Computers in Biology and Medicine 145 (2022) 105442

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A ensemble methodology for automatic classification of chest X-rays using deep learning☆

Luis Vogado [a,*], Flávio Araújo [b], Pedro Santos Neto [a], João Almeida [c], João Manuel R.S. Tavares [d], Rodrigo Veras [a]

[a] Departamento de Computação, Universidade Federal do Piauí, Teresina, Brazil
[b] Curso de Bacharelado em Sistemas de Informação, Universidade Federal do Piauí, Picos, Brazil
[c] Departamento de Informática, Universidade Federal do Maranhão, São Luís, Brazil
[d] Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial, Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

ARTICLE INFO

2010 MSC:
00-01
99-00

Keywords:
Image analysis
Machine learning
Image classification
Computer aided diagnosis

ABSTRACT

Chest radiographies, or chest X-rays, are the most standard imaging exams used in daily hospitals. Responsible for assisting in detecting numerous pathologies and findings that directly interfere in the patient's life, this exam is therefore crucial in screening patients. This work proposes a methodology based on a Convolutional Neural Networks (CNNs) ensemble to aid the diagnosis of chest X-ray exams by screening them with a high probability of being normal or abnormal. In the development of this study, a private dataset with frontal and lateral projections X-ray images was used. To build the ensemble model, VGG-16, ResNet50 and DenseNet121 architectures, which are commonly used in the classification of Chest X-rays, were evaluated. A Confidence Threshold (CTR) was used to define the predictions into High Confidence Normal (HCn), Borderline classification (BC), or High Confidence Abnormal (HCa). In the tests performed, very promising results were achieved: 54.63% of the exams were classified with high confidence; of the normal exams, 32% were classified as HCn with an false discovery rate (FDR) of 1.68%; and as to the abnormal exams, 23% were classified as HCa with 4.91% false omission rate (FOR).

1. Introduction

Imaging exams are practical tools to aid in the diagnosis of diseases. Among them, radiography or X-ray examination is a low-cost and easy-to-operate technique. Due to its low cost, it is the most performed imaging exam in the world. For example, according to the National Health Service, in England, it has been performed about 16 million times between April 2020 and March 2021.[1] Present in developing or difficult-to-access regions, it is commonly used as a primary diagnostic tool, allowing specialists to observe pathologies that are difficult to trace. X-ray is one of the few imaging modalities that cover all regions of the human body. Among the main areas, the chest is where several pathologies associated mainly with the lungs and heart are found, such as pneumonia, pleural effusion, cardiomegaly and pulmonary nodules.

The development of computational methodologies that identify the aforementioned diseases would enable the development of computer-aided diagnostic (CAD) system that can assist in detecting and following up patients. In recent years, several works based on deep learning models have been successfully applied to problems commonly found in the medical field [1–3]. For example, mainly due to the pandemic caused by COVID-19 [4], researchers used techniques based on deep learning models to identify patients infected with this disease in computed tomography (CT) [5] and X-ray images [6–8].

Hospitals can use CAD systems to reduce costs and help to prioritize more extreme care, speeding up service lines. However, to provide such benefits, these systems must have an error rate close to 1% from a

☆ ☆ Fully documented templates are available in the elsarticle package on CTRAN.

* Corresponding author.
E-mail addresses: lhvogado@gmail.com (L. Vogado), flavio86@ufpi.edu.br (F. Araújo), pasn@ufpi.edu.br (P.S. Neto), jdallyson@nca.ufma.br (J. Almeida), tavares@fe.up.pt (J.M.R.S. Tavares), rveras@ufpi.edu.br (R. Veras).

[1] https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/diagnostic-imaging-dataset-2020-21-data (accessed in September 2021.).

https://doi.org/10.1016/j.compbiomed.2022.105442
Received 23 December 2021; Received in revised form 12 March 2022; Accepted 20 March 2022
Available online 23 March 2022

Table 1
Summary of the related works found in the literature.

Work	Year	Methodology	Number of Images	Availability	Performance results
Normal/Abormal Classification					
Yates et al. [15]	2019	InceptionV3 + Fine-tuning	53,149	Public	Acc of 94.6%
Dunnmon et al. [17]	2019	DenseNet121	216,431	Private	Acc of 91%
Ellis et al. [16]	2020	DenseNet121	7000	Private	Acc of 82%
Wong et al. [18]	2020	VGG16+ResNet50 pyramid	128,886	Public	AUC of 0.821
Tang et al. [19]	2020	ResNet18	141,617	Public	Acc of 94.64%
Dyer et al. [20]	2021	Ensemble (DenseNet121 and EffIcientNet b4)	3887	Private	15% of all examinations with HCn of 97.7%

medical point of view [9]. In addition, there are other challenges for the correct prediction of exams, such as: obtaining technically limited exams that tend to make diagnosis difficult, problems related to resolution, positioning, and the existence of exams produced using different equipment.

Besides to the challenges, there are numerous pathologies or changes detectable in chest X-rays. However, specific abnormal findings are located in close regions or have similar characteristics, making it difficult to identify them correctly. Another adversity for developing CAD systems is the acquisition of properly labeled datasets or errors from the PACS system, such as receiving other types of exams. Given the numerous challenges for building a methodology that can generalize all existing changes, a new methodology that tends to help reduce errors and consequently increase accuracy is proposed.

The proposed methodology has as main objective to provide high confidence chest X-ray predictions for normal (High Confidence Normal - HCn) and abnormal (High Confidence Abnormal - HCa) cases, i.e., classes. Initially, we pre-process the images and perform deep fine-tuning of pre-trained architectures. We used an ensemble with the architecture originating from the frontal projection (posteroanterior (PA)) and another one trained only with lateral projections in order to make predictions with high confidence. Based on the ensemble's results, we used confidence factors to define which exams belong to the HCn and HCa classes. Here, the normal class is represented by exams that do not have any finding or alteration, regardless of whether they are benign. In contrast, an abnormal class represents the complement of the normal class.

Among the main contributions of this study, we can highlight: (1) evaluation in a heterogeneous database with different abnormalities and findings that are not present in state-of-the-art public databases; (2) pre-processing task for resizing images with different resolutions and preserving the input size ratio; (3) new ensemble methodology using different state-of-the-art Convolutional Neural Network (CNN) architectures and different projections; (4) new evaluation methodology considering the probability of CNNs to generate ratings based on confidence factors; and (5) proposal of a fully automatic solution that can be easily implemented in different medical facilities, mainly in hospitals.

The remainder of this article is organized as follows. Related state-of-the-art works are presented in Section 2, and the dataset used in the experiments is described in Section 3.1. The proposed methodology is detailed in Section 3.2, as well as the built classifier models. In Sections 4 and 5, results and discussions about the challenges and comparisons with related works found in the literature are presented. Finally, in Section 6, the main contributions and conclusions achieved in this study are pointed out.

2. Related works

The current literature as to the development of methodologies that identify changes in chest X-rays is vast. According to Çallı et al. [10], most of the published works presented methodologies based on image classification. Among the primary approaches studied, we can highlight the screening of exams and their binary classification and the prediction/detection of specific abnormalities. Therefore, we surveyed works in the literature aligned with the objective proposed in this study.

One of the main difficulties in large hospitals is the patient screening system. Difficulties generated by the delay in patients care can lead to severe consequences in some cases. Therefore, diverse authors have proposed methodologies for differentiating healthy and abnormal chest X-rays. The proposed solutions intend to streamline patient screening automatically. However, we observed that in the literature, the primary methodologies still have limitations regarding the number of alterations found in chest exams [11–13]. This difficulty is also related to the availability of datasets in the literature. The ChestX-ray14 dataset [14], with about 112,000 images and 13 types of pathologies, is the most popular.

Among the works analyzed, we can highlight the work of Yates et al. [15], were two distinct datasets were used to build the two classes that were addressed separately with 94.6% of accuracy. It is noteworthy that this approach favors getting high correctness rates since each dataset has a distinct characteristic in terms of resolution, imaging equipment and applied pre-processing techniques, which tends to favor the learning of the used CNN. The work proposed by Ellis et al. [16] was the only one among those analyzed that applied lateral and frontal projections in the classification of the exams, concatenating both images for later classification with a CNN. In Dunnmon et al. [17], the authors used a very extensive evaluated dataset with 216,431 images and obtained 91% of accuracy using the DenseNet121 architecture. In Wong et al. [18], the authors used the concatenation of CNNs to develop a multi-model feature pyramid approach; two databases were evaluated, and a Receiver operating characteristic (ROC) of 0.821 was obtained.

In Tang et al. [19], the authors proposed evaluating different CNNs for the classification of frontal X-ray exams into healthy or abnormal. The evaluated VGG-19, ResNet18, ResNet50, InceptionV3 and DenseNet121 architectures did not present significant differences in their results. Furthermore, images with different resolutions, ranging from 256 $\times$ 256, 512 $\times$ 512 and 1024 $\times$ 1024 pixels, were evaluated and the results were not significantly different. However, among the used architectures, the one that obtained the highest average of results was ResNet18, with a accuracy of 94.64%.

In the work of Dyer et al. [20], an algorithm for identify healthy X-ray exams with a high confidence factor was presented. The approach intended to reduce the workload of radiologists, offering results with a high probability of being healthy. The methodology is based on an ensemble formed by DenseNet and EfficientNet B4 architectures, which reduced by up to 15% the number of exams evaluated by the physician with an HCn with an error of 2.3%. The authors used 3887 images to develop their solution.

When analyzing the problem involving the classification into healthy and abnormal, we observed that the best performance achieved by different methodologies was not reached by combining various abnormalities, which is critical for implementation in actual conditions of use, such as the ones found in common hospitals and clinics [21].

Table 1 presents the state-of-the-art methodologies found with particular relevance in this study.

From Table 1, one can realize that some of the works found in the literature used private datasets and others used public ones. However, none of the works used datasets with a good balance between the healthy and pathological classes. This confirms the demand for datasets that better represent both classes. Even in the ChestX-ray 14 dataset,

Table 2
Details about the dataset used in the experiments.

	PA	Lateral	Labels	Labeling method	Binary classification
Built Dataset	224,042	128,418	2	RP, RIR	Normal: 236,350 Abnormal: 116,110

which has a large number of images with pathologies, the "no findings" class does not guarantee that the exams are normal. In addition, the heterogeneity of the used dataset is crucial for obtaining a model capable of generalizing different pathologies in real scenarios, where any finding can appear. This heterogeneity is only present in the works of Tang et al. [19] and Yates [15], where the proposed methodologies were developed and evaluated using multiple datasets. Even so, these authors did not explore the use of lateral projection images, which can be decisive for an efficient diagnosis.

In the found state-of-the-art works, a common factor is the use of fine-tuning techniques in convolutional neural networks and the evaluation of different architectures. The authors achieved results with consolidated state-of-the-art CNNs, such as DenseNet, ResNet, and Inception. This demonstrates that, despite the proposal of numerous CNNs, for implementation in real systems in order to aid the diagnosis, the consolidated architectures are the ones that tend to achieve better effectiveness.

