NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATION

PROGRAM OF POSTGRADUATE STUDIES
"BIOINFORMATICS-BIOMEDICAL DATA SCIENCE"

MASTER PROGRAM'S THESIS

Explainable Artificial Intelligence for Deep Learning Methods in Chest X-Ray Classification

Theodora Chrysoula

Supervisors: **Dr. Theodore Dalamagas**
Research Director, Information Management Systems Institute,
ATHENA Research Center

Dr. Christos Diou
Assistant Professor, Department of Informatics and Telematics,
Harokopio University of Athens

ATHENS

FEBRUARY 2023

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

"ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ-