can be evaluated with a medical imaging exam, like a chest radiograph (Figure 1). The clinical evaluation of cardiomegaly is based on the calculation of the cardiothoracic ratio (CTR), a widely used radiographic index to assess cardiac size and provide prognostic information in both congenital and acquired heart diseases [Li et al. 2019]. The CTR on a chest x-ray (CXR) image is basically the relationship between cardiac and thoracic diameters. Unfortunately, it is manually measured by the specialists, which is a very time-consuming task because of the large volume of images produced nowadays [Koenigkam-Santos et al. 2019]. A better approach is to automate this measurement to improve medical care.

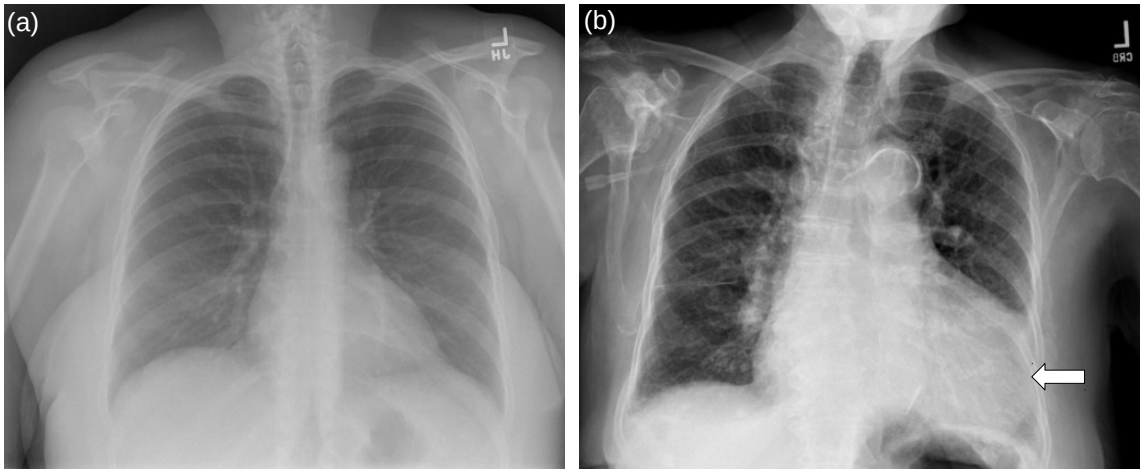

Figure 1. Posterior-anterior CXR images from a normal subject (a) and a patient who presented irregular heartbeat (b). The arrow indicates a markedly severe cardiomegaly.

Computer-based tools can analyze large volumes of medical images and accurately detect disease patterns automatically [Ferreira Junior et al. 2020b]. Recently, the use of deep learning has been gaining considerable importance in medical image analysis as it has the potential to improve the efficiency of specialists [LeCun et al. 2015]. This artificial intelligence branch uses raw data (i.e., image pixels) as input and abstracts layer-wise the original imaging data into a final feature vector that is used for disease detection without manual procedures by the specialist [Liang and Zheng 2019].

Some works have detected cardiomegaly on CXR by employing deep learning and convolutional neural networks (CNNs). Chamveha et al. proposed an algorithm for calculating the CTR based on the U-Net segmentation of the heart and further detection of cardiomegaly [Chamveha et al. 2020]. The authors combined the NIH ChestX-Ray14 [Wang et al. 2017] and Stanford CheXpert datasets [Irvin et al. 2019] for both training and testing. Que et al. also developed an algorithm based on the U-Net to measure the CTR from CXRs and used the NIH ChestX-Ray14 dataset for experiments [Que et al. 2018]. Li et al. proposed a customized U-Net model to segment the cardiac region on the CXR for the CTR calculation and the heart enlargement [Li et al. 2019]. The authors used images from a local repository for the experiments. Candemir et al. developed a CNN model to detect the cardiomegaly and associated softmax probabilities with severity grades of the heart enlargement [Candemir et al. 2018]. The authors used images from the datasets NIH ChestX-Ray14 for training and from OpenI for both training and testing [Demner-Fushman et al. 2015]. However, none of those studies have used an external dataset to test the generalization of the proposed solutions.

Generalization, or external validation, is a significant challenge for machine learning models, including those based on deep learning [Koenigkam-Santos et al. 2019]. Testing the model generalization is an essential step to validate CNN algorithms, especially to important methodological regulation agencies like FDA (Food and Drug Administration) and IMDRF (International Medical Device Regulators Forum). Once a machine-learning-based method has been appropriately validated on an independent external dataset, it could be used routinely as a Soft-

ware as a Medical Device (SaMD) to potentially support clinical decision-making [International Medical Device Regulators Forum 2013, Ferreira Junior et al. 2020b].

Therefore, our purpose in this work is to develop deep-learning methods to detect cardiomegaly on radiographic images and to test the generalization potential of the proposed models on an independent multicenter CXR dataset. To the best of our knowledge, this is the first work proposing to validate state-of-the-art artificial intelligence methods for cardiomegaly on CXRs, both independently and externally.

2. Material and Methods

2.1. CXR Datasets

This study used retrospectively posterior-anterior (PA) CXR images from patients of two public image datasets: PadChest and OpenI [Bustos et al. 2019, Demner-Fushman et al. 2015]. PadChest is a labeled large-scale Spanish CXR dataset with image-associated reports. To this date, PadChest included 160,868 images from 67,625 patients that attended the San Juan de Alicante Hospital, University of Alicante, Spain. Manual report labeling (gold standard) provided by the dataset's specialists was used as ground truth in this work, and it was performed on 39,039 images, from which we selected 10,566 in the PA projection view, without external objects like prosthesis, catheter, metal objects, pacemaker, among others. We used PadChest to model the CNNs for image classification, and thus, we randomly split this PA-CXR cohort in training, validation, and testing sets (with the proportion of 75:19:6, Table 1) [Ferreira Junior et al. 2020a].

Table 1. Sampling of the CXR image datasets used in this work.

	PadChest				OpenI
	Training set	Validation set	Testing set		Entire set
Normal	6,728	1,682	297	Normal	1,388
Cardiomegaly	1,249	313	297	Cardiomegaly	321

Furthermore, OpenI is a publicly available American repository and web-service that enables search and retrieval of images from the National Library of Medicine, National Institutes of Health (NIH). OpenI currently has 7,470 frontal and lateral CXR examinations from various hospitals of the Indiana University School of Medicine, USA, along with 3,955 anonymized radiology reports [Candemir et al. 2018]. Manual report labeling (gold standard) was also used as ground truth to images from OpenI. We used the Medical Subject Heading (MeSH) indexing descriptors included on the image-associated reports to label the CXRs. OpenI was used only for generalization testing purposes, and hence, this study comprised all 1,709 PA exams from OpenI (Table 1). It is important to note that images from patients with cardiomegaly could also present other chest abnormalities, like lung opacities and mediastinal complications. Some of the most commons were consolidation, atelectasis, pneumothorax, pleural effusion, aortic elongation, hilar enlargement, among others.

2.2. Image Pre-Processing

To improve the training efficiency, we previously developed a model to remove structures from the radiographic exams that may not interest the analysis

[Ferreira Junior et al. 2020a]. This model is based on the U-Net CNN that removes anatomical parts like the head, neck, and arms, along with exam objects from the CXR image. The algorithm first segments the lungs using transfer learning and pre-trained U-Net weights to create a binary mask of the chest cavity region [Pazhitnykh and Petsiuk 2017, Ronneberger et al. 2015]. It then creates a bounding box from the extreme points on the lungs mask to crop the chest cavity and segment only the region of interest (Figure 2).

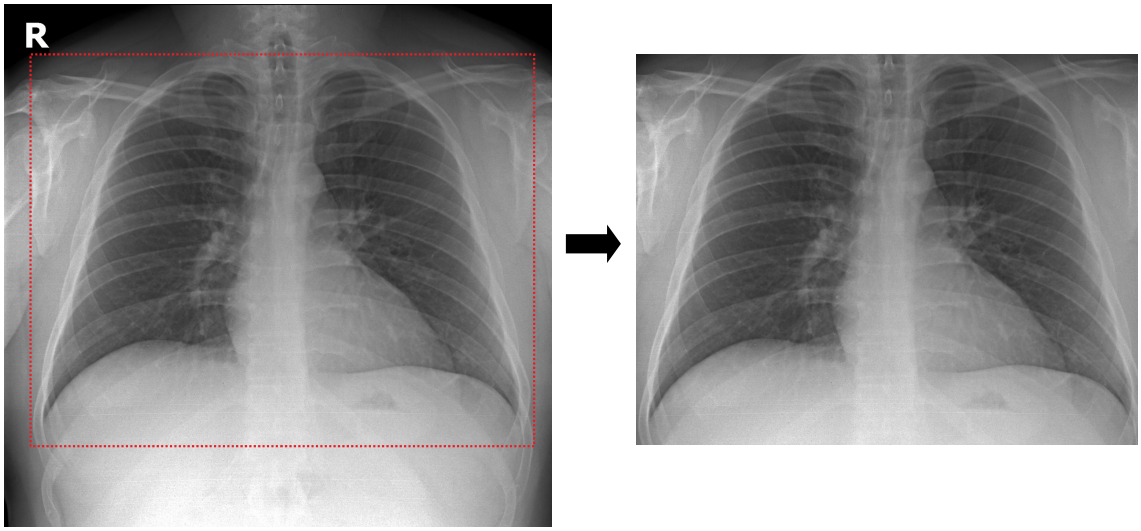

Figure 2. Automatic cropping procedure used in this work to pre-process the images to serve as input for the CNNs. Left: original image. Right: cropped image.

