Explainable AI in Deep Learning-based Detection of Aortic Elongation on Chest X-ray Images

Estela Ribeiro[1, 2, *], **Diego A. C. Cardenas**[1], **Felipe M. Dias**[1, 3], **José E. Krieger**[1], **and Marco Antônio Gutierrez**[1, 2, 3]

[1]Heart Institute (InCor) – Clinics Hospital University of Sao Paulo Medical School (HCFMUSP) – Sao Paulo – SP – Brazil
[2]University of Sao Paulo Medical School (FMUSP) – Sao Paulo – SP – Brazil
[3]Polytechnique School, University of Sao Paulo (POLI USP) – Sao Paulo – SP – Brazil
[*]estela.ribeiro@hc.fm.usp.br

ABSTRACT

Aortic Elongation can result from age-related changes, congenital factors, aneurysms, or conditions affecting blood vessel elasticity. It is associated with cardiovascular diseases and severe complications like aortic aneurysms and dissection. In this work, we evaluated the performance of deep learning models (DenseNet and EfficientNet) for aortic elongation detection based on transfer learning and fine-tunning techniques using Chest X-Rays (CXR) as input. DenseNet achieved higher accuracy (84.7% $\pm$ 0.7), precision (75.6% $\pm$ 1.3), sensitivity (88.7% $\pm$ 2.7), specificity (82.3% $\pm$ 1.6), F1-score (81.6% $\pm$ 1.0), and AUROC (93.1% $\pm$ 0.4) than EfficientNet. To gain insights into the decision-making process of the deep learning models, we employed Grad-CAM and LIME explainability methods. Through these techniques, we were able to successfully identify the expected location of aortic elongation in the x-ray images. Moreover, we used the pixel-flipping method to assess quantitatively the interpretations providing valuable insights into models behavior. By incorporating explainable AI techniques, we enhanced the interpretability and understanding of the models' decisions. This approach holds promise for aiding clinicians in timely and accurate diagnosis, potentially improving patient outcomes in clinical practice.

1 Introduction

Aortic Elongation is a medical condition in which the aorta, the largest artery in the human body, is longer than normal[1]. This can occur for various reasons, including age-associated changes[2], congenital factors[3–7], aneurysms[8], or underlying medical conditions that affect the elasticity of blood vessels, resulting in Cardiovascular Diseases (CVD). Common diseases associated to aortic elongation include aortic aneurysms[8], Marfan Syndrome[3], Ehlers-Danlos Syndrome[4], Loeys-Dietz Syndrome[6] and Turner Syndrome[7]. The consequences of an elongated aorta can vary depending on the severity and cause of the condition. In some cases, an elongated aorta may not present symptoms or significant problems. In other cases, however, it can lead to serious complications such as aortic aneurysms, aortic dissection, aortic rupture, and other cardiovascular conditions[1].

The identification of aortic elongation can be done through imaging exams such as Chest X-Rays (CXR), Computerized Tomography (CT) or Magnetic Resonance Imaging (MRI). Works focused on automatic detection of diseases using these imaging exams are largely benefited with the growth of Deep Learning (DL) field that provides a powerful set of tools for processing and analyzing medical images.

In comparison to conventional machine learning methods, Convolutional Neural Networks (CNNs)[9,10], a type of Deep Neural Network that is commonly used in Computer Vision tasks such as image and video recognition, have several significant advantages. CNNs are able to extract features from raw data without any prior knowledge of the input data. This is particularly useful for image recognition tasks where traditional feature extraction methods may not be effective. CNNs also learn multiple layers of abstraction that represent features at increasing levels of complexity. This allows them to recognize complex patterns and objects. Furthermore, they have the ability to recognize patterns that are difficult for the human eye to notice by learning from large amounts of data.

Overall, Deep Learning has the potential to greatly improve the accuracy and efficiency of medical imaging, leading to better patient outcomes and more effective healthcare. However, it is important to carefully validate and interpret the results of DL models, as they can be sensitive to biases in the training data and may not always generalize to new patient populations.

Since many hospitals are remote from major centers and lack access to radiologists to review and report exams, these DL models make it possible to obtain results with a high success rate that aid inexperienced doctors' diagnosis. In this sense, Computer-aided diagnostic (CAD) systems can help hospitals reduce their expenses, prioritize patients with more critical

conditions, and shorten longer waiting services. However, due in part to the following issues, this kind of application is still quite rare in clinical practice: (i) the majority of existing CXR datasets rely on automated disease label extraction from free-text radiology reports that usually uses Natural Language Processing (NLP) methods, which can lead to inconsistencies and errors; (ii) there aren't many CXR datasets with labeled abnormality annotation, and the ones that do tend to be too small; (iii) current research ignores difficulties with application, primarily focusing on enhancing the performance of machine learning models; and (iv) most of the proposed methods employ complicated machine learning algorithms that do not always convey precise and comprehensible knowledge about how the models arrive at their predictions, due to the lack of interpretability.

Regarding interpretability, the black-box nature of these deep learning models prevents transparency in their decision-making processes, despite the fact that they are capable of achieving high classification accuracy[11]. This issue has gained greater attention recently, raising questions about whether AI systems should be used in high-risk situations like medicine[12,13]. Given the lack of interpretability, this issue may limit the applications of AI systems in hospitals[14]. Nowadays, many researchers focus on developing techniques that could help to understand how the deep learning models reach their decisions[15–18]. Some of them rely on the analysis of gradients[15,16], which can be difficult to interpret, while others are based on perturbation[17,18], which can be specific to a particular instance.

Since CXR exams are quick and inexpensive, being the most commonly used imaging test for the chest, its usage for the detection of Aortic Elongation is important, given that the determination this condition can give indicatives of the aforementioned diseases. By implementing such a model, low-income hospitals that might not have access to experienced and knowledgeable radiologists might lessen the burden on their medical infrastructure. Furthermore, even the most experienced expert may be susceptible to errors, thus this model can assist in handling the arduous and time-consuming task of interpreting and evaluating CXR images. By quickening the diagnostic process and lessening the workload of doctors, the integration of DL algorithms and medical knowledge can enhance the performance of medical staff and could also decrease patient waiting times. In this study, we aim to assess qualitatively and quantitatively the explainable methods in order to understand the decisions of a deep learning model for Aortic Elongation detection with CXR images.

2 Results

We trained two different network architectures (EfficientNet B4 and DenseNet121) for the aortic elongation detection. Our training process involved a two-step approach, using a public (VinDr-CXR) and a private (InCor-CXR-DB) datasets. Initially, we trained the models using the VinDr-CXR dataset, employing pre-trained weights from ImageNet. Then, we fine-tuned the models using the InCor-CXR-DB dataset. Before training, the images were properly preprocessed by first segmenting the lung regions using a U-Net network. Next, we adjusted the contrast of the resulting images and resized them appropriately. Finally, two interpretation methods (GradCAM and LIME) were applied to a subset of 70 CRX images from the InCor-CXR-DB dataset and the pixel-flipping method was employed to quantify the explainability of these methods. Details on the proposed methodology can be seen in Section 4 (Methods).

In the following, we present our obtained results for aortic classification. Then, we showed examples of the application of the LIME and Grad-CAM methods to interpret our trained models. Finally, we presented our quantitative evaluation results of such interpretation methods using the pixel-flipping method.