The critical point is that the authors did not present evidence that the proposed methodologies can be used in real scenarios. The exception is the proposal by Dyer et al. [20], which prioritized the accuracy of the methodology to assist the radiologist in issuing normal reports. Therefore, observing the need for approaches with reliable answers, consolidated architectures, heterogeneous datasets, and the use of all incidences, we developed a methodology based on committees with CNNs to help radiologists with the production of predictions with high precision.

3. Materials and methods

This section describes the development of the proposed methodology. Hence, we emphasize the images dataset used in the experiments and its labeling method, the deep learning approaches employed, and the evaluated architectures. In the end, we present the evaluation methodology that seeks to validate the results obtained according to the main metrics found in the literature.

3.1. Image dataset

The need for large amounts of data to train methodologies based on the Deep Learning paradigm is well known. However, there are other challenges to the development of diagnostic aid tools for real world use. One of the main problems are the data labeling and pre-processing. Among the public datasets used in the found state-of-the-art works, there are inaccuracies regarding data labeling, especially when multiple classes are taken into account. Another fact that usually affects public databases is the overlap and dependency between the studied classes [10].

The data collection was performed in 84 Brazilian hospitals, totalizing 217,302 collected exams. The dataset has 352,460 anonymous images, and the image resolutions range from 727×692 to 4892×4020 pixels. The exams were obtained in DICOM format and converted into ".png" format for further processing. The Photometric Interpretation attribute came with the Monochrome1 parameter for some exams, making that the lowest pixel value is displayed by the color white, contrasting most of the exams that came with Monochrome2. Therefore, the images have been adjusted to satisfy the Monochrome2 standard.

The exams were obtained according to the frontal (anteroposterior/posteroanterior) and lateral views. It is worth noting that it is not always that the protocols require the physician to request the lateral projection. Thus, all exams have at least one frontal image and one or no lateral image, in a total of 224,042 images of frontal projection and 128,418 of lateral projection. In Table 2, the main characteristics of the used dataset are presented. Fig. 1 gives examples of the included chest X-rays, with

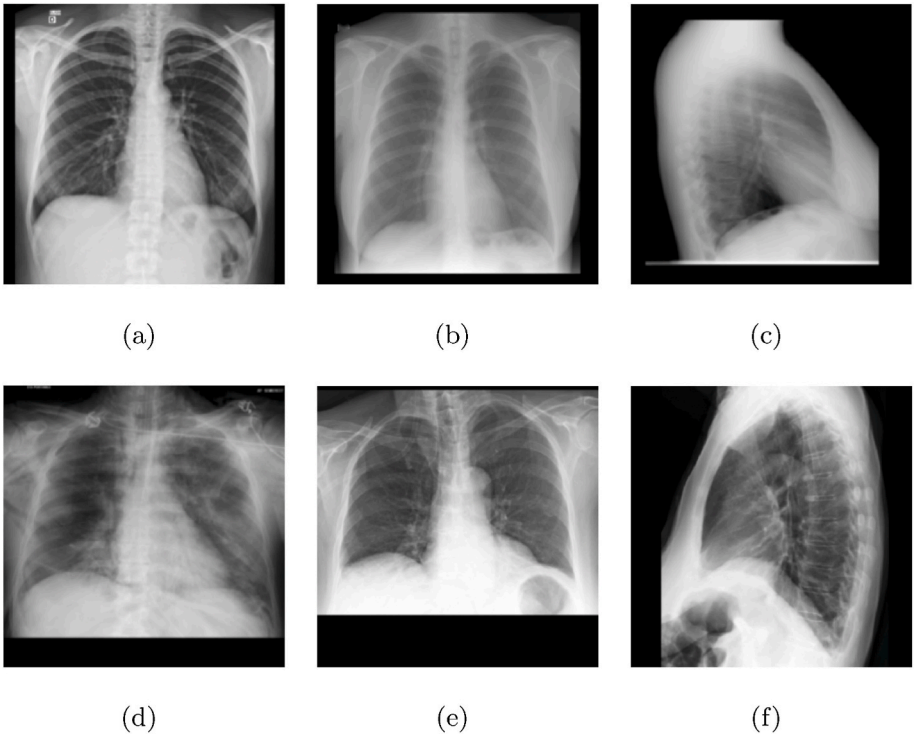


Fig. 1. Examples of normal (a)–(c) and abnormal (d)–(f) chest x-ray images belonging to the used dataset.

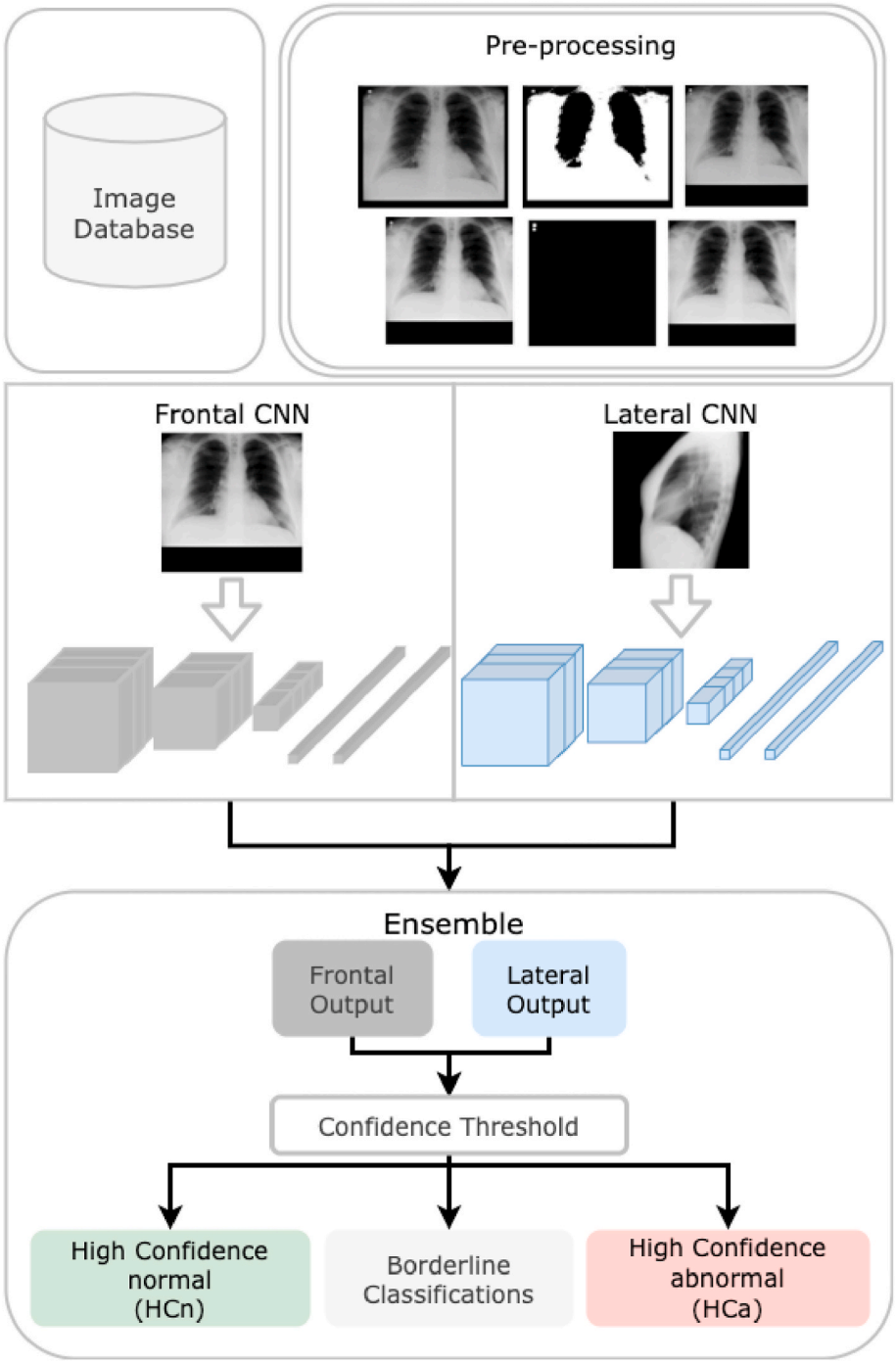


Fig. 2. Flowchart of the proposed classification methodology.

their frontal and lateral projections.

We used two labeling methodologies for the exams included in the dataset. The first was the Report Interpretation Radiologist (RIR) [10], where the specialist analyzes the medical report and classifies it according to content. This methodology was used for all exams in the normal class, and for part of the abnormal exams. The second used labeling methodology was Report Parsing, where we applied automatic techniques to classify the reports. In this work, we used a dictionary collected from radiologists with terms that denote abnormalities or findings, and we searched the reports for terms that were not classified as normal in the first methodology. A fact to be highlighted is the definition of the term "abnormal". Here, since we intended to simulate the most accurate medical knowledge, we used any term that escapes normality as abnormal. Thus, unlike other datasets, findings such as calcified nodules, granulomas, accessions and pacemakers are considered abnormal.

3.2. Proposed methodology

In this study, we propose a methodology based on an ensemble of CNNs to aid the diagnosis of chest X-ray exams through screening exams with a high probability of being normal or abnormal, Fig. 2. Initially, the images go through a pre-processing task in order to adapt them for the input of CNNs. We refined and evaluated three widely used CNNs architectures for Chest X-ray classification, ranked the exams according to the confidence factor, and defined three classes of responses according

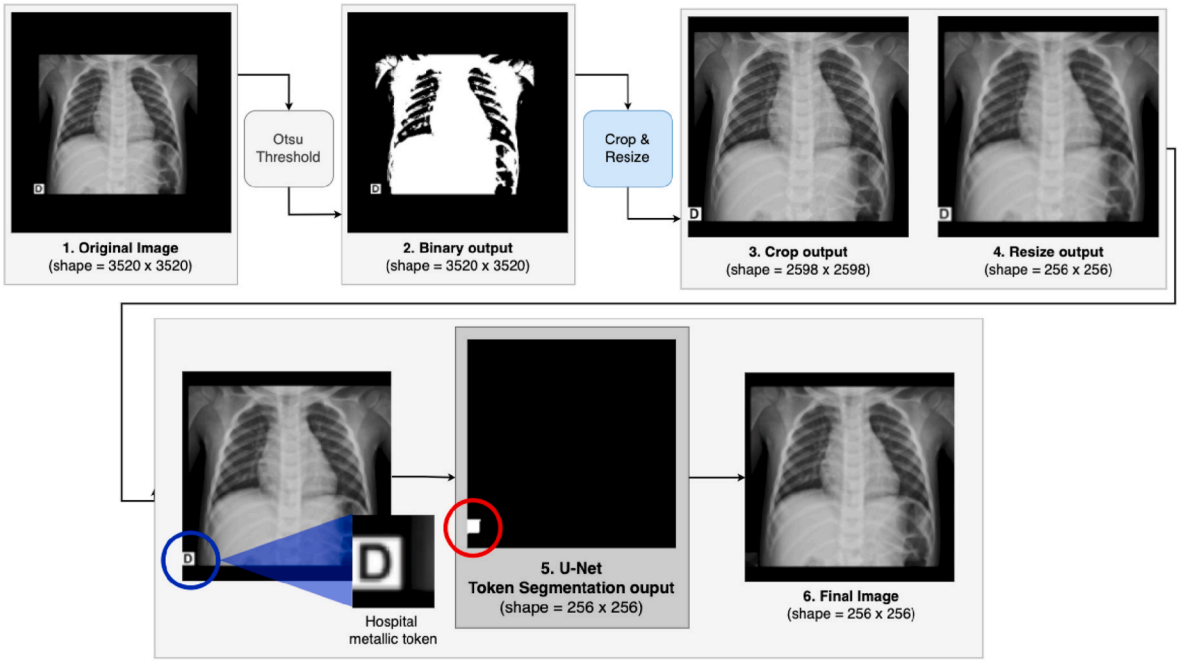
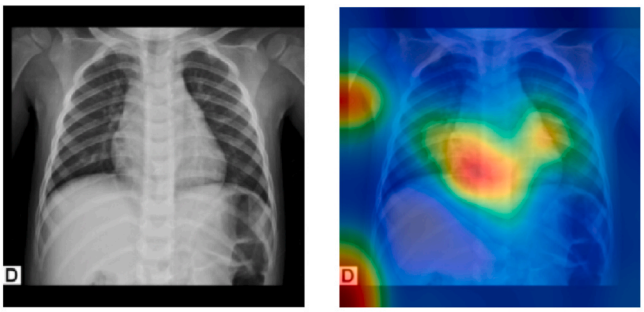
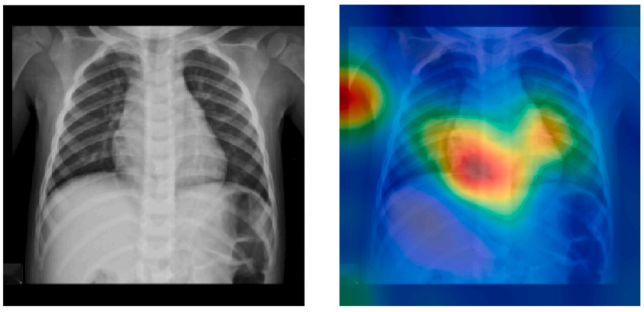


Fig. 3. Steps of the proposed pre-processing task.



(a) Image with a metallic token

(b) Activation map for the image with the token (a)



(c) Resulting image of a token removal process

(d) Activation map for the tokenless image (b)

Fig. 4. Heatmaps with the regions that had the most significant influence on the prediction of the used CNN: After the token removal, the used CNN architecture does not consider the region of the token as a relevant region for the final result.

to the combined probability between the evaluated architectures.

3.2.1. *Pre-processing*

Due to the high dimensionality, different resolutions of the input images, and a metallic token with the original medical examination result, we implemented a pre-processing task to identify the region of interest (ROI) and adapt the input images to the input pattern of CNNs. Fig. 3 depicts the implemented pre-processing task.

The default input for CNNs is square images. In addition, the large number of operations performed makes it challenging to process images with large dimensions. Thus, based on empirical tests, evaluations with medical specialists, and the evaluation results of several dimensions presented in TANG, we defined the input images' size at 256 × 256 pixels. However, resizing images to this dimension distorts the image regions, as there is a big difference between the images' height and width. Thus, we identified the region of interest: the lung area, before resizing the input image, by applying an Otsu threshold [22]. Then, the image's background region is removed, and the image is cropped to contain only the chest region. After this process, the image is transformed into a square image using zero paddings. Finally, the square image is resized to the desired size.

According to Refs. [23,24], the presence of metallic tokens in X-ray images can bias the learning of CNNs. Fig. 4 shows the influence that a metallic token can have on the learning process of CNNs. Hence, it illustrates a VGG-16 activation map for the frontal scan in a scenario with and without a token. In the scenario where the input image has the token (Fig. 4(a)), it is observed that one of the regions where the heat map is more intense (red) is the region pf the token (Fig. 4(b)). This fact indicates that the CNN considered the token essential for decision-making. When the input image for the CNN is the result of a token removal process (Fig. 4(c)), the activation map (Fig. 4(d)) in the region where the token was removed is predominantly blue, which suggests that the region is not considered relevant for the prediction.

As such, the deep learning architecture will learn the token pattern and not critical features that exist in the exam. Therefore, we propose token segmentation through U-Net [25]. We chose a CNN to segment this region because there are no standards for the tokens' size and location in the exams. Thus, training a U-Net with few convolutional filters and less complexity is more effective for different situations.

Table 3
Characteristics of the studied deep learning models.

Model	Topological depth	Number of parameters	Year
VGG-16	23	138,357,544	2014
ResNet50	168	25,636,712	2015
DenseNet121	159	8,062,504	2017

For the U-net training and evaluation, the specialist manually labeled the token of 239 images; then, we used 80% of this data to train the network. We used the remaining 20% of the data to evaluate the segmentation of the token regions. This methodology obtained an accuracy of 99.96%. Then, the token was removed from the image under study using the Fast Marching Method (FMM) proposed by Telea et al. [26], which considers the information of neighboring pixels, starting from the edges of the region of interest. The region's pixels are replaced by the weighted and normalized sum of all pixels in the neighborhood.

3.2.2. Transfer learning

The use of techniques based on deep learning has been used over the years in solutions to the most diverse problems. In the literature, Convolutional Neural Networks have been applied to develop solutions that involve the diagnosis of medical images. These networks have high generalizability, overlaying traditional techniques commonly presented in the literature [10]. Among the main techniques involving CNNs, we can highlight the transfer learning technique, where a CNN is pre-trained in a generic dataset and then used as the basis for changes in the architecture by fine-tuning to a new problem.

In this work, we used the fine-tuning technique to train the proposed architecture. In Vogado et al. [27], the authors used different fine-tuning techniques to develop an architecture that correctly classifies blood slides with or without leukemia. Among the approaches presented, the modified Deeply Fine-Tuning (mDFT) consists of fine-tuning the entire CNN architecture and readjusting the fully connected layers. This fine-tuning technique achieved the best results for the given problem considering challenges such as the dataset size and the presented problem. Therefore, we used mDFT to train the architectures used in the proposed methodology.

3.2.3. Evaluated CNNs

Defining the fine-tuning technique to be used was just one step in developing the proposed methodology. Another fundamental step was the definition of the base CNN for the development of the proposed architecture.

Over the years, several CNN architectures have been proposed to solve computer vision and machine learning problems. These architectures can generalize different datasets and provide accurate results. From the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [28], architectures with distinct characteristics such as depth, number of parameters and convolutional layers, were presented. Among them, we can highlight VGG-16 [29], ResNet50 [30] and DenseNet121 [31]. In works presenting methodologies for the classification of chest X-ray images, these three deep learning models are the most commonly used [10]. In Table 3, the main characteristics of these three architectures are indicated.

We observed that the number of parameters needed to train the architectures was reduced over the years, from over 138 million to 8 million. However, the topological depth increased from 23 to 168, and then 159 layers. In the experiments, we applied mDFT with these three architectures on front and lateral image projections.

3.2.4. Ensemble

An ensemble of classifiers consists of combining different predictions from different classifiers to issue a single answer about the input data [32]. Among the main advantages of using ensembles, we can mention the reduction in overfitting, variance, and the minimization of the instability of the learning algorithms.

The ensemble application can be compared with the assessment carried out by radiologists in identifying whether an exam is normal or not. This common feature is due to one or more projections, mainly, frontal and lateral, on chest X-rays. The combination of the evaluation of each projection helps the physician in the final decision. Therefore, the ensemble employed in this work acts similarly since two CNNs are trained respectively with two projections of one exam. So, it is possible to output a final prediction using predefined rules.

To provide answers with a high probability of being normal or abnormal. We present the use of Confidence Threshold (CTR) to define the predictions into High Confidence Normal (HCn), Borderline classification (BC), or High Confidence Abnormal (HCa), according to the probability given by each CNN. The CTR helps to reduce errors and, consequently, increase the proposed methodology's accuracy. In this way, the approach can reduce the physician's workload and help to improve the quality of medical reports through the issuance of pre-

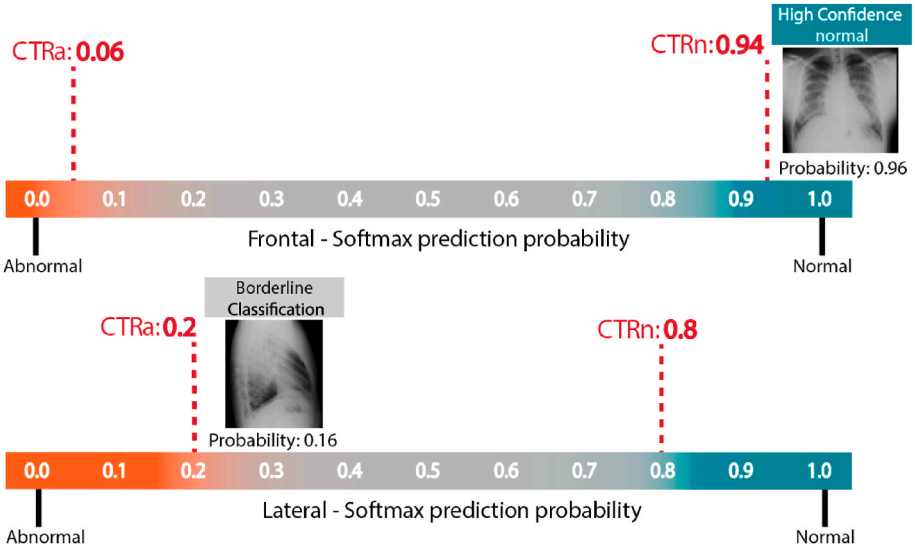


Fig. 5. Example of how the first stage of the ensemble works with confidence factors: The image was classified as normal in the first line, because the prediction probability (0.96) was higher than CTRn (0.94). In the second line, the image was classified as Borderline, because the prediction probability (0.16) was between CTRa (0.2) and CTRn (0.8) values.

Table 4
Best results achieved for frontal projection images using the validation set with different architecture configurations (best values in bold).

Approach	Fc layers	Acc	P	R	S	K
VGG-16	1024	**88.43%**	86.06%	81.40%	92.46%	**0.7472**
VGG-16	1024–512	87.99%	84.06%	**82.67%**	91.04%	0.7397
VGG-16	1024–256	88.29%	87.14%	79.57%	93.28%	0.7423
ResNet50	256	87.89%	**87.29%**	78.10%	**93.50%**	0.7324
DenseNet121	512	87.67%	86.89%	77.85%	93.28%	0.7275

reports. Thus, the physician will only review the response given by the methodology. Therefore, we define an image as one of the three classes according to:

$$R_i = \begin{cases} HC_n & \text{if } P_i > CTRn, \\ HC_a & \text{if } P_i < CTRa, \\ BC & \text{if } P_i > CTRa \text{ and } P_i < CTn, \end{cases} \quad (1)$$

where R is the answer obtained from the ensemble and P_i is the probability of the first neuron obtained by the softmax activation layer for an i image. The confidence factor represented by CTR can assume values according to the class, being n normal and a abnormal.