To highlight the importance of image pre-processing and the previously developed algorithm [Ferreira Junior et al. 2020a], Figure 3 presents a global distribution of the datasets using a bi-dimensional space, where each point represents a single image. The t-Distributed Stochastic Neighbor Embedding (tSNE) method was used to allow the visualization of image representations [Maaten and Hinton 2008]. It can be seen from Figure 3 that the datasets seem better clustered with cropped images. Moreover, visualization of clusters in the tSNE can correlate with the performance of CNNs in the identification of cardiomegaly.

2.3. Convolutional Neural Network Training

In this work, we assessed five CNNs on the automated detection task: EfficientNetB2, DenseNet121, Xception, InceptionV3, and MobileNet [Tan and Le 2019, Huang et al. 2017, Chollet 2017, Szegedy et al. 2016, Howard et al. 2017]. These CNNs have been widely used in medical image classification works, including CXR [Rajpurkar et al. 2017, Wang et al. 2017, Irvin et al. 2019, Candemir et al. 2018]. We used ImageNet weights to initialize the training [Russakovsky et al. 2015] and the default architecture topology before classification layers for all CNNs. However, we replaced the original fully connected layers with a customized 3-layer multilayer perceptron to standardize classification layers for all CNNs. A Dropout Regularization was used between each MLP layer with a rate of 30%. The MLP had a ReLU activation function except for the last layer for which the activation function was sigmoid. All CNN inputs also used a resolution of 224×224 for standardization purposes.