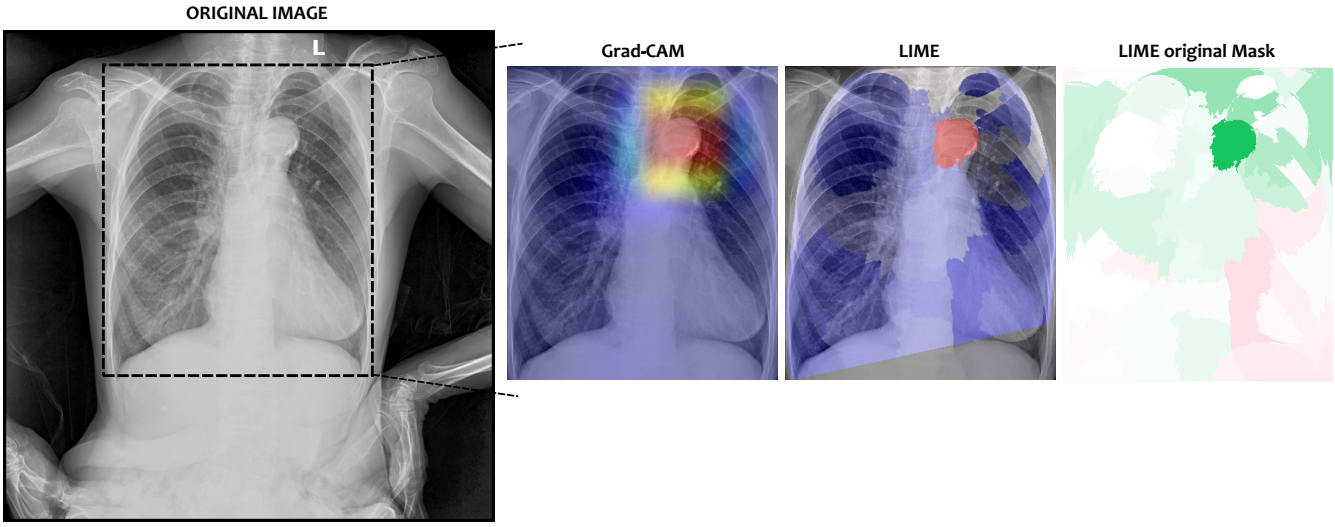
The classification performance of our models is presented in Table 1. It can be observed that the DenseNet model achieved better results than the EfficientNet model in identifying Aortic Elongation. Our DenseNet model achieved an Accuracy of 0.847 ($\pm$ 0.007), Precision of 0.756 ($\pm$ 0.013), Sensibility of 0.887 ($\pm$ 0.027), Specificity of 0.823 ($\pm$ 0.016), F1-score of 0.816 ($\pm$ 0.010) and an AUROC of 0.931 ($\pm$ 0.004). Our EfficientNet model achieved an Accuracy of 0.830 ($\pm$ 0.009), Precision of 0.771 ($\pm$ 0.020), Sensibility of 0.792 ($\pm$ 0.014), Specificity of 0.854 ($\pm$ 0.017), F1-score of 0.781 ($\pm$ 0.009) and an AUROC of 0.909 ($\pm$ 0.007). These results demonstrate the effectiveness of our approach in developing a reliable model for Aortic Elongation detection, which can potentially have significant implications for clinical practice.

Model	Acc	Prec	Se (TPR)	Spe (TNR)	F1	AUROC
EfficientNet B4	0.830 (0.009)	0.771 (0.020)	0.792 (0.014)	0.854 (0.017)	0.781 (0.009)	0.909 (0.007)
DenseNet 121	0.847 (0.007)	0.756 (0.013)	0.887 (0.027)	0.823 (0.016)	0.816 (0.010)	0.931 (0.004)

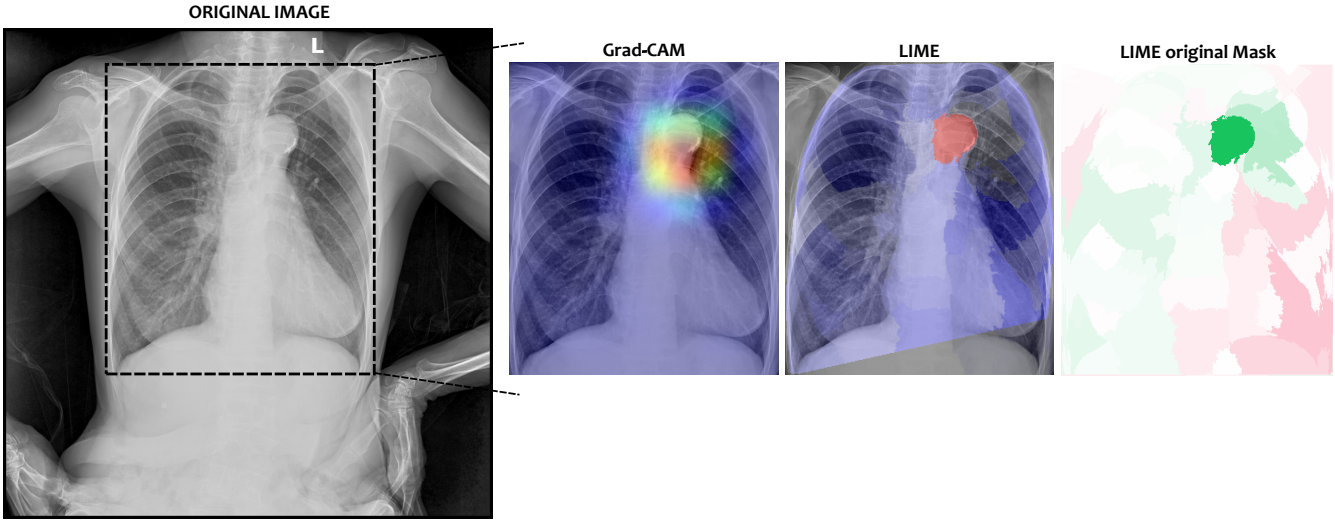
Table 1. Performance of our proposed models.

Figure 1 displays the results of the Grad-CAM and LIME methods for an exam labeled as Aortic Elongation (DenseNet prediction probability: 99.44%, EfficientNet prediction probability: 99.64%). Both explainable methods correctly identified the location of the aortic elongation.

Similarly, Figure 2 shows the results of the explainable methods for an exam labeled as non-Aortic Elongation (DenseNet prediction probability: 0.60%, EfficientNet prediction probability: 3.06%). Neither method highlighted any specific region