It is noteworthy that the values for CTRa and CTRn were defined empirically, ranging from 0.5 to 1.0 for the normal class, and from 0.5 to 0 for the abnormal class. Fig. 5 presents an example of how the built ensemble works in the classification of images using random values for CTRa and CTRn for a frontal and a lateral projection.

After classifying all exam images according to the confidence factor values, we carry out the second ensemble stage, which decides the final class. Likewise the assessment by confidence factor, we follow the same pattern of classes presented in the decision by image. However, we implemented rules according to medical knowledge for decision-making:

$$R_e = \begin{cases} HC_a & \text{if at least one } R_i == HC_a, \\ HC_n & \text{if all } R_i == \mathrm{HC_n} \text{ or the number of } \mathrm{HC_n} > \mathrm{BC}, \\ \mathrm{BC} & \text{if all } \mathrm{R_i} == \mathrm{BC} \text{ or the number of BC} \geq \mathrm{HC_n}, \end{cases} \quad (2)$$

where e represents the evaluated exam. In the implemented rules, if there is at least one i image classified as HCa, the entire exam will be HCa. Therefore, to classify the exam in HCn, the quantity of HCn images must be greater than the quantity of BC, or all of them are HCn.

3.3. Evaluation metrics

The metrics generally used to assess diagnostic aid methodologies are based on the confusion matrix. Based on this matrix, it is possible to visualize and evaluate the performance of a prediction algorithm through the verification of predictions. Thus, for binary problems, we can represent the confusion matrix according to the following values: true positive (TP), false positive (FP), false negative (FN) and true negative (TN).

In the problem addressed in this study, the normal class is represented as negative and the abnormal as positive. Thus, TP represents what was correctly classified as abnormal, and FP what is normal but classified as abnormal, TN represents the images correctly classified as abnormal, and FN the abnormal images classified as normal. From these values, we can calculate the valuation metrics. We evaluated the following metrics to select the best models from the performed experiments: accuracy (Acc), precision (P), recall (R), specificity (S), kappa (K), and area under the ROC curve.

Metrics aligned with the results from the classification ensemble were also used. The first is the False Discovery Rate (FDR), and the second is the False Omission Rate (FOR):

$$FDR = \frac{FN}{TN + FN}, \quad (3)$$

$$FOR = \frac{FP}{TP + FP}. \quad (4)$$

These metrics are calculated according to the confusion matrix resulting from the classification ensemble, agreeing with the number of HCn and HCa responses.

In addition to the FDR and FOR metrics, we propose two metrics for selecting the best ensembles. Since the main objective is to increase the percentage of responses with high confidence (HC) and reduce FDR and FOR errors, we propose the Commitment (CM) metric that represents the weighted average of the number of responses and the error obtained in that class, which were defined according to:

$$CM_n = (0.4 * HC_n) + (0.6 * FDR), \quad (5)$$

$$CM_a = (0.4 * HC_a) + (0.6 * FOR), \quad (6)$$

where a represents the abnormal class, and n the normal class. We defined 0.4 as the weight for HC_n and HC_a, and 0.6 for the respective errors by class: FDR and FOR. In this way, the error has a substantial influence on the decision of the best ensemble approach. From selecting the best commitment for each class and with the defined approaches, it is still necessary to make a final decision among the combinations of different ensembles. For this, we used the weighted average of the Commitment values (ACM) for each class, giving greater weight to the normal class, with 0.6 and 0.4 for the abnormal class.

4. Results

For the development of the proposed methodology, we split the experiments into three phases, namely: (1) selection of the best models for frontal projections, (2) selection of the best models for lateral projections, and (3) building of the evaluation ensemble between the models. The used criteria for selecting the models in the first two phases took into account metrics from the literature such as accuracy, precision, recall, specificity, kappa, and AUC. To rank the results, we considered AUC as the primary metric. After choosing the models, the ensemble was evaluated using the confidence factors for the HCa and HCn responses. The selection criteria for the best combinations of the ensemble's models was the best result for the CMa and CMn metrics.

The results presented in this section were obtained without using data augmentation. Despite data augmentation being a widely used

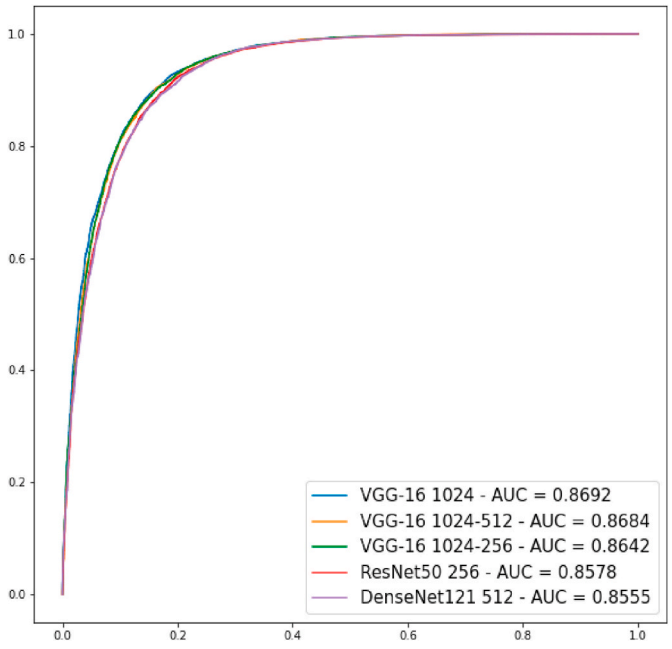


Fig. 6. ROC curve and AUC of the best results obtained for the frontal projection images.

technique to improve CNNs training and avoid overfitting, it was not necessary in this work. We performed tests with and without data augmentation in the initial experiments and found that they performed statistically equal. Thus, we chose not to use data augmentation, reflecting a scenario closer to the one commonly found by physicians. In addition, we verified that, due to the number of available images, the fine-tuned CNNs did not suffer from the overfitting problem.

4.1. Results for frontal projection images

Table 4 presents the best results achieved for the three architectures evaluated in the validation set for the frontal projection images. The architectures are presented according to the metrics obtained and the number of fully connected layers used. In addition, we ranked the results according to the AUC metric. As a tiebreaker, we selected the model with the best performance in the kappa index. Fig. 6 presents the ROC curve and AUC for each architecture indicated in Table 4. To calculate the ROC and AUC curve, we used the probability of each image belonging to the positive class and its actual class. This information was extracted from the CNNs softmax layer. The threshold used to differentiate the positive and negative classes was 0.5.

Also, from Table 4, we can realize that the best validation performance was achieved by VGG-16 with 1024 neurons in the fully connected layer, with 88.43% of accuracy, 0.7472 as to kappa, which is considered very good, and 0.8693 for AUC. For the VGG-16 with two fully connected layers of 1024 and 512, we got a AUC of 0.8685. However, a higher sensitivity was obtained for the best result. Among the architectures with the greatest depth, the ResNet50 256 achieved only 87.89% of accuracy, but had the highest precision and specificity values.

A fact to be observed in the obtained results is that with the increase in the amount of fully connected layers, VGG did not obtain superior results than its configuration with only one layer with 1024 neurons. The same can be observed for ResNet50, whose best result was achieved with only one layer of 256 neurons, and DenseNet121 with 512 neurons.

It is noteworthy that among the studied architectures, VGG-16, even being older and shallower than the others, achieved superior results, noting that it was the architecture that best generalized the images in the validation set for this dataset. This can be justified due to shallower CNNs producing better results for binary problems or with few classes than deeper architectures. This is because deeper architectures are more susceptible to overfitting in binary problems [27].

Table 5 illustrates the results of an ablation study to verify the relevance of metallic token removal in the five best approaches defined in Table 4. From the results, we observed that the use of tokens added a bias in CNN's learning and did not present superior results in any of the

Table 6
Best results achieved for the lateral projection images using the validation set with different architecture configurations (best values in bold).

Approach	Fc layers	Acc	P	R	S	K
ResNet50	1024	83.62%	66.97%	**76.25%**	86.3%	0.5991
VGG-16	1024–512	**85.0%**	**71.96%**	71.83%	89.8%	**0.6167**
ResNet50	512	84.1%	68.88%	73.79%	87.86%	0.6028
ResNet50	256	84.95%	71.77%	71.9%	89.7%	0.6157
VGG-16	1024–256	84.72%	71.75%	70.55%	**89.88%**	0.6075

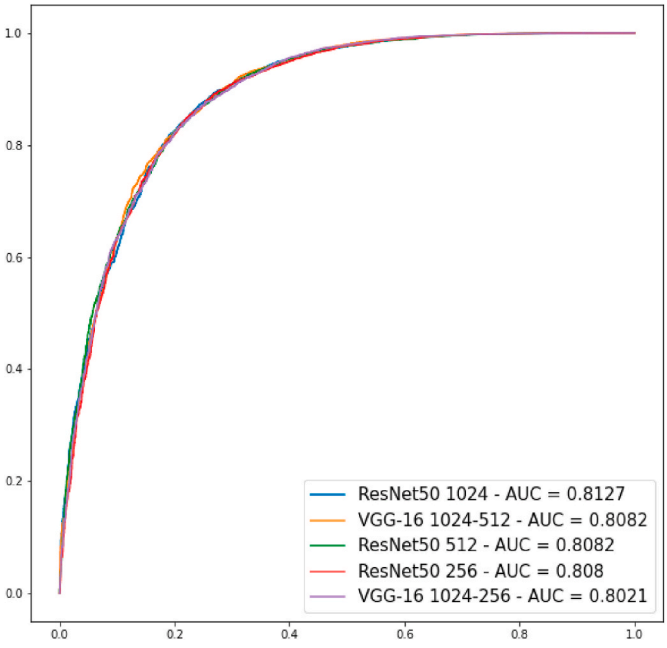


Fig. 7. ROC curve and AUC of the best results obtained for lateral projection.

Table 5
Comparison between the best results with and without token removal in frontal projections images.

		With Token			Without Token		
Approach	Fc layers	Acc	K	AUC	Acc	K	AUC
VGG-16	1024	87.67%	0.7212	0.8445	**88.43%**	**0.7472**	**0.8692**
VGG-16	1024–512	87.84%	0.7171	0.8544	**87.99%**	**0.7397**	**0.8684**
VGG-16	1024–256	87.46%	0.7296	0.8433	**88.29%**	**0.7423**	**0.8642**
ResNet50	256	87.52%	0.7243	0.8543	**87.89%**	**0.7324**	**0.8578**
DenseNet121	512	87.07%	0.7158	0.8518	**87.67%**	**0.7275**	**0.8555**

Table 7
Comparison between the best results with and without token removal in lateral projections images.