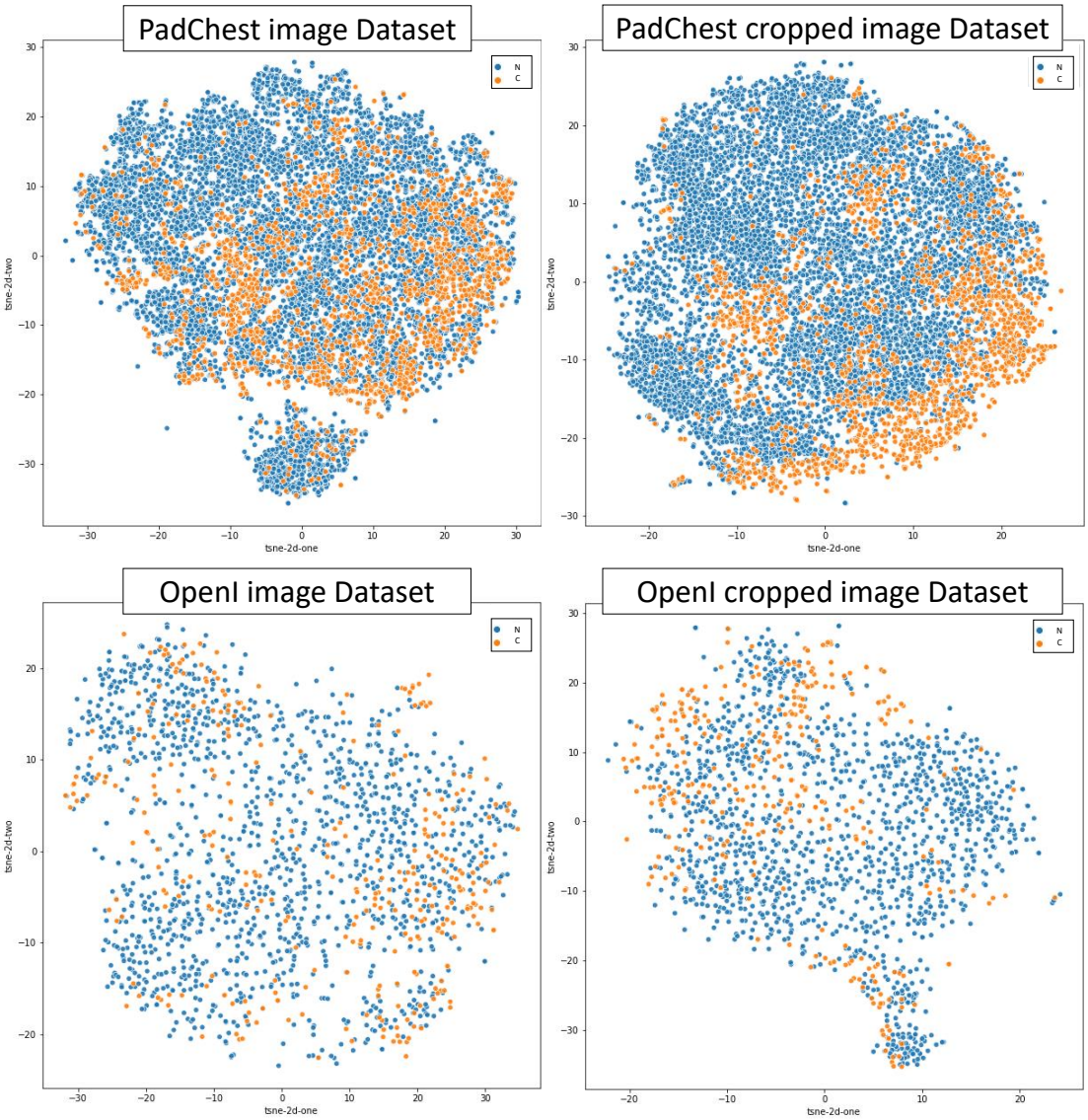

Figure 3. Global distribution of the datasets PadChest (n = 10,566) and OpenI (n = 1,709) with original and cropped CXR images.

Data augmentation techniques (i.e., image rotation, shift, shear, scale, and flip) were used to reduce the training set imbalance for CNN modeling. Therefore, augmentation was performed only in images with cardiomegaly. The training was performed with 30 epochs by stochastic gradient descent in batches of 16 images per step using RMSprop. This optimizer had a learning rate of 0.001 and a callback to reduce it by a factor of 0.5 every five epochs with no improvement in validation loss.

To improve modeling sensitivity, we employed a strategy to increase the weight of the abnormal class for the training [Ferreira Junior et al. 2020b]. For this purpose, normal images weighted 0.25, and augmented images from patients with cardiomegaly weighted 0.75 (Table 1). Those weights were defined empirically. To evaluate any performance improvement from this strategy, we also trained the models without weighted class relevance, i.e., with default 0.5 weights for both classes after the augmentation. We will use

the notation "75-25-weighing" to refer to the first strategy and "50-50-weighing" for the second strategy in the remainder of the paper.

2.4. Performance Evaluation

The receiver operating characteristic (ROC) curve and the confusion matrix assessed the experiments. The metrics of the area under the ROC curve (AUC), sensitivity, specificity, accuracy, precision, and F1 score were calculated to evaluate the analysis. The Keras framework v2.2.5 with TensorFlow backend v1.14.0 was used for modeling. The DeLong's test measured the statistical difference in ROC curves. Tests with p value < 0.05 were considered significant. Statistical analysis was performed using R v3.4.4. All experiments were performed on a Foxconn HPC M100-NHI with an 8-GPU cluster of NVIDIA Tesla V100 16GB cards.

3. Results

EfficientNetB2 obtained the highest overall performance with AUC of 0.912 on the test set to detect cardiomegaly, yielding the following metrics: sensitivity of 0.892, specificity of 0.933, accuracy of 0.912, precision of 0.930, and F1 score of 0.911. Figure 4 presents the AUC of the five CNNs for the test data with two strategies (50-50 and 75-25). Figure 5 presents the AUC of the same five CNNs for the external dataset with also both strategies. The default class 50-50-weighing yielded higher performance than the proposed strategy of class 75-25-weighing on the test set only with the EfficientNetB2 model. On the other hand, the other four CNN models obtained higher performance when using the 75-25-weighing, although no statistically significant difference was found with the DenseNet121, Xception, and InceptionV3 models. Furthermore, the largest improvement in sensitivity (the main reason to use the 75-25-weighing strategy) was obtained by the MobileNet model, with an increase of 0.255 percentage points from 0.569 to 0.824.