(a) DenseNet



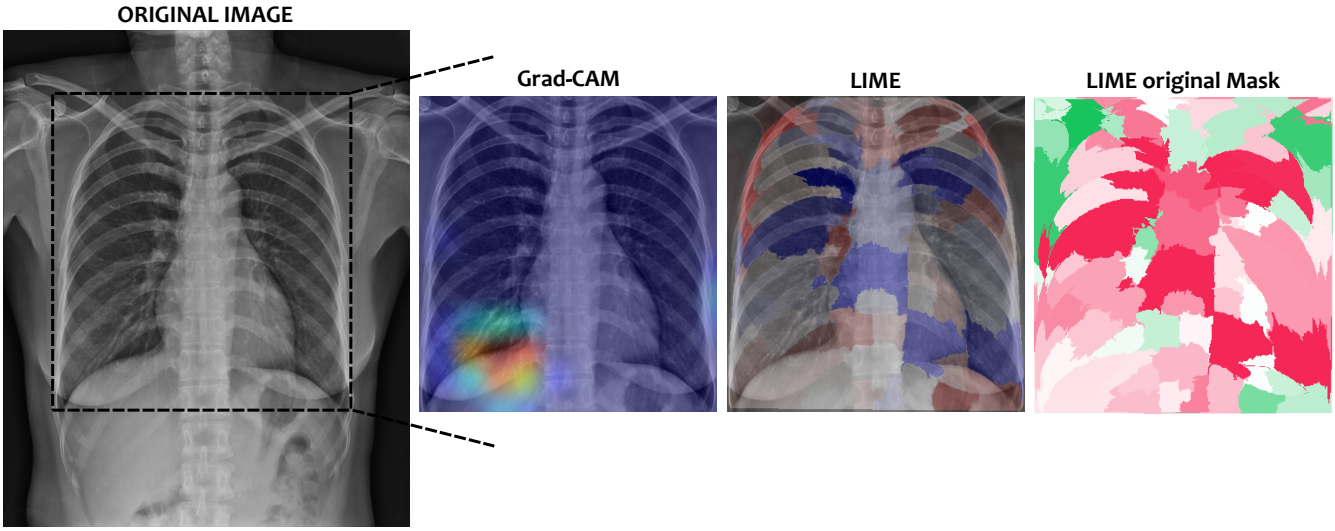
(b) EfficientNet

Figure 1. Explainable methods results for an exam correctly labeled as Aortic Elongation for both models. DenseNet prediction probability of aortic elongation: 99.44%. EfficientNet prediction probability of aortic elongation: 99.64%. In the original image, a region of interest is demarcated by a dashed box. The output of each method, Grad-CAM and LIME, is superimposed on the chest X-ray (CXR) image. The most relevant regions are highlighted in red, while the least relevant regions are indicated in blue. The "LIME original Mask" refers to the mask generated by the LIME method.

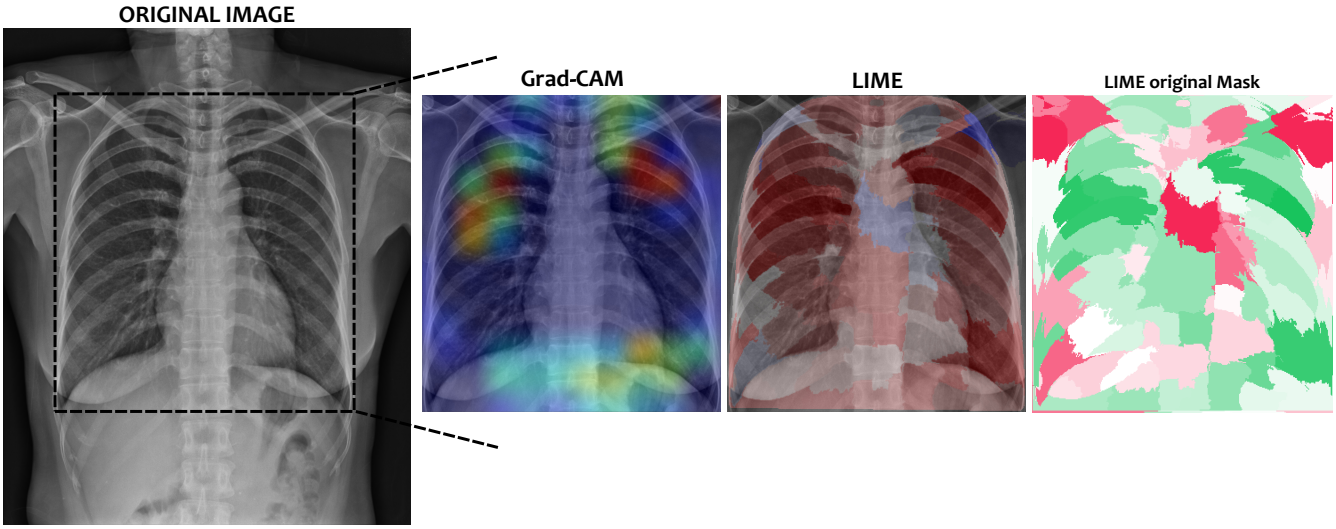
since this is a normal CXR exam. For the DenseNet model, LIME highlighted some edge locations, while the Grad-CAM highlighted the lower right region of the thorax. In the case of the EfficientNet model, LIME highlighted various places in the lung region, while the Grad-CAM highlighted regions without the heart.

Figure 3 presents the results of an exam that was wrongly predicted as Aortic Elongation but labeled as cardiomegaly in the dataset (DenseNet prediction probability: 80.16%, EfficientNet prediction probability: 92.81%). Here, all explainable methods identified the aorta region, except for the LIME method on the EfficientNet model, which highlighted the aorta region along with other parts of the chest, indicating the upper right region (near the shoulder) as the most important segment.

To assess the reliability of our studied explainable methods, we utilized the pixel-flipping method. The results of this quantitative method for our classification models are illustrated in Figure 4. There are significant differences in the results of both models. When relevant pixels/segments are removed, the DenseNet model experiences a rapid decline in its prediction performance (output score) for both Grad-CAM and LIME methods, which aligns with the desired behavior. However, for the EfficientNet model, the pixel-flipping method didn't distinctly differentiate from the baseline of random removal.



(a) DenseNet



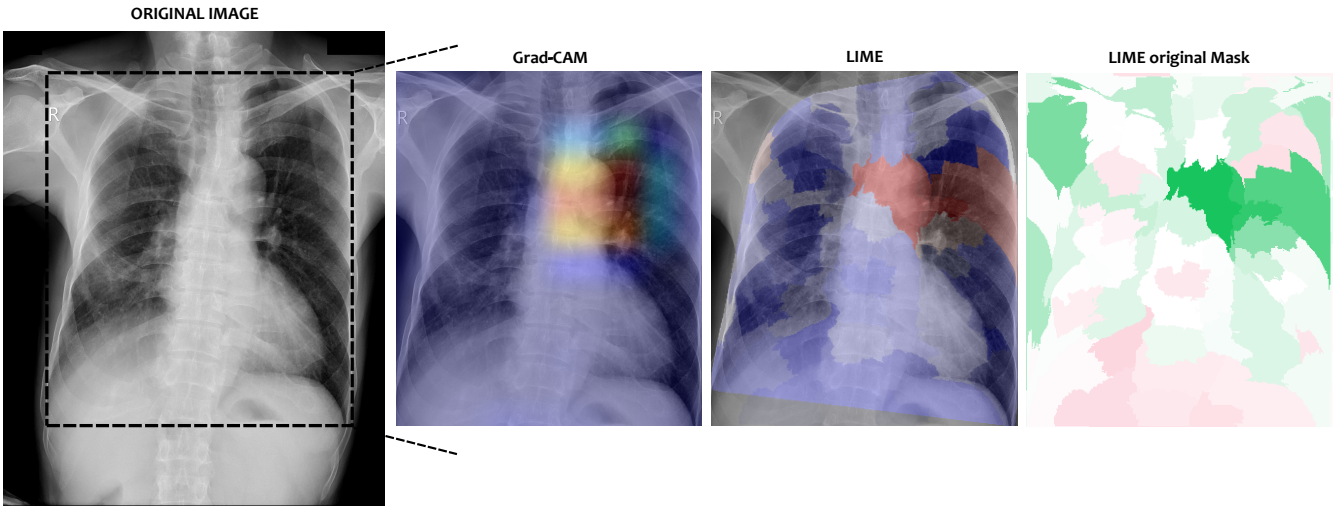
(b) EfficientNet

Figure 2. Explainable methods results for an exam correctly labeled as non-Aortic Elongation. DenseNet prediction probability of aortic elongation: 0.60%. EfficientNet prediction probability of aortic elongation: 3.06%. In the original image, a region of interest is demarcated by a dashed box. The output of each method, Grad-CAM and LIME, is superimposed on the chest X-ray (CXR) image. The most relevant regions are highlighted in red, while the least relevant regions are indicated in blue. The "LIME original Mask" refers to the mask generated by the LIME method.