		With Token			Without Token		
Approach	Fc layers	Acc	K	AUC	Acc	K	AUC
ResNet50	1024	**84.1%**	0.5981	0.8025	83.62%	**0.5991**	**0.8127**
VGG-16	1024–512	84.89%	0.6083	0.7999	**85.0%**	**0.6167**	**0.8082**
ResNet50	512	82.29%	0.5725	0.8032	**84.1%**	**0.6028**	**0.8082**
ResNet50	256	83.41%	0.5922	0.8079	**84.95%**	**0.6157**	**0.8080**
VGG-16	1024–256	79.95%	0.5345	0.7944	**84.72%**	**0.6075**	**0.8021**

Table 8
Ensemble results ranked considering different architectures for frontal and lateral projection images in the validation set (best values in bold).

CTRn - CTRa												
Frontal CNN	Lateral CNN	Frontal	Lateral	BC	Responses	HCn	FDR	HCa	FOR	CMn	CMa	ACM
VGG-16 1024 VGG-16 1024-512	VGG-16 1024 -512	0.94–0.06	0.8–0.2	8959 (47.07%)	10074 (52.93%)	29.57% (5629)	2.58% (145)	23% (4445)	4.5% (200)	70.28	**67.36**	**69.112**
VGG-1024	ResNet50 1024 VGG-16 1024-512	0.95–0.16	0.81–0.13	8222 (43.2%)	10811 (56.8%)	29.91% (5692)	2.78% (158)	27% (5119)	5.92% (303)	**70.29**	67.2	69.054
VGG-16 1024 VGG-16 1024-512	ResNet50 1024	0.93–0.15	0.81–0.04	8659 (45.49%)	10374 (54.51%)	29.62% (5637)	2.79% (157)	25% (4737)	4.33% (205)	70.17	67.35	69.042
VGG-16 1024	VGG-16 1024-512	0.95–0.16	0.84–0.07	8385 (44.06%)	10648 (55.94%)	29.82% (5675)	2.91% (165)	26% (4973)	5.43% (270)	70.18	67.19	68.984
VGG-16 1024	ResNet50 1024	0.94–0.14	0.83–0.04	8510 (44.71%)	10523 (55.29%)	29.81% (5673)	3.0% (170)	25% (4850)	5.03% (244)	70.12	67.17	68.94

Table 9
Ensemble results considering different architectures for frontal and lateral images in the test dataset.

CTRn - CTRa												
Frontal CNN	Lateral CNN	Frontal	Lateral	BC	Number of answers	HCn	FDR	HCa	FOR	CMn	CMa	ACM
VGG-16 1024 VGG-16 1024-512	VGG-16 1024 -512	0.94–0.06	0.8–0.2	8524 (45%)	10263 (55%)	32% (6010)	1.68% (101)	23% (4253)	4.91% (209)	71.78	66.10	68.97

illustrated metrics [24]. Among the evaluated metrics, only the accuracy presented similar results with a significance level of 5%

4.2. Results for lateral projection images

Table 6 presents the results achieved by the models under comparison for the validation set in the used dataset with lateral projection images. Fig. 7 presents the ROC curve and AUC for each architecture indicated in Table 6.

From the best achieved results, one can highlight ResNet50 with 1024 neurons, which obtained 0.8127 of AUC and a recall of 76.25%. However, even with the best AUC, this CNN configuration had the lowest accuracy, precision, specificity and kappa. In contrast, the second-best result obtained by VGG-16 with 1024-512 achieved the best results in three metrics, and the second-best AUC tied with ResNet50 with 512 neurons. However, VGG-16 is considered superior according to the tiebreaker defined by the value obtained as to kappa.

Table 7 illustrates the results of an ablation study to verify the relevance of metallic token removal in the five best approaches defined in Table 6. As discussed in the frontal view, the results obtained with the token presence are inferior to those obtained without the presence. For the lateral view, only in the ResNet50 configuration with 1024 neurons, the token approach obtained a higher accuracy with 84.1% against 83.62%. However, looking at the value of K and especially the AUC, we verify that the approach with removal was superior.

4.3. Exam screening ensemble

From the results obtained in the validation set and the selection of the five best models of CNNs, we evaluated the ensemble classifier for the exams screening. Due to the number of existing combinations between all models, we reduced the scope to only the two best results with frontal projection images, and the two best with lateral projection images. Hence, from the selected models, we studied eight combinations. Table 8 indicates the ensemble results for the validation set along with the five best achieved matches.

Among the objectives outlined for building the ensemble classifier, one of the main ones is to provide high confidence, that is, a high probability of actually belonging to a specific class. Therefore, we increase class precision and decrease error. On the other hand, there is a reduction in the number of responses issued, now released as BC. Furthermore, to develop and select the best combinations, it was necessary to impose rules to limit the amount of normal and abnormal responses. We observed through empirical experiments that with HCa and HCn rates above 30%, the values for FOR and FDR would be above 5% in most cases. To limit the error to be obtained, we established HCa and HCn between 20 and 30%, always looking for the best FOR and FDR value between these two response ranges.

During the tests to build the ensemble classifier, we varied the CTRn and CTRa confidence factors for the frontal and lateral models and used the CM as a parameter to balance the number of responses issued by class and the error obtained. Thus, we observed that we obtained the best confidence factors from the best CM values.

Among the obtained results, we verified that in the five combinations

Table 10
Comparison of the proposed methodology with state-of-the-art ones on the binary classification of chest X-ray exams.

Work	Year	Images	Acc	AUC
Yates et al. [15]	2019	53,149	94.6%	–
Dunnmon et al. [17]	2019	216,431	91.0%	–
Ellis et al. [16]	2020	7000	82.0%	–
Wong et al. [18]	2020	128,886	–	0.8210
Tang et al. [19]	2020	141,617	94.64%	0.9824
Proposed Methodology	2022	352,460	87.54%	0.8721

Table 11
Comparison between the proposed methodology and the one suggested by Dyer et al. [20].

Work	Year	Images	HCn	FDR
Dyer et al. [20]	2021	3887	15%	2.3%
Proposed Methodology	2022	352,460	32%	1.68%

presented, the number of responses was higher than 50%. For the frontal CNNs, the CTRn values were higher than 0.9 of probability, while for the lateral CNNs, none of the CTRn exceeded the 0.85 range. We believe that the confidence factors of the lateral CNNs were lower because some of the findings or alterations are not visible on lateral images. However, it does not decrease system performance because, according to the medical protocol, only the frontal view is necessary for some instances, and the examination with the lateral view is not required. Thus, for cases of exams with only frontal projection, the probability emitted by the frontal CNN is the only one taken into account in the classification into HCa, HCN, or BC.

When analyzing the HCn values, we observed that they were closer to the 30% defined as the limit, while the FDR values were, until then, equivalent to or less than 3%. This same behavior was not observed in the abnormal class, which presented HCa values far from 30% and STR rates higher than 5% in some cases. This fact demonstrates that models' ability to classify normal exams and that the unbalance between the classes, even mitigated, interfered in the final result.

Initially, for the evaluation of the architectures, we considered the combination of two models, a frontal and a lateral model. However, we calculated the average between the probabilities obtained by two models of the same projection to obtain better results. To rank the results, we calculated the Average Commitment Metric (ACM) according to the values of CMn and CMa. With the ensemble between the front and lateral models, we found the best result using the frontal VGG-16 1024 and VGG-16 1024-512 models, and the lateral VGG-16 1024-512 model. Furthermore, we observed that, among the five best results, the three that obtained the highest ACM values had a combination of models.

We calculated the results for the test set from the best combination of models and definition of CTR values. In Table 9, one can observe that some metrics were superior to the validation set, demonstrating that the approach managed to generalize the results to a set never seen by the models.

For the test dataset, the results for the normal class were higher than those obtained for the validation set, with HCn equal to 32% and FDR of 1.68%. In abnormal class, the results obtained were close to those in the validation set, with 23% HCa and FOR of 4.91%. It is noteworthy that the CTR values were defined using the validation set, so these are replicated to the test set without limitations regarding the amount of HCa or HCn responses. The proposed methodology presented a number of answers superior to 50% of the total cases as in the validation dataset.

We performed a Z-test [33] to statistically compare and evaluate the results among the proposed ensemble (Table 8) and the individual CNNs, Frontal VGG-16 1024, Frontal VGG-16 1024 512 and Lateral VGG-16 1024-512 (Tables 4 and 6), with a significance level of 5% to assess if the results were significantly different from each other. The results shown that the ensemble approach achieved markedly higher

Table 12
Comparison among the results obtained by the proposed methodology and the ones obtained by state-of-the-art methodologies in public image datasets (best values in bold).

Work	Year	Acc	P	R	S	AUC
Indiana + NIH Datasets						
Yates et al.	2018	94.60%	**99.80%**	**94.60%**	93.4%	**0.98**
Proposed Methodology	2021	**98.85%**	62.00%	85.00%	**98.62%**	0.92
NIH-RSNA Dataset						
Tang et al.	2020	**92.34%**	88.09%	**97.40%**	87.55%	**0.9871**
Proposed Methodology	2022	89.63%	**92.83%**	86.23%	**93.21%**	0.8972

Table 13
Comparison among the results obtained by the proposed methodology and the ones obtained by state-of-the-art methodology in our image dataset (best values in bold).

Work	Year	Methodology	Accuracy	AUC
Yates et al. [15]	2018	Fine tuning with InceptionV3	86.89%	0.8364
Dunnmon et al. [17]	2019	Fine tuning with DenseNet121	86.95%	0.8456
Tang et al. [19]	2020	Fine tuning with ResNet18	87.19%	0.8450
Proposed Methodology	2022	mDFT VGG-16 1024	**88.43%**	**0.8693**

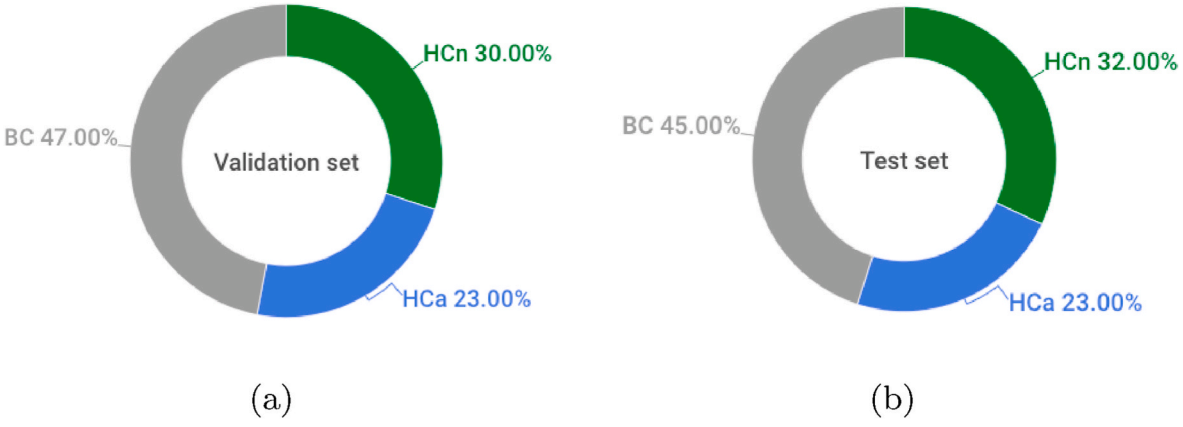


Fig. 8. Percentage of responses with high confidence for the validation (a) and test (b) sets obtained by the proposed methodology.