Concerning the generalization potential, most of the results obtained on the test set were confirmed on the independent external dataset used. The default 50-50-weighing yielded higher performance on the external dataset (Figure 5) when using the EfficientNetB2 and Xception architectures, corroborating the EfficientNet invariance to class weighing. Moreover, the InceptionV3 and MobileNet models also confirmed the statistically significant performance improvement with a higher weight for the cardiomegaly class during the training. Furthermore, the most significant improvement in sensitivity was again obtained by the MobileNet model, with an increase of 0.358 percentage points from 0.464 to 0.822. Finally, the best generalization performance was obtained by the InceptionV3 model with AUC of 0.879, achieving the following efficiency: sensitivity of 0.869, specificity of 0.890, accuracy of 0.886, precision of 0.646, and F1 score of 0.741.

Figure 6 presents traditional class activation maps to allow explain visually which regions of the CXR image were related to each model output and to corroborate the quantitative results previously explained. The figure also shows another mapping approach to locally visualize which pixels from the CXR that are most relevant for image classification [Wang et al. 2020, Springenberg et al. 2015]. This so-called guided backpropagation map learns class-descriptive regions using gradient propagation of the prediction scores. By using a backward pass after passing through the network forward, it is possible to compute the gradient of the activation of the neuron [Wang et al. 2020]. The method also adds

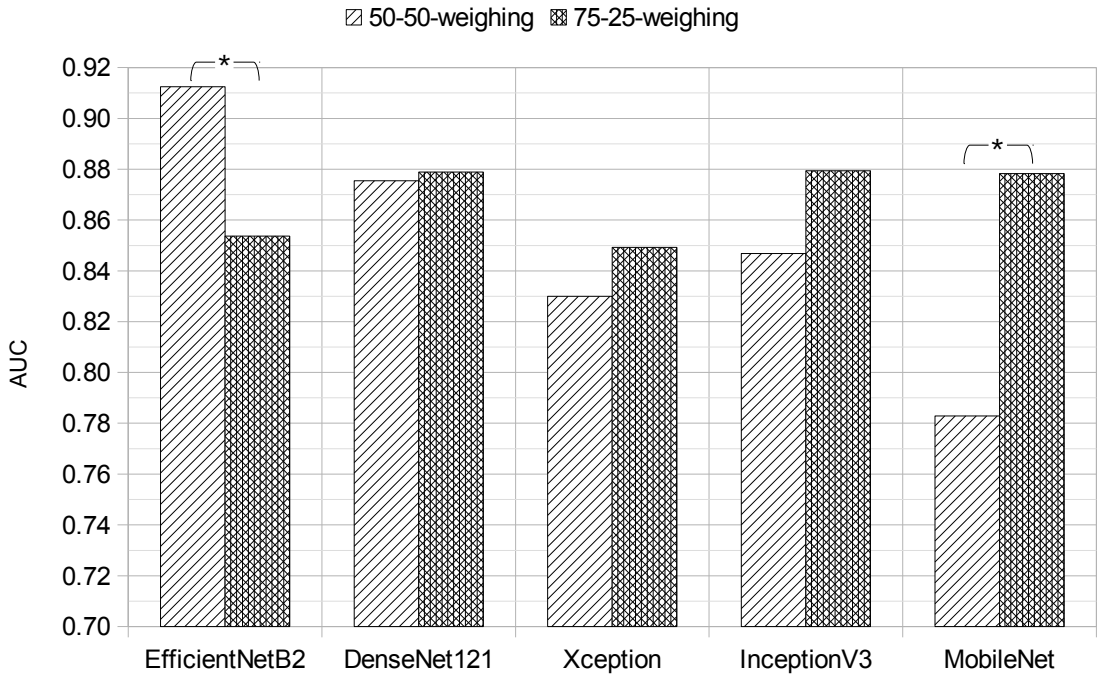

Figure 4. Efficiency of the models on the test set. The asterisk at the end of the bar indicates statistically significant difference.

another signal from higher layers to prevent the backward flow of negative gradients for enhanced class mapping [Springenberg et al. 2015]. Figure 6 highlights the importance of class weighing on training to detect cardiomegaly as the heart-located region received more global and local activation when using the 75-25-weighting strategy. The default class 50-50-weighing produces both globally and locally scattered activated regions, resulting in lower performance on the detection of cardiomegaly.

4. Discussion

In this work, we developed, tested, and externally validated convolutional neural network models to detect cardiomegaly automatically on chest radiograph images. Cardiomegaly is a health complication that has several etiologies and may result in heart failure. Despite the development of new therapies, mortality remains high in patients with symptomatic heart failure, and hence, cardiomegaly needs to be cautiously and rapidly assessed [Amin and Siddiqui 2019]. CXR images are accessible, inexpensive, and dose-effective compared to other imaging modalities, like computed tomography and magnetic resonance [Candemir et al. 2018]. However, differential diagnosis of cardiomegaly on CXR includes disorders that can result in an enlarged cardiomediastinal silhouette, hindering, even more, the clinical decision making [Li et al. 2019, Amin and Siddiqui 2019].