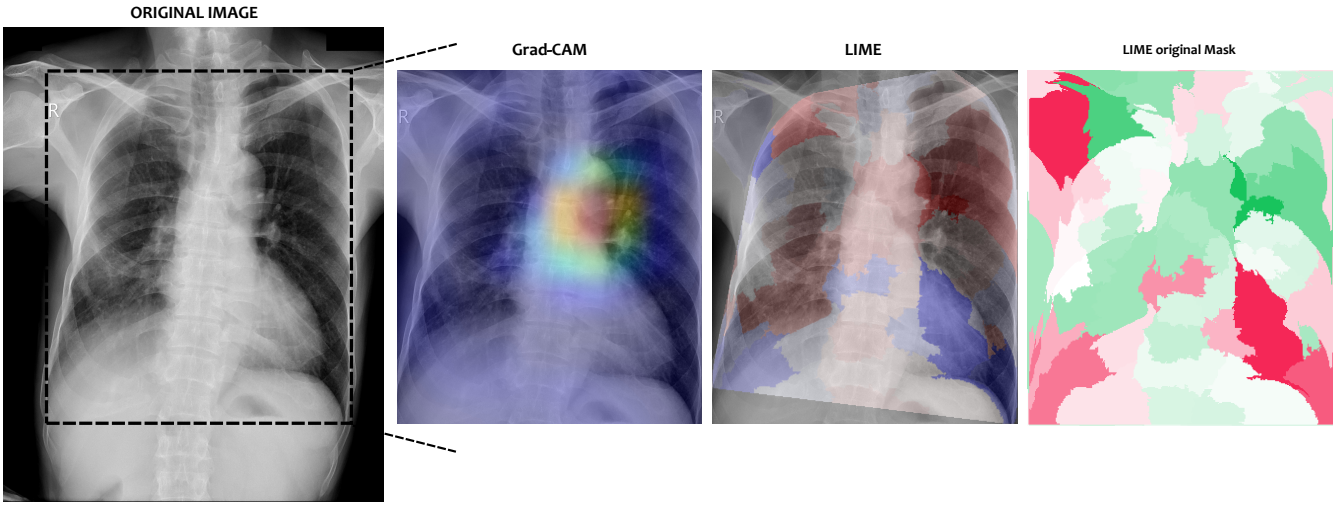
3 Discussion

We have developed a comprehensive strategy, by joining the Aortic Elongation model with the explainable methods, to analyze Chest X-rays. This approach has the potential to enhance diagnostic accuracy and aid in early detection of this cardiac condition. Through transfer learning and fine-tunning processes, our models were able to accurately identify Aortic Elongation. By using explainable methods, we can provide valuable insights for physicians into the decision making process of our models. Chest X-rays are often considered a cost-effective and time-efficient option, making them an ideal initial screening tool. By utilizing this approach, high-risk patients can be efficiently triaged during their routine care. Subsequently, appropriate lifestyle modification counseling can be provided, and further examinations or interventions can be pursued if necessary.

In a qualitative analysis, Grad-CAM and LIME explainable methods correctly pinpointed the expected aorta region on exams correctly predicted as Aortic Alongation. On the first example (Figure 1), it is interesting to observe that both models presented different explanations. One could argue that different models learns the problem (Aortic Elongation prediction, here) differently, and the visual explanations could assess if they learned the problem correctly. On the second example (Figure 2),



(a) DenseNet



(b) EfficientNet

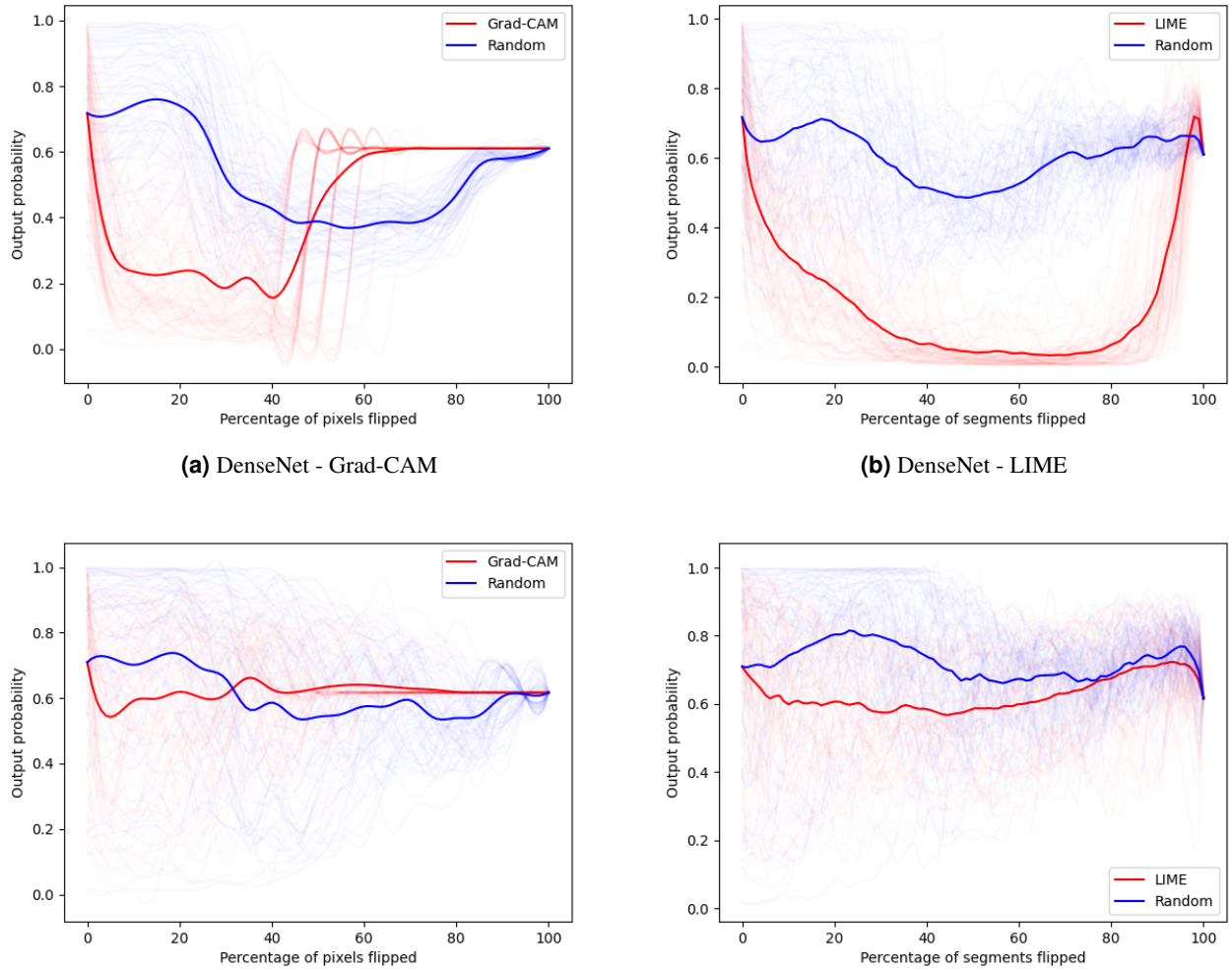
Figure 3. Explainable methods results for an exam wrongly labeled as Aortic Elongation. DenseNet prediction probability of aortic elongation: 80.16%. EfficientNet prediction probability of aortic elongation: 92.81%.

we can see that for this normal CXR the models seems to not know where to look on the image. One could suggest that the model didn't find the aorta problem, making the explainable methods highlighting spurious regions of the image. Moreover, again we can see the difference between both models, where DenseNet only highlights few segments, whereas EfficientNet indicate more regions on the CXR. In the case where the models wrongly predicted Aortic Elongation (Figure 3) both models indicated the aorta region as the most relevant.