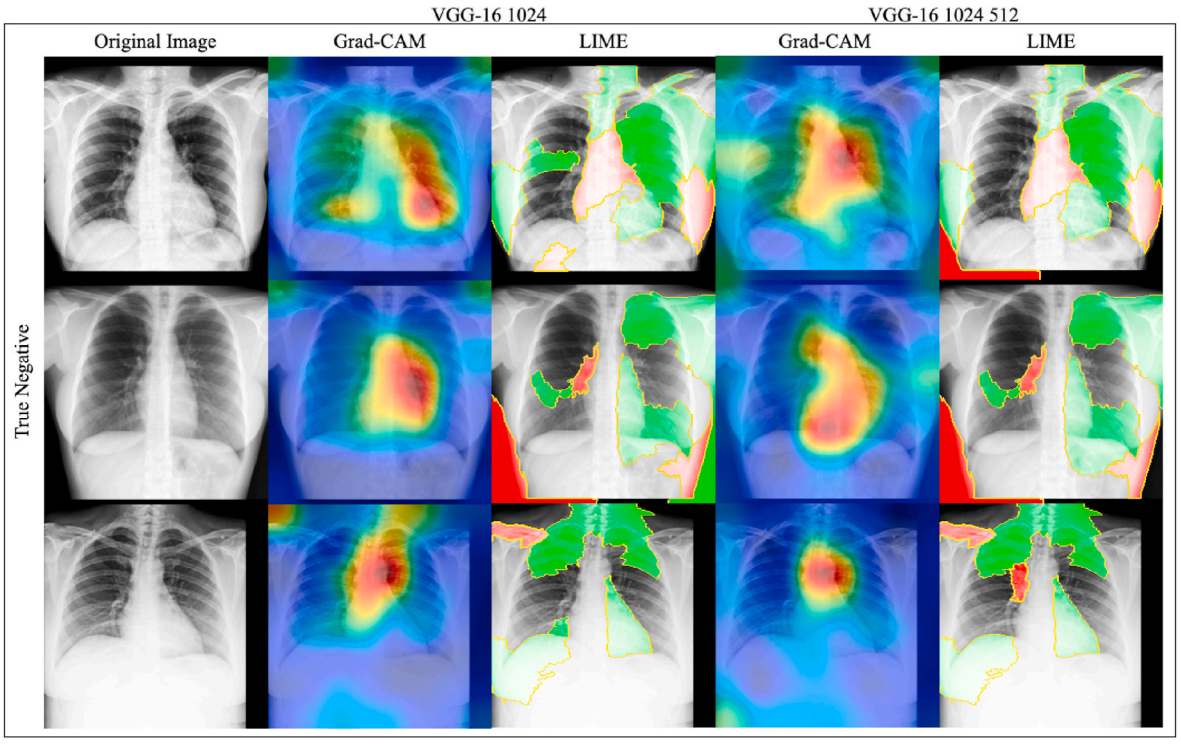


Fig. 9. Samples with visual interpretations for images correctly classified as normal.

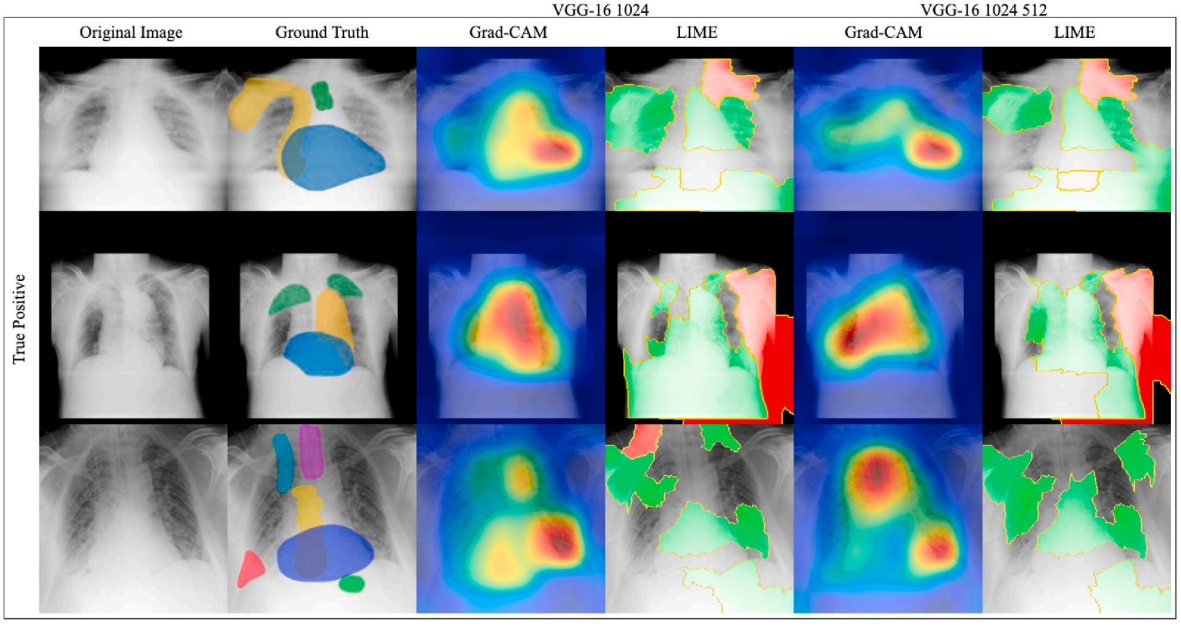


Fig. 10. Samples with visual interpretations for images correctly classified as abnormal.

performance than the individual CNNs.

5. Discussion

From Table 10, it is possible to compare the results achieved by the proposed and state-of-the-art methodologies on the binary classification of exams into normal and abnormal. On the other hand, from Table 11, it is possible to compare the HCn results obtained ny the proposed methodology with the ones of Dyer et al. [20], the only work found in the literature that, like the proposed methodology, proposes to issue an initial diagnosis in exams with high confidence.

In Table 10, it is indicated the results obtained for the test set with the proposed methodology and CTR values equal to 0.5, which is the default value found in the literature as to binary problems. This methodology obtained 87.54% of accuracy and 0.8721 as to ROC, which highlights the generalizability of the proposed models for the test set. Furthermore, when compared with methodology in the literature, we observed that the results achieved by the proposed methodology are in the range of the results presented in the state-of-the-art.

A fact noted during the results analysis was the number of images

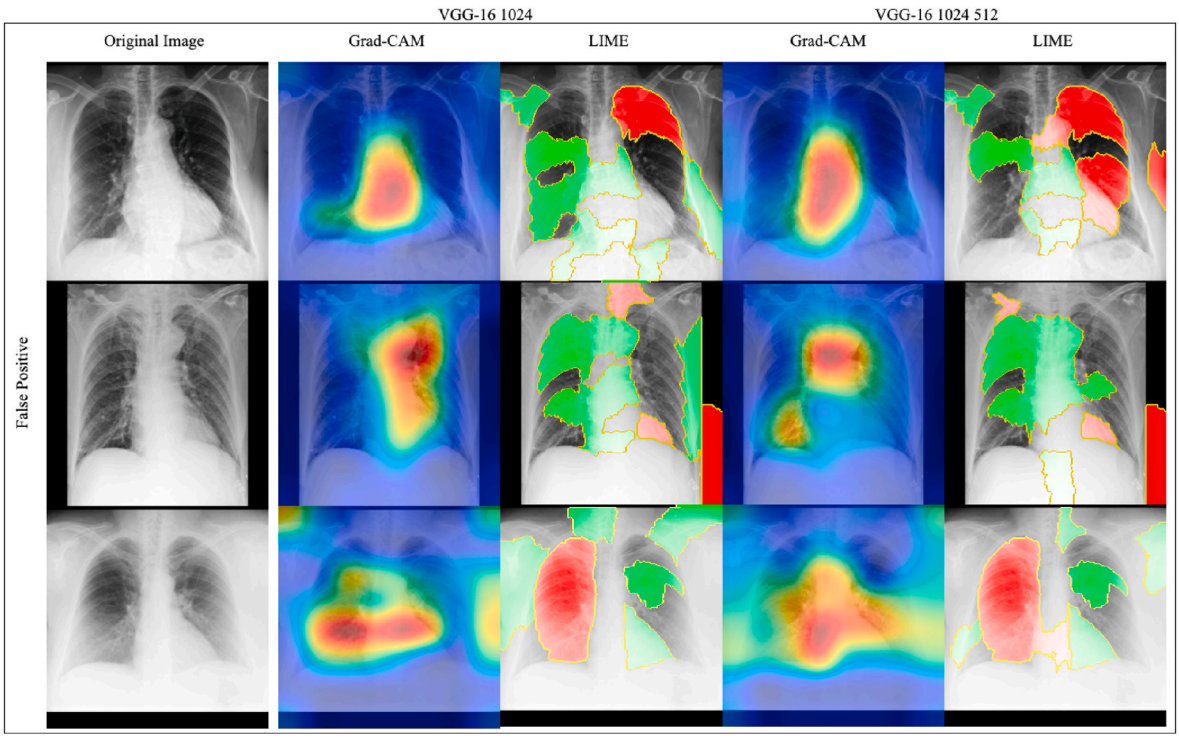


Fig. 11. Samples with visual interpretations for images incorrectly classified as abnormal.

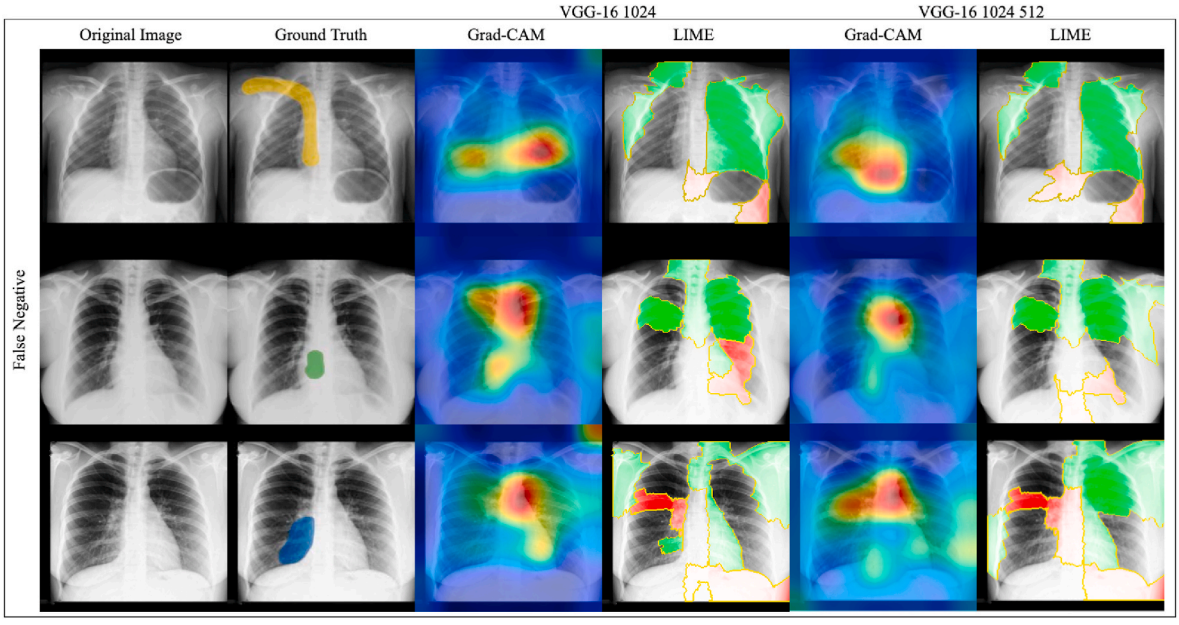


Fig. 12. Samples with visual interpretations for images incorrectly classified as normal.