The proposed models developed in this work automatically detect cardiomegaly patterns, which could bring great benefits to the clinical routine at the beginning of care as they could prioritize abnormal exams for further reading from a specialist, ultimately optimizing examination time and reporting. They were validated on an independent multicentric CXR dataset and presented great potential to be embedded in a SaMD tool. To the best of our knowledge, this is the first work to validate deep learning models for car-

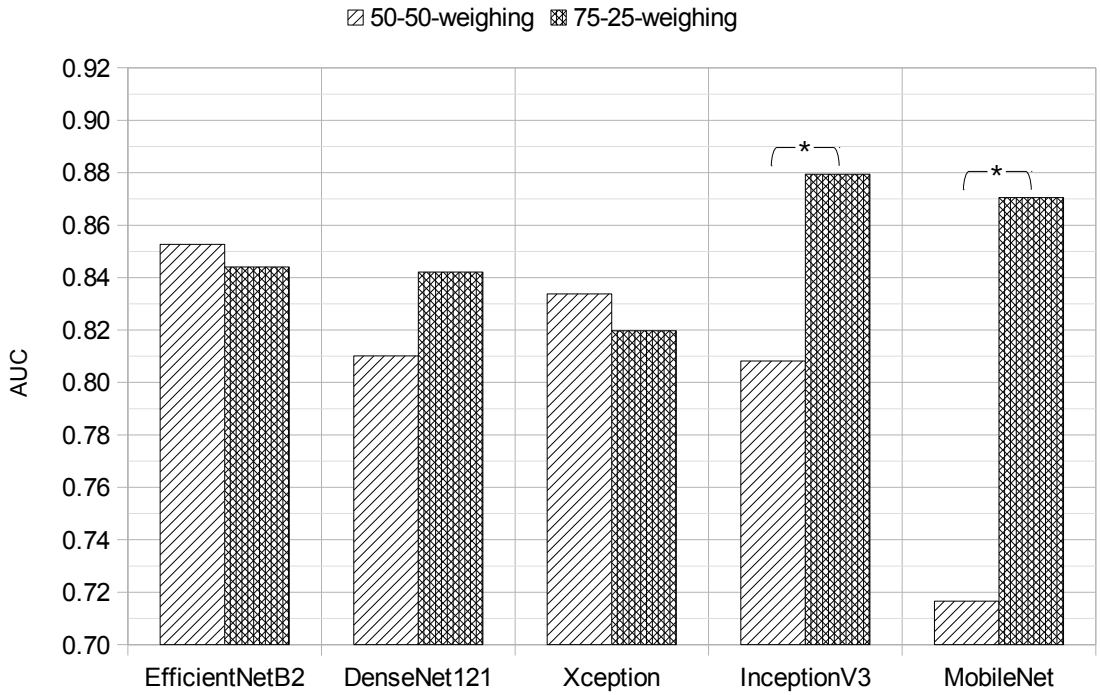

Figure 5. Efficiency of the models on independent external dataset. The asterisk at the end of the bar indicates statistically significant difference.

diomegaly that were trained and externally tested with exams from patients, x-ray scanners, and institutions totally different. Table 2 presents the performance of those validated models that obtained AUC of at least 0.85 on the independent dataset to be further used as a SaMD.

Table 2. Performance of the validated CNN models for automated detection of cardiomegaly on chest radiograph.

CNN	AUC	Sensitivity	Specificity	Accuracy	Precision	F1 Score
InceptionV3	0.879	0.869	0.890	0.886	0.646	0.741
MobileNet	0.871	0.822	0.919	0.901	0.700	0.756
EfficientNetB2	0.853	0.760	0.945	0.910	0.763	0.761

Comparing our work with the literature, most of the approaches used U-Net segmentation to calculate the CTR and then identified cardiomegaly. Chamveha et al. [Chamveha et al. 2020] combined the NIH ChestX-Ray14 and Stanford CheXpert datasets for both training and testing, and they obtained accuracies on a test set of 0.671 using the former and 0.698 using the latter. Que et al. [Que et al. 2018] used the NIH ChestX-Ray14 dataset yielded an AUC of 0.935 on the classification task. Li et al. [Li et al. 2019] used images from a local repository for the experiments and obtained an accuracy on a testing set of 0.953. Other works, like the proposed by Candemir et al. [Candemir et al. 2018] using a CNN model and combined the NIH ChestX-Ray14 for training and OpenI for both training and testing, achieved an AUC of 0.95. However, it is essential to highlight that none of those externally validated the proposed solutions with independent datasets.