Quantitatively, it is expected that the explainable methods which accurately highlight the image pixels/segments influencing the model's decision will exhibit a rapid decline in prediction performance (mean output score) as these pixels/segments are progressively perturbed (set to zero). To establish a baseline, random explanations (blue lines) are also included on Figure 4, as it is anticipated that perturbing image pixels/segments randomly will result in a slower decrease in performance. In addition, it is important to clarify that the results obtained through the pixel-flipping method reflect the output score of the models and not their accuracy. The output score represents the confidence or certainty assigned by the model to a specific prediction. It is distinct from accuracy, which measures the overall correctness of the model's predictions.

Based on our findings presented in Figure 4, several observations can be highlighted:

(i) The DenseNet model demonstrated the expected behavior for both explainable methods; (ii) Grad-CAM applied to the DenseNet model showed a rapid decrease in prediction performance. When 50% of the pixels were removed, the prediction performance reached a basal output score level on average; (iii) LIME performed better than Grad-CAM on the DenseNet model and significantly distinguished itself from the baseline (random removal). Removing segments rapidly led to a drop in prediction



(a) DenseNet - Grad-CAM

(b) DenseNet - LIME

(c) EfficientNet - Grad-CAM

(d) EfficientNet - LIME

Figure 4. Pixel-flipping curves for a subset of individual CXR exams (thin lines) and their average mean curves (thick line) for both explainable methods, gradually removing the segments ordered from the most relevant (red lines) or randomly removing the segments (blue lines).

performance. Removing the most important segments made the model to change prediction to non-Aortic Elongation. However, with the removal of the final segments, the performance reached the same basal level as aforementioned for the Grad-CAM; (iv) The EfficientNet model did not clearly differentiate itself from the baseline (random removal) for both explainable methods; and (v) Both models concluded with a basal output score prediction of 0.61. This suggests that if we take an CXR image and set all pixels/segments to zero, the model, on average, would classify it as Aortic Elongation using a threshold of 0.5. However, it's important to note that a completely black image (with all pixels set to zero) falls outside the domain on which the model was trained. Therefore, the model is not fitted to recognize this pattern since it hasn't learned it. These observations provide valuable insights into the behavior and performance of the DenseNet and EfficientNet models when subjected to the explainable and pixel-flipping methods.

To the best of our knowledge, there are only a limited number of studies that have focused on the automated detection of Aortic Elongation specifically from chest X-rays. It is important to note that some studies employ a multilabel paradigm, where Aortic Elongation may be included as one of the labels depending on the dataset. However, the number of works specifically targeting the detection of Aortic Elongation in chest X-rays remains relatively scarce. Rosenwasser et al. (2022)[19] used an Inception V3 and a DenseNet169 models on the VinDr-CXR dataset, relating this cardiac condition with Marfan's syndrome. Likewise, Lee et al. (2022)[20] propose the detection of aortic dissection based on a ResNet 18 CNN, using Grad-CAM to visually interpret their results.

It is indispensable that the models manage to generate explanations about their decision making. The issue of interpretability cannot be a limiting factor for systems already developed. We must make use of the great existing capacity of AI systems, which evolves every day, to generate more efficient and reliable models. Grad-CAM and LIME methods can provide valuable insights into how AI models make decisions and identify important image regions for their predictions, which can be useful for improving model transparency, fairness, and trustworthiness. Moreover, Pixel-flipping can give a quantitative metric to the explainable methods, allowing to assess their interpretations.

Taken together, our results demonstrate the effectiveness of our approach in developing a reliable model for Aortic Elongation detection. The high accuracy, precision, and AUROC scores achieved by the DenseNet model, along with the supporting visualizations from the explainable methods, underscore its potential value in clinical practice. Further research and development in this area could lead to improved diagnostic capabilities and enhance patient care in the field of cardiac imaging.

4 Methods

4.1 Study Design

This is a study aimed at learning and detecting aortic elongation using DL to analyse chest-X ray images obtained from a public available dataset and a private dataset. Moreover, we aim at assessing the proposed DL models using explainable AI methods qualitatively and quantitatively, providing insights on the models predictions. In this section, we describe the dataset, the preprocessing steps and the deep neural network architecture used for classifying CXR images into Aortic Elongation and Non-Aortic Elongation. Moreover, we describe the explainable methods used to interpret the model's results and assess qualitatively and quantitatively their interpretability.

The general structure of the proposed method is shown in Figure 5. In summary, our approach involved utilizing two CNN networks, namely DenseNet and EfficientNet. We initiated the training process by leveraging the ImageNet weights and then employed transfer learning using a publicly available CXR dataset to enable our models to learn the new domain. Subsequently, we performed fine-tuning using a private CXR dataset. To assess the performance of our models, we implemented a 5-fold cross-validation strategy. Once we gained insights into their performance, we proceeded to retrain the models using the entire dataset, without employing the cross-validation process. This allowed us to create comprehensive models that were subsequently utilized for the proposed explainable AI methods. Finally, we conducted a visual analysis of the interpretations provided by the explainable methods, in conjunction with a quantitative analysis using the pixel-flipping method, on a subset of the test set from our private dataset.

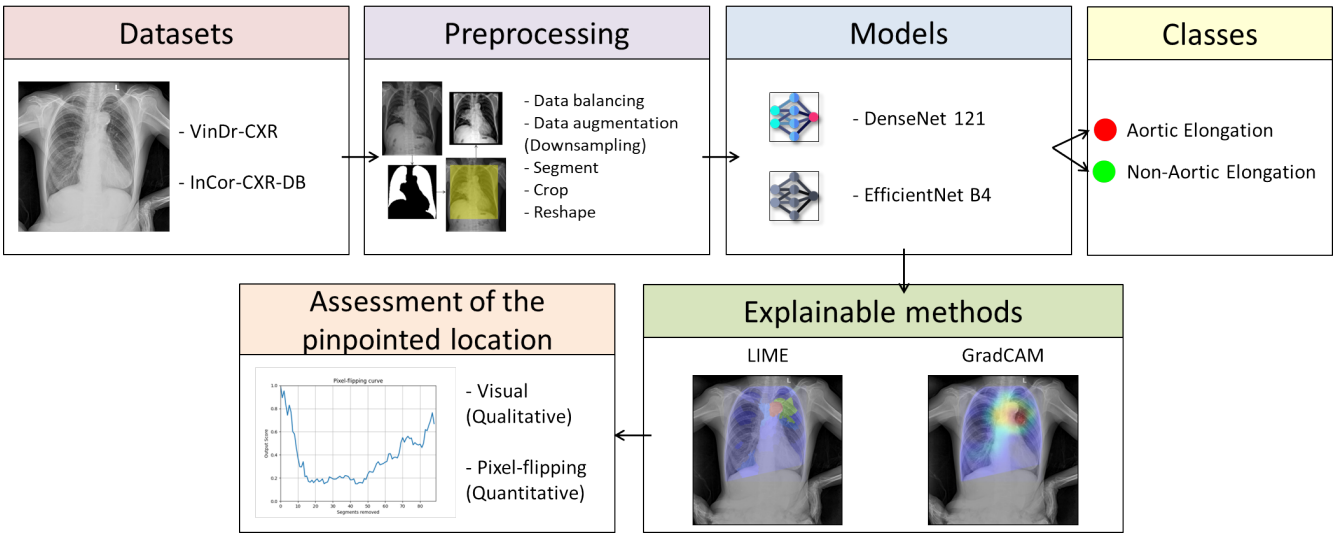


Figure 5. General structure of the proposed methodology.