Table 14
FDR and FOR values obtained with different assessment sets.

Assessment set	FDR	FOR
Validation	2.58%	4.50%
Test	1.68%	4.91%
Supervision	1.40%	7.02%

Table 15
Top five pathologies identified in the exams incorrectly classified as normal in the supervision assessment.

Pathology	Number of exams	(% of total errors)
Nodules	23	13.21%
Granuloma	21	12.06%
Cardiomegaly	19	10.91%
Opacities	16	9.19%
Consolidation	9	5.17%

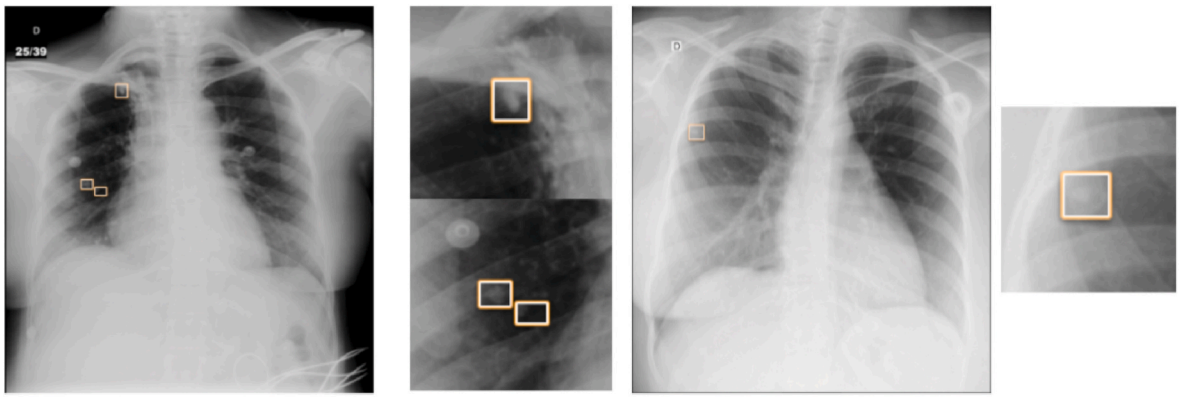


Fig. 13. Sample of chest X-rays images with the presence of pulmonary nodules.

and the different datasets used in the found state-of-the-art works. While the proposed approach used more than 300,000 images to develop the methodology, some authors presented methodologies with only frontal images. Only the work of Ellis et al. [16] made use of the two projections in their experiments. However, the authors evaluated the proposed methodology using a small dataset with only 7000 images.

Taking into account the answers given by the proposed ensemble based methodology, we compared its results with the ones achieved by the methodology of Dyer et al. [20], Table 11. Analyzing the percentage of responses with HCn, the methodology suggested by Dyer et al. [20] reached 15%, while the proposed methodology obtained 32%. The error for the normal class of the proposed methodology was lower than that the one presented by Dyer et al. [20].

Fig. 8 illustrates the percentage of HCn and HCa responses for the validation and test sets. We observed that the proposed methodology classified more than 50% of the exams with a precision greater than 95% for both classes.

In addition to the previous indirect comparison, Tables 12 and 13 present direct comparisons with state-of-the-art methodologies.

Formerly, we investigated the performance of the proposed methodology on the public image sets used by Yates et al. [15], and Tang et al. [19]. The first set of images was composed of the Indiana Dataset and NIH databases, while the second one is a simplified version of the NIH-RSNA dataset. In this situation, we did not select the best parameters. We simply executed the proposed methodology with the parameters previously defined based on the training of our dataset. The obtained results are presented in Table 12, where it is possible to observe that the proposed methodology achieved competitive results, being better in at least two metrics.

We also evaluated the performance of state-of-the-art methodologies on our image dataset. Among the authors who provided sufficient details to reproduce or make the source code publicly available, we found: Yates et al. [15], Dunnmon et al. [17] and Tang et al. [19]. For this evaluation, the implementations were carried out considering all the information presented in the respective works and only the frontal projection was used since these works were developed to be applied only in this projection. In addition, the training and test sets were the same used in the definition of the proposed methodology.

Table 13 presents the results of this evaluation. The methodology by Tang et al. [19] achieved an accuracy of 87.19%, being considered the best among the three works found in the literature. However, the performance was lower than that the one obtained by the proposed methodology, which obtained an accuracy of 88.43%. In fact, the state-of-the-art methodologies performed worse than the five best architectures evaluated in this study using frontal projection (Table 4).

5.1. *Visual interpretation*

CNNs currently provide excellent learning and generalization capabilities. However, due to their complexity, they do not present transparency of what was learned. Therefore, a crucial aspect of understanding the most relevant features used in the prediction is the visual data interpretation.

Visual interpretation can be classified into two categories: interpretation of an instance and general network interpretation. The first category is divided into gradient-based and perturbation-based methods. Gradient-based ones such as Class Activation Mapping (CAM) and Grad-CAM [34] use the latest convolutional layer to provide a visual interpretation at the pixel level and have class discrimination capability. On the other hand, pertubation-based methods consider an element as essential for decision making if its removal changes the output considerably. The importance of this disturbing element can be estimated by comparing the network output with and without the element. In images, for example, it is intuitive that changes in the pixels that most contribute to a result lead to a different prediction. The Local Interpretable Model-agnostic Explanations (LIME) [35] method is one of the perturbation-based methods and represents the discriminative importance of the class using superpixels.

To help interpret the results obtained by the proposed methodology, we implemented and evaluated examples correctly and incorrectly classified as normal and abnormal. In addition, we also present the ground-truth with color marking defined by a radiologist for different findings and pathologies. We used Grad-CAM and LIME to provide a visual interpretation of the exams. The Grad-CAM was represented employing the heat map, with the most intense region being the one that contributed the most to the prediction. The LIME represented the top 10 superpixels that positively (represented in green) and negatively (represented in red) contributed to the prediction. In addition, we used the two best frontal architectures obtained during the development of the methodology and which are part of the proposed ensemble classifier.

The images were randomly selected for this analysis, and examples of True Negatives and True Positives images are shown in Fig. 9 and Fig. 10, respectively.

In the exams of Fig. 9, one can observed that, in both visualizations, the models considered the pulmonary and cardiac regions as the critical areas. In addition, in the third example, it can be noted that the region of interest was the aorta, where most of the cases of alterations were concentrated. We concluded that this is a region of great relevance for classification as HCa.

In the examples in Fig. 10, one can observe three cases with very evident alterations when compared to TN examples. The first exam contains a sternorrhaphy and cardiac pacemaker and an enlargement in the heart area. Through interpretation, we observed that the Grad-CAM highlighted the heart regions, denoting that cardiomegaly was the

primary influence on the prediction as abnormal. LIME also highlighted the area of the pacemaker. In the second example, the highlighted points of interest escaped the abnormality, concentrating on regions close to the heart that were configured as a thickening and the altered aorta. The third example shows opacities in both lungs and reduced transparency of the right lung base. In addition, tubes, center access and clips are visible. Both methods highlighted the lungs regions as the most relevant for classification. Furthermore, additional findings were considered in Grad-CAM of VGG-16 1024 512.

It is noteworthy that in all examples, Grad-CAM generated different heatmaps. This denotes the ensembles' ability to use distinct image features to obtain the exam's final prediction, corroborating the proposed methodology's robustness.

Figs. 11 and 12 illustrate exams incorrectly classified as abnormal or normal, that is, False Positive and False Negative results. Among the two types of errors, we highlight a greater severity in FNs, as they can present risks to the patient. It is noteworthy that, in these cases, the exams tend to resemble normal cases visually. Still, they were configured as abnormal through the identification of barely perceptible changes in the images.

For FN cases, we observed visually similar results to those shown in Fig. 9 with the lung region free and changes in cardiac volume. Furthermore, when comparing the areas of interest demarcated by Grad-CAM and LIME with the Ground Truth marked by the radiologist, we observed that the CNNs considered the regions of the findings as relevant for the prediction. To predict the exams as normal, the considered region of interest was the cardiac one. In the first examination presented, the finding visualized was the central venous access highlighted in yellow. In the second exam, the doctor observed the arthrosis in the spine. However, the change can be considered invisible due to the reduction in the exam quality due to the resizing. In the third example, infiltration in the right lung was observed using the LIME with the VGG-16 1024 architecture. However, the network considered that this finding contributed positively to the normal classification.

In the examples illustrated in Fig. 12, the interpretations provided by Grad-CAM and LIME are focused on regions that present risks to the patient, such as the heart and lungs. Factors contributing to the appearance of these errors are the quality of examination acquisition, the radiologist's interpretation of what should or should not be considered a severe alteration, or even the need for additional tests to confirm the alterations.

5.2. Validation with supervising physicians

We implemented the proposed methodology on a Picture Archiving and Communication System (PACs) server to validate and certify the quality of the responses with radiologists. Two image sets were available, with 12,804 exams classified as HC_n and 7183 classified as HC_a. The agreement achieved for the class HC_n was 98.60% (12,630) and as to FDR was 1.4% (174). While for the class HC_a, we obtained 92.98% (6679) of agreement and a FOR of 7.02% (502). Table 14 presents the FDR and FOR values obtained in the three evaluation scenarios: validation, testing, and supervision.

For normal exams, the FDR value in the validation set did not match the test and supervision sets' results. The lowest FDR value was obtained in the supervision. The FOR obtained in the validation and test sets were similar, while in the supervision, a higher value was obtained, indicating more false positives.

For abnormal exams (FOR metric), we observed that improvements in the proposed methodology are still required to increase the agreement with experts. However, we can emphasize that the physicians' interpretation of what should or should not be considered a pathology is a determining factor in the accuracy of a computational methodology. Other factors can contribute to the difference in the results obtained with the same CTRa and CTRn values, such as the existing diversity in the dataset with different pathologies.