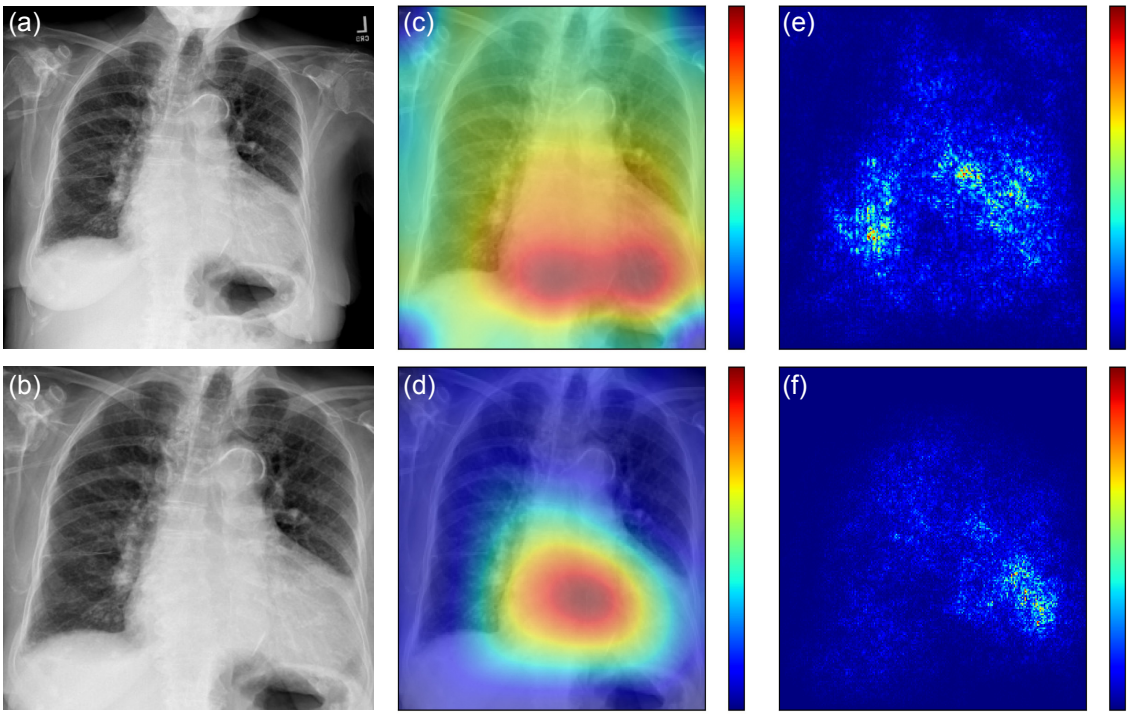

Figure 6. Discriminative parts from chest-cavity-cropped images mapped from the last layer of MobileNet models: (a) original CXR image; (b) resulted image after automatic chest cavity cropping; (c-d) class activation maps obtained from 50-50 and 75-25 weighing strategies, respectively; (e-f) guided-backpropagated maps obtained from 50-50 and 75-25 weighing strategies, respectively.

It is worth mention the InceptionV3, MobileNet, and EfficientNetB2 models from this work could play a key role in computer-aided detection of heart anomalies related to size. All of those models yielded statistically equivalent high efficiency of generalization. Moreover, their performances (AUC on independent external testing > 0.85) are acceptable for method validation by the FDA and IMDRF. Furthermore, for simple architectures, like InceptionV3 and MobileNet, the 75-25 weighing strategy was essential to improve pattern recognition and increase the performance for validation. For robust architectures, like EfficientNetB2, class weighing was not a significant factor to improve training efficiency. Therefore, class weighing is an important parameter to take into account the CNN architecture and, mostly, to image classification and efficient detection of abnormalities on medical exams.

Our main limitation in this study was the lack of clinical data available from the public datasets. The institutions only made the bitmap images available, which, in the first moment, limited the methods as the CXRs are not in the original DICOM (Digital Imaging and Communications in Medicine) format. Moreover, it did not allow the possibility to investigate the causes of the heart enlargement and to associate radiographic features with clinical outcomes (investigation approach known as radiomics [Koenigkam-Santos et al. 2019]) from the patients with cardiomegaly. There are other CXR datasets on the literature, like the previously cited NIH ChestX-Ray14 and Stanford CheXpert. However, neither had patient clinical data publicly available. More-

over, they are known to be inconsistent as natural language processing on radiology reports performed the image labeling, which could lead to text-mining errors as labels may not accurately reflect the visual content of the images [Sabottke and Spieler 2020, Oakden-Rayner 2020, Candemir et al. 2018]. Another limitation of this work is that the analysis was performed with cardiomegaly (plus other chest diseases) *vs.* normal cases. Another experimental approach could be the classification of positive *vs.* negative cases for cardiomegaly.