4.2 Datasets

VinDr-CXR[21–23] is a dataset of CXR scans obtained retrospectively from two major hospitals in Vietnam. All images are in DICOM format and in postero-anterior (PA) view. It contains 18,000 images manually annotated by a group of radiologists. This dataset is divided into a training set with 15,000 scans independently labeled by three radiologists, and a test set with 3,000 scans labeled by the consensus of five radiologists. From these 15,000 exams, 10,606 are labeled as "Normal" and 4,376

have some abnormal condition. Thus, we downsampled this "Normal" class, randomly selecting 4,376 normal exams. Thus, our resulting dataset consists of 8,752 exams, where 2,350 exams labeled as Aortic Elongation. It should be stressed that due to the lack of patients ID information, images from the same patient may exist throughout train and test sets.

Besides the VinDr-CXR dataset, we used a private CXR dataset (InCor-CXR-DB) collected from the Picture Archiving and Communication System (PACS) of a tertiary referral hospital in Brazil specialized in cardiology (Heart Institute Hospital), with patients older than 18 years old. All images are in DICOM format in a postero-anterior (PA) view. The text-based chest radiography reports were assessed by 2 experienced observers. This dataset contains 473 exams labeled as Aortic Elongation and 760 exams labeled as non-Aortic Elongation, including equally sampled exams labeled as normal and abnormal conditions. This private dataset complies with all relevant ethical regulations and was approved by the Research Ethics Committee. Table 2 summarize the information regarding both datasets.

Table 2. Number of CRX exams labeled as Aortic Elongation from VinDr-CXR dataset and our InCor-CRX-DB private database.

Dataset	Total	Aortic Elongation	Non-Aortic Elongation
VinDr-CXR	8,752	2,350	6,402
InCor-CXR-DB	1,233	473	760

4.3 Image Preprocessing

To address the issue of the imbalanced data, we employed a downsampling technique on the larger class by randomly selecting exams labeled as non-aortic elongation. Subsequently, we applied a data augmentation technique to compensate for the reduced number of samples.

Our preprocessing steps are based on previous works[24,25] for cardiomegaly classification. In order to improve the efficiency of the training step and reduce the computational cost, we used a previously developed and validated model[26] that segments the lungs, to be able to create a binary mask of the chest cavity region, using a fine-tuned UNet-based convolutional neural network. We used the extreme points of this lung mask to create a bounding box and crop the chest cavity, segmenting only this region of interest, that was used as input in our classification model. We rescaled the intensity of the cropped image with a contrast stretching method, including all intensities that fall within 1st and 99th percentiles of the image histogram. The image is transformed into a square image using zero padding, since the networks expect square input images. Moreover, normally the pre-trained models expect images normalized and with 3-channel RGB images shape. Thus, the square image is resized to 384x384x3. Figure 6 displays the preprocessing steps described.

4.4 Deep Learning model

The small size of datasets in deep learning for analyzing medical images is a major constraint. Consequently, it is frequently difficult to train a CNN from scratch[27]. In this case, Transfer-learning is one solution. To do so, one can use a pre-trained network as a feature extractor and re-trained it with the data from the new domain, such as CXR images. Here, we initiated the training with the ImageNet weights[28]. Then, we trained the models using the VinDr-CXR dataset to learn the patterns specific to the CXR domain. Subsequently, we performed fine-tuning using our InCor-CXR-DB private dataset. Only the fully connected layers of the model were unfrozen during this fine-tuning process. Figure 7 depicts the overall proposed model based on transfer learning and fine-tunning techniques.

For our experiments, we choose two different Deep Neural Network architectures that are widely used on the classification of CXRs, being them the DenseNet 121 and the EfficientNet B4 architectures. We used the default Keras architecture topology for the networks with an input resolution of 384x384x3 for the DenseNet 121 model, and an input resolution of 380x380x3 for the EfficientNet B4 model. For the fully connected layers, we used a customized 2-layer perceptron with dropout regularization of 30%, ReLU activation in the first layer (n = 256), and a sigmoid function in the last layer. Our model was trained over 40 epochs, 16 batch size per step using Adam optimizer with a learning rate of 0.0001 and a callback to reduce it by a factor of 0.2 every six epochs in case of no improvement in validation loss. Moreover, the data were augmented via: (1) rotation, (2) horizontal flipping, and vertical flipping.

We implemented a robust 5-fold cross-validation strategy during the fine-tuning step of our model development process. To prevent any potential data leakage, we employed data augmentation techniques on each fold of the cross-validation. By doing so, we ensured that the same image did not appear in multiple folds, thus maintaining the integrity of our evaluation process. Our augmentation strategy involved applying a $2\times$ augmentation factor to the major class and an appropriately computed factor, denoted as n, to the minor class. The purpose of this approach was to balance the representation of classes and address any potential class imbalance issues.

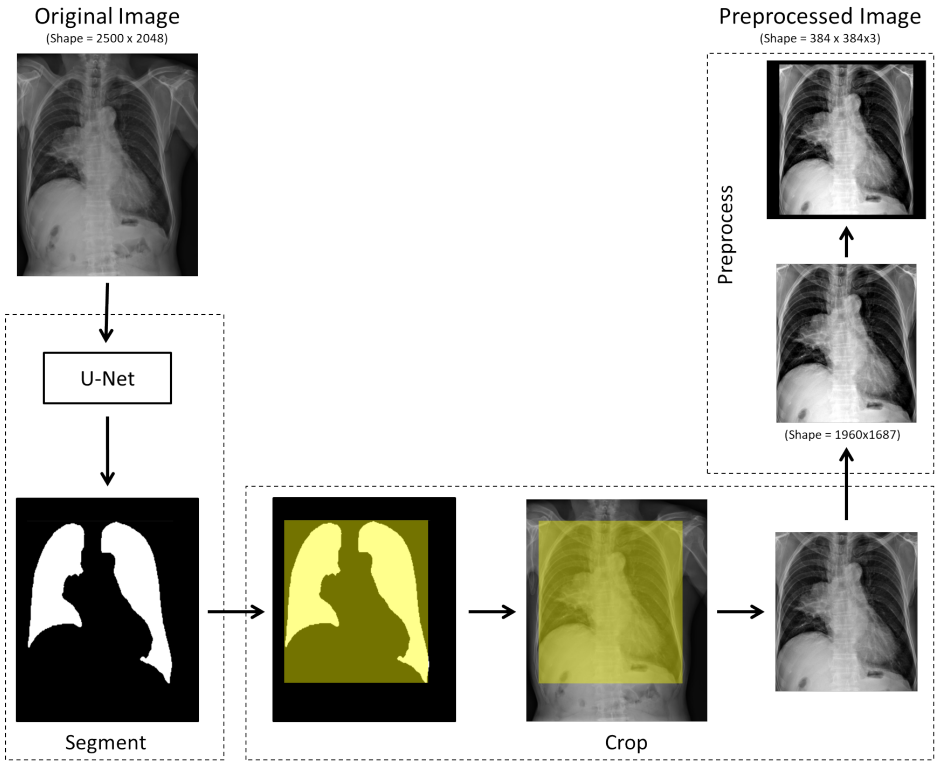


Figure 6. Steps for the proposed preprocessing task, being them: (i) chest cavity segmentation; (ii) cropping; (iii) contrast stretching; (iv) fill borders to get squared images; and (v) resize.