Among the 172 exams classified as abnormal by the specialist, a large part contains granulomas (12.06%), which appear due to the healing process of previous diseases and generally do not present any risk to the patient. Table 15 presents the top five pathologies or findings among the false negatives classified by the specialist in the supervision assessment.

Among the pathologies and findings presented in Table 15, nodules and granulomas are challenging to identify by computational methodologies based on image classification due to usual performed image resizing. In addition, opacities and consolidations, representing 9.19 and 5.17% of the total cases, respectively, in situations where they are more discrete in the lung base, represent a challenge for the proposed methodology. Besides, Cardiomegaly can vary according to the physician's interpretation, but it still is a critical error.

5.3. Strengths and limitations of the proposed methodology

Based on the presented results, we can observe that the proposed methodology developed based on the definition of confidence factors is relevant in this scientific context and applicable in a practical scenario. Compared with methodologies found in the state-of-the-art, we proposed a more robust pre-processing of the images used in training and not just a common resizing of the input images. Another interesting feature is the use of the lateral projection in the ensemble, which is not commonly taken into account in the literature. It is worth noting that this incidence, when present in the examination, is crucial for medical analysis since several abnormalities cannot be adequately observed with only the frontal projection.

One of the limitations of the proposed methodology is the classification of the input images into only two classes (binary). Despite this, it has benefits for physicians and hospitals. To the physicians because the use of confidence factors. Thus, it is able to issue more accurate answers, helping them to reduce the response time for normal exams and giving a prior opinion on the presence of pathologies. Besides, it is possible to use precise answers in queue control for the screening of exams, optimizing the service flow. In addition, several hospitals are isolated from big centres and do not have radiologists available to evaluate/report exams, so, with the proposed methodology, it is possible to obtain results with a high success rate that help inexperienced physicians in their diagnosis. Additionally, one can note that the screening of healthy or pathological exams is crucial for a later stage where it will be possible to detect specific pathologies.

We emphasize that pulmonary nodules (solid and round lesions) are one of the main challenges faced in developing computational methodologies to diagnose chest X-rays. Following the flow of the proposed methodology, during image resizing, the characteristics of this finding are usually missed, especially when they are close to the lung base. In some exams, the nodule is the only existing alteration, and the rest of the exam is entirely healthy. Cases like this increase the number of false negatives and require more specific methodologies for their detection. Fig. 13 shows examples of chest X-rays with the presence of pulmonary nodules.

6. Conclusion

This work presented a methodology based on an ensemble of classifiers and CNNs for classifying chest X-ray projections into normal or abnormal. In the development of the proposed methodology, we evaluated different architectures of CNNs and hyper-parameters on a multi-example and heterogeneous image dataset. In addition, we carry out assessments that tend to reduce the error per class and, consequently, achieve more accurate results.

Comparing the proposed methodology with the ones found in the state-of-the-art, our methodology classified more than twice the number of exams with a lower error rate, thus indicating better reliability in its predictions. We believe that using a heterogeneous image dataset in the training stage gave the proposed methodology the capacity to extract

features considered essential for the correct prediction, which leads to a high generalization power. Through answers based on high confidence, the methodology can reduce the medical effort to classify normal exams or organize an optimized and prioritized workflow.

The proposed methodology is susceptible to errors, mainly in the presence of nodules and discrete opacities, due to the variety of findings and their distribution in the dataset. In these cases, it is necessary to implement a further step for the specific classification of some changes. As a next step for developing a complete methodology, we will evaluate the detection of different pathologies or findings resulting from a normal responses. We intend to provide a hierarchical methodology less susceptible to errors, and perform new experiments to increase the tradeoff between the percentage of responses and accuracy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Yanas, E. Triantaphyllou, A systematic survey of computer-aided diagnosis in medicine: past and present developments, Expert Syst. Appl. 138 (2019), 112821.

[2] S. Itani, F. Lecron, P. Fortemps, Specifics of medical data mining for diagnosis aid: a survey, Expert Syst. Appl. 118 (2019) 300–314, https://doi.org/10.1016/j.eswa.2018.09.056.

[3] F. Gao, H. Yoon, T. Wu, X. Chu, A feature transfer enabled multi-task deep learning model on medical imaging, Expert Syst. Appl. 143 (2020), 112957, https://doi.org/10.1016/j.eswa.2019.112957.

[4] S. Zhang, K. Amahong, X. Sun, X. Lian, J. Liu, H. Sun, Y. Lou, F. Zhu, Y. Qiu, The mirna: a small but powerful rna for covid-19, Briefings Bioinf. 22 (2021) 1137–1149, https://doi.org/10.1093/bib/bbab062.

[5] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Y. Chen, J. Su, G. Lang, Y. Li, H. Zhao, K. Xu, L. Ruan, W. Wu, Deep Learning System to Screen Coronavirus Disease 2019 Pneumonia, 2020, 09334 arXiv:2002.

[6] A.M. Ismael, A. Şengür, The investigation of multiresolution approaches for chest x-ray image based covid-19 detection, Health Inf. Sci. Syst. 8 (2020) 1–11, https://doi.org/10.1007/s13755-020-00116-6.

[7] J.C. Gomes, V.A. de F. Barbosa, M.A. Santana, J. Bandeira, M.J.S. Valença, R.E. de Souza, A.M. Ismael, W.P. dos Santos, Ikonos: an intelligent tool to support diagnosis of covid-19 by texture analysis of x-ray images, Res. Biomed. Eng. (2020) 1–14, https://doi.org/10.1007/s42600-020-00091-7.

[8] A.M. Ismael, A. Şengür, Deep learning approaches for covid-19 detection based on chest x-ray images, Expert Syst. Appl. 164 (2021), 114054, https://doi.org/10.1016/j.eswa.2020.114054.

[9] C. Qin, D. Yao, Y. Shi, Z. Song, Computer-aided detection in chest radiography based on artificial intelligence: a survey, Biomed. Eng. Online 17 (113) (2018) 1–23, https://doi.org/10.1186/s12938-018-0544-y.

[10] E. Çallı, E. Sogancioglu, B. van Ginneken, K.G. van Leeuwen, K. Murphy, Deep Learning for Chest X-Ray Analysis: A Survey, Medical Image Analysis, 2021, 102125, https://doi.org/10.1016/j.media.2021.102125.

[11] H. Behzadi-khormouji, H. Rostami, S. Salehi, T. Derakhshande-Rishehri, M. Masoumi, S. Salemi, A. Keshavarz, A. Gholamrezanezhad, M. Assadi, A. Batouli, Deep learning, reusable and problem-based architectures for detection of consolidation on chest x-ray images, Comput. Methods Progr. Biomed. 185 (2020), 105162, https://doi.org/10.1016/j.cmpb.2019.105162.

[12] Q. Guan, Y. Huang, Multi-label chest x-ray image classification via category-wise residual attention learning, Pattern Recogn. Lett. 130 (2020) 259–266, https://doi.org/10.1016/j.patrec.2018.10.027, image/Video Understanding and Analysis (IUVA).

[13] B. Chen, J. Li, X. Guo, G. Lu, Dualchexnet: dual asymmetric feature learning for thoracic disease classification in chest x-rays, Biomed. Signal Process Control 53 (2019), 101554, https://doi.org/10.1016/j.bspc.2019.04.031.

[14] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, https://doi.org/10.1109/cvpr.2017.369.

[15] E. Yates, L. Yates, H. Harvey, Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification, Clin. Radiol. 73 (9) (2018) 827–831.

[16] R. Ellis, E. Ellestad, B. Elicker, M.D. Hope, D. Tosun, Impact of hybrid supervision approaches on the performance of artificial intelligence for the classification of chest radiographs, Comput. Biol. Med. 120 (2020), 103699, https://doi.org/10.1016/j.compbiomed.2020.103699.

[17] J.A. Dunnmon, D. Yi, C.P. Langlotz, C. Ré, D.L. Rubin, M.P. Lungren, Assessment of convolutional neural networks for automated classification of chest radiographs, Radiology 290 (2) (2019) 537–544, https://doi.org/10.1148/radiol.2018181422.

[18] K.C.L. Wong, M. Moradi, J. Wu, A. Pillai, A. Sharma, Y. Gur, H. Ahmad, M. S. Chowdary, J. Chiranjeevi, K.K. Reddy Polaka, V. Wunnava, D. Reddy, T. Syeda-Mahmood, A robust network architecture to detect normal chest x-ray radiographs, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1851–1855, https://doi.org/10.1109/ISBI45749.2020.9098671.

[19] Y.-X. Tang, Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B.A. Redd, C.J. Brandon, Z. Lu, M. Han, J. Xiao, R.M. Summers, Automated abnormality classification of chest radiographs using deep convolutional neural networks, Digit. Med. 3 (70) (2020) 1–8.

[20] T. Dyer, L. Dillard, M. Harrison, T.N. Morgan, R. Tappouni, Q. Malik, S. Rasalingham, Diagnosis of normal chest radiographs using an autonomous deep-learning algorithm, Clin. Radiol. 76 (2021) 473.e9–473.e15, https://doi.org/10.1016/j.crad.2021.01.015.

[21] G. Chassagnon, M. Vakalopoulou, N. Paragios, M.-P. Revel, Artificial intelligence applications for thoracic imaging, Eur. J. Radiol. 123 (2020), 108774, https://doi.org/10.1016/j.ejrad.2019.108774.

[22] N. Otsu, A threshold selection method from gray-level histograms, in: IEEE Transactions on Systems, Man and Cybernetics 9, 1979, pp. 62–66, https://doi.org/10.1109/TSMC.1979.4310076, 1.

[23] J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano, E.K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study, PLoS Med. 15 (11) (2018) 1–17, https://doi.org/10.1371/journal.pmed.1002683.

[24] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut Learning in Deep Neural Networks, 2020, 7780 arXiv:2004.

[25] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), vol. 9351, Springer, 2015, pp. 234–241, of LNCS.

[26] A. Telea, An image inpainting technique based on the fast marching method, J. Graph. Tool. 9 (1) (2004) 23–34, https://doi.org/10.1080/10867651.2004.10487596.

[27] L. Vogado, R. Veras, K. Aires, F. Araújo, R. Silva, M. Ponti, J.M.R.S. Tavares, Diagnosis of leukaemia in blood slides based on a fine-tuned and highly generalisable deep learning model, Sensors 21 (9) (2021), https://doi.org/10.3390/s21092989.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[29] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-scale image Recognition, CoRR abs/1409.1556, http://arxiv.org/abs/1409.1556.

[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2016, pp. 770–778.

[31] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269.

[32] B.V. Dasarathy, B.V. Sheela, A composite classifier system design: concepts and methodology, Proc. IEEE 67 (5) (1979) 708–713.

[33] R.G. Congalton, K. Green, in: Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, second ed., CRC Press, Boca Raton, 2008.

[34] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626, https://doi.org/10.1109/ICCV.2017.74.

[35] M.T. Ribeiro, S. Singh, C. Guestrin, why should I trust you?": explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1135–1144.