5. Conclusion

In summary, the proposed models disclosed great potential to be used as a SaMD for cardiomegaly. Moreover, it could promote the development of a teleradiology tool to aid clinical routine in distant places with limited medical resources, where x-ray scanners are the only imaging option of health care. Finally, some directions we could take to improve this research are to first expand the investigation by clinically validating the models and assessing whether they could aid imaging diagnosis routinely; and to design a robust radiomics model able to associate CXR features with cardiomegaly outcomes to potentially support clinical decisions and precision imaging of cardiothoracic diseases.

Acknowledgements

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

References

Amin, H. and Siddiqui, W. J. (2019). *Cardiomegaly*. StatPearls Publishing.

Bui, A. L., Horwich, T. B., and Fonarow, G. C. (2011). Epidemiology and risk profile of heart failure. *Nature Reviews Cardiology*, 8(1):30.

Bustos, A., Pertusa, A., Salinas, J.-M., and de la Iglesia-Vayá, M. (2019). Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv preprint arXiv:1901.07441*.

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

Chamveha, I., Promwiset, T., Tongdee, T., Saiviroonporn, P., and Chaisangmongkon, W. (2020). Automated cardiothoracic ratio calculation and cardiomegaly detection using deep learning approach. *arXiv preprint arXiv:2002.07468*.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258.

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. (2015). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Ferreira Junior, J. R., Cardenas, D. A. C., Moreno, R. A., Rebelo, M. F. S., Krieger, J. E., and Gutierrez, M. A. (2020a). Multi-view ensemble convolutional neural network to improve classification of pneumonia in low contrast chest x-ray images. In *42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society*. Accepted for publication.

Ferreira Junior, J. R., Santos, M. K., Tenório, A. P. M., Faleiros, M. C., Cipriano, F. E. G., Fabro, A. T., Näppi, J., Yoshida, H., and Azevedo Marques, P. M. (2020b). CT-based radiomics for prediction of histologic subtype and metastatic disease in primary malignant lung neoplasms. *International Journal of Computer Assisted Radiology and Surgery*, 15:163–172.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708.

International Medical Device Regulators Forum (2013). Software as a Medical Device (SaMD): Key Definitions. Last access on March 5th, 2020. Available at `http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-definitions-140901.pdf`.

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

Koenigkam-Santos, M., Ferreira Junior, J. R., Wada, D. T., Tenório, A. P. M., Barbosa, M. H. N., and Azevedo Marques, P. M. (2019). Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine. *Radiologia Brasileira*, 52(6):387–396.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Li, Z., Hou, Z., Chen, C., Hao, Z., An, Y., Liang, S., and Lu, B. (2019). Automatic cardiothoracic ratio calculation with deep learning. *IEEE Access*, 7:37749–37756.

Liang, G. and Zheng, L. (2019). A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer Methods and Programs in Biomedicine*. Ahead of print. DOI:10.1016/j.cmpb.2019.06.023.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Mayo Foundation for Medical Education and Research (2020). Enlarged heart. Last access on Feb 28th, 2020. Available at `https://www.mayoclinic.org/diseases-conditions/enlarged-heart/symptoms-causes/syc-20355436`.

Oakden-Rayner, L. (2020). Exploring large-scale public medical image datasets. *Academic Radiology*, 27(1):106–112.

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

Que, Q., Tang, Z., Wang, R., Zeng, Z., Wang, J., Chua, M., Gee, T. S., Yang, X., and Veeravalli, B. (2018). CardioXNet: automated detection for cardiomegaly based on deep learning. In *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 612–615.

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

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sabottke, C. F. and Spieler, B. M. (2020). The effect of image resolution on deep learning in radiography. *Radiology: Artificial Intelligence*, 2(1):e190015.

Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2015). Striving for simplicity: The all convolutional net. In *Proceedings of the 3rd International Conference on Learning Representations*, page arXiv:1412.6806.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114.

Wang, H., Jia, H., Lu, L., and Xia, Y. (2020). Thorax-Net: An attention regularized deep neural network for classification of thoracic diseases on chest radiography. *IEEE Journal of Biomedical and Health Informatics*, 24:475–485.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). 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*, pages 2097–2106.