In order to evaluate the employed model, we assessed five different metrics, including Accuracy (Acc), Precision (Prec), Sensitivity (Se), Specificity (Spe), F1-score (F1), and Area Under Operational Receipt Curve (AUROC).

Our experiment was performed using a Foxconn High-Performance Computer (HPC) M100-NHI with an 8 GPU cluster of 16 GB NVIDIA Tesla V100 cards. The methodology was implemented using the Python framework and Keras v2.4.3 with TensorFlow backend v2.3.0.

4.5 Explainable AI methods

We selected two widely used local explainable methods, i.e., Grad-CAM and LIME, to examine individual CXR images. LIME (Local Interpretable Model-agnostic Explanations)[17] and Grad-CAM (Gradient-weighted Class Activation Mapping)[16] are two methods used in explainable AI, which provide insights into the decision-making process of complex machine learning models. Visualizations result in a map with the most significant parts are highlighted in red color for both Grad-CAM and LIME methods.

LIME[17] is a model-agnostic method that explains the predictions of any machine learning model in a way that is easily interpretable to humans. LIME works by creating a local linear model around a specific data point to approximate how the original model makes its predictions. It then identifies the features, or superpixels (segments on the image instance) in the context of image inputs, of the input that are most important in determining the outcome of the prediction. In our work, we used a linear regression as the surrogate model, and segmented the images with a *quickshift* segmentation method from the skimage segmentation library on python, using the following parameters: kernel_size = 5; max_dist = 100; and ratio = 0.5. Moreover, we used 1000 perturbations to perform LIME.

On the other hand, Grad-CAM[16] is a visualization method that helps in understanding the regions of an image that contribute to the prediction made by a convolutional neural network (CNN). It uses the gradients of the predicted class with respect to the feature maps of the last convolutional layer to generate a heatmap that highlights the regions of the input image that are most relevant to the final prediction.

Both LIME and Grad-CAM are important tools for explainable AI as they help in identifying the features or regions of the input that the model is relying on to make its predictions. These methods can be used to identify potential errors in the model, identify biases, or simply to provide insights into how the model is processing the input data. By providing a more transparent and interpretable view into the inner workings of machine learning models, LIME and Grad-CAM help in building

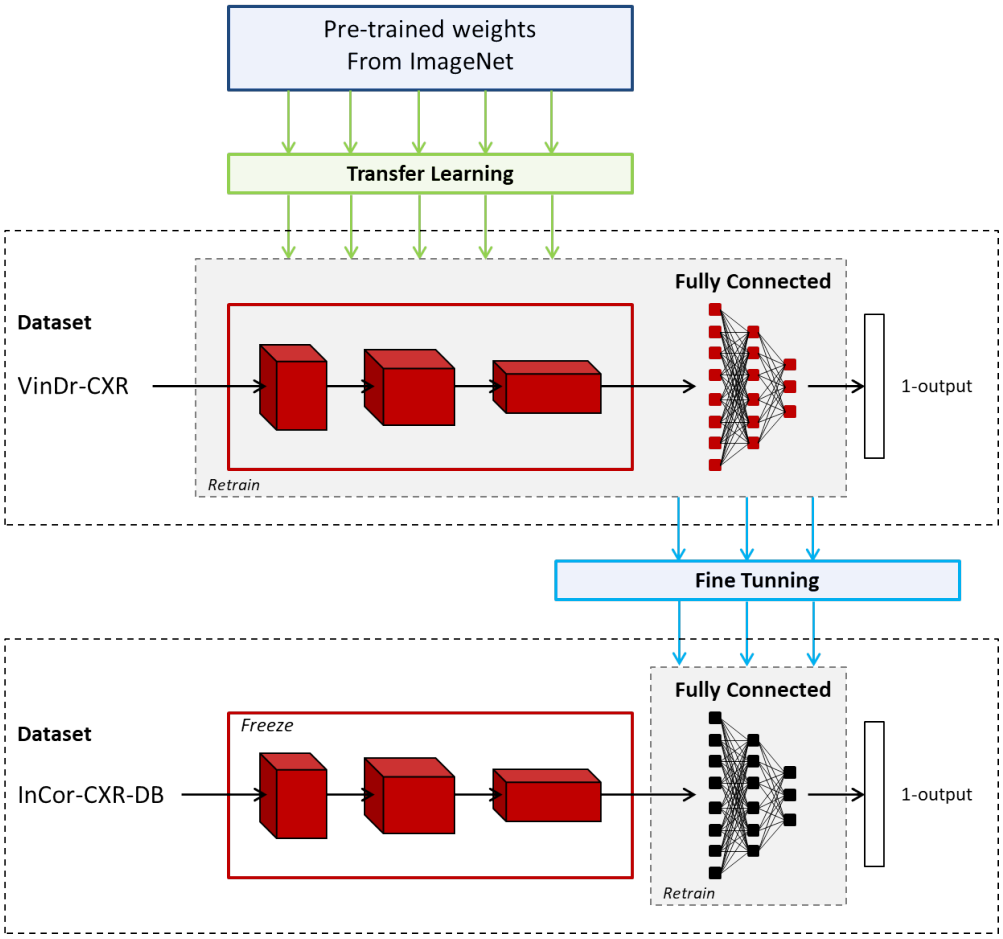


Figure 7. The transfer learning and fine-tuning techniques used in the development of the proposed CNN models.

more trustworthy and reliable AI systems.

Moreover, first we utilized a cross-validation approach to estimate the performance of our models, which involved splitting our data into multiple subsets to train and evaluate the models iteratively. Once we obtained an understanding of their performance, we proceeded to retrain the models using the entire dataset, without the cross-validation process. This allowed us to create complete models that we then applied to perform both the proposed explainable AI methods aforementioned.

4.6 Quantification of Explainable AI methods

The pixel-flipping method[29,30] is a technique used to quantify the explainability of AI methods that operate on image data, particularly in the context of computer vision tasks such as image classification. The goal of this method is to assess how sensitive the AI model's predictions are to changes in individual pixels of an input image, which can provide insights into how the model makes decisions and which image regions are important for its predictions. The pixel-flipping method typically involves the following steps:

1. Input Image Selection: A set of input images from the InCor-CXR-DB dataset used to test the AI model are selected for analysis. These images should represent the task of interest and cover a range of different samples to ensure a comprehensive assessment. Here we selected a subset of the InCor-CXR-DB test set with 70 CXR exams labeled as Aortic Elongation;

2. Perturbation: Each selected image is then perturbed by flipping the value of individual pixels, either by inverting their intensity (e.g., changing a pixel from black to white or vice versa) or by changing their color value (e.g., changing a pixel from red to green or vice versa). In this work we set a pixel to zero for Grad-CAM method, or set a segment to zero for LIME method. This perturbation is done based on their importance as determined by the explainable methods, ordered from the most relevant to the least according to the particular explanation, showing how quickly the prediction score decreases. Likewise, for a baseline, we randomly selected the pixels or segments for comparison;

3. Prediction Evaluation: The perturbed images are then fed into the DL model, and the resulting predictions are compared to the original predictions on the unperturbed image. The discrepancy between the original and perturbed predictions is used as a measure of the model's sensitivity to pixel-level changes. For example, if flipping a certain pixel significantly changes the model's prediction, it indicates that the model may rely heavily on that pixel for its decision-making process. In this work we present the curves of the DL model output score for the perturbed images;

4. Interpretation: Finally, the results of the pixel-flipping analysis can be interpreted to gain insights into the model's behavior. For example, if certain pixels have a high impact on the model's predictions, it may indicate that those pixels are important for the model's decision-making process and should be further investigated. This information can help understand the inner workings of the DL model and identify potential biases or limitations in its decision-making process.

Overall, the pixel-flipping method is a useful tool for quantifying the explainability of AI methods in computer vision tasks by assessing their sensitivity to pixel-level changes. Here, for LIME we gradually removed (set a segment to zero) the segments of an individual input image, ordered from the most relevant to the least according to the particular explanation, showing how quickly the prediction score decreases. For Grad-CAM, we removed $10, 20, \ldots, 100\%$ of the pixels, based on the most relevant on the resulting heatmap. Pixel-flipping curves are computed for a subset of the test set of individual images, and then averaged to create a mean curve. Finally, for comparison (baseline), we performed pixel-flipping by randomly removing the segments and pixels for LIME and Grad-CAM methods.

Acknowledgements

This study was supported by São Paulo Research Foundation (FAPESP) – grant n° 2021/12935-0, the Zerbini Foundation, and Foxconn Brazil, as part of the research project "Machine Learning in Cardiovascular Medicine".

Ethics Statement

This research was approved by the Internal Review Board (IRB), registration CAAE 45070821.3.0000.0068, as part of the Machine Learning in Cardiovascular Medicine Project.

Author contributions statement

E.R. Conceptualization, Methodology, Implementation and Writting. D.A.C.C. Conceptualization, Methodology, Implementation and Writting. F.M.D. Methodology and Writting. J.E.K. and M.A.G. acted as project leaders. All authors analyzed the results and revised critically the manuscript. All authors read and approved the submitted manuscript.

Additional information

Competing interests: The authors declare no competing interests.

References

1. A decade of discovery in the genetic understanding of thoracic aortic disease. Can. J. Cardiol. **32**, 13–25, DOI: 10.1016/j.cjca.2015.10.017 (2016).

2. Adriaans, B. P. et al. Aortic elongation part i: the normal aortic ageing process. Heart **104**, 1772–1777, DOI: 10.1136/heartjnl-2017-312866 (2018).

3. Graham Stuart, A. & Williams, A. Marfan's syndrome and the heart. Arch. Dis. Child. **92**, 351–356, DOI: 10.1136/adc.2006.097469 (2007).

4. Akutsu, K. et al. Vascular involvements are common in the branch arteries of the abdominal aorta rather than in the aorta in vascular ehlers-danlos syndrome. CJC Open **5**, 72–76, DOI: https://doi.org/10.1016/j.cjco.2022.11.001 (2023).

5. Callewaert, B., Malfait, F., Loeys, B. & De Paepe, A. Ehlers-danlos syndromes and marfan syndrome. Best Pract. / Res. Clin. Rheumatol. **22**, 165–189, DOI: 10.1016/j.berh.2007.12.005 (2008). Orphan Skeletal Diseases.

6. MacCarrick, G. et al. Loeys–dietz syndrome: a primer for diagnosis and management. Genet. Medicine **16**, 576–587, DOI: 10.1038/gim.2014.11 (2014).

7. Bondy, C. A. Aortic dissection in turner syndrome. Curr. Opin. Cardiol. **23**, 519–526, DOI: 10.1097/HCO.0b013e3283129b89 (2008).

8. Goldfinger, J. Z. et al. Thoracic aortic aneurysm and dissection. J. Am. Coll. Cardiol. **64**, 1725–1739, DOI: 10.1016/j.jacc.2014.08.025 (2014).

9. O'Shea, K. & Nash, R. An introduction to convolutional neural networks (2015). 1511.08458.

10. Yamashita, R., Nishio, M., Do, R. K. G. & Togasho, K. Convolutional neural networks: An overview and application in radiology. Insights into imaging **9**, 611–629, DOI: 10.1007/s13244-018-0639-9 (2018).

11. Molnar, C. Interpretable Machine Learning (2019).

12. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**, 206–215, DOI: 10.1038/s42256-019-0048-x (2019).

13. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence n health care. Lancet Digit. Heal. **3**, e745–e750, DOI: 10.1016/S2589-7500(21)00208-9 (2021).

14. Kundu, S. AI in medicine must be explainable. Nat. Medicine **27**, 1328–1328, DOI: 10.1038/s41591-021-01461-z (2021).

15. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps (2014). ArXiv: 1312.6034, 1312.6034.

16. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In 2017 IEEE International Conference on Computer Vision (ICCV), 618–626, DOI: 10.1109/ICCV.2017.74 (2017).

17. Ribeiro, M. T., Singh, S. & Guestrin, C. "why should i trust you?": Explaining the predictions of any classifier (2016). ArXiv: 1602.04938, 1602.04938.

18. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions (2017). ArXiv: 1705.07874, 1705.07874.

19. Aortic enlargement detection using chest x-rays to identify potential marfan syndrome. Procedia Comput. Sci. **207**, 2125–2133, DOI: 10.1016/j.procs.2022.09.272 (2022). Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 26th International Conference KES2022.

20. Lee, D. K. et al. Detection of acute thoracic aortic dissection based on plain chest radiography and a residual neural network (resnet). Sci. Reports **12**, 21884, DOI: 10.1038/s41598-022-26486-3 (2022).

21. Nguyen, H. Q. et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations (2020). 2012.15029.

22. Nguyen, H. Q. et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations (version 1.0.0), DOI: https://doi.org/10.13026/3akn-b287 (2020).

23. Pham, H. H., Nguyen, H. Q., Nguyen, H. T., Le, L. T. & Khanh, L. An accurate and explainable deep learning system improves interobserver agreement in the interpretation of chest radiograph. IEEE Access **10**, 104512–104531, DOI: 10.1109/ACCESS.2022.3210468 (2022).

24. Cardenas, D. et al. Multicenter validation of convolutional neural networks for automated detection of cardiomegaly on chest radiographs. In Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde, 179–190, DOI: 10.5753/sbcas.2020.11512 (SBC, Porto Alegre, RS, Brasil, 2020).

25. Junior, J. R. F. et al. A general fully automated deep-learning method to detect cardiomegaly in chest x-rays. In Mazurowski, M. A. & Drukker, K. (eds.) Medical Imaging 2021: Computer-Aided Diagnosis, vol. 11597, 537 – 542, DOI: 10.1117/12.2581980. International Society for Optics and Photonics (SPIE, 2021).

26. Pazhitnykh, I. & Petsiuk, V. Lung segmentation 2d (2017).

27. Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T. & Saalbach, A. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. Sci. Reports **9**, 6381, DOI: 10.1038/s41598-019-42294-8 (2019).

28. Russakovsky, O. et al. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. **115**, 211–252, DOI: 10.1007/s11263-015-0816-y (2015).

29. Samek, W., Binder, A., Montavon, G., Lapuschkin, S. & Müller, K.-R. Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks Learn. Syst. **28**, 2660–2673, DOI: 10.1109/TNNLS.2016.2599820 (2017).

30. Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J. & Müller, K.-R. Explaining deep neural networks and beyond: A review of methods and applications. Proc. IEEE **109**, 247–278, DOI: 10.1109/JPROC.2021.3060483 (2021).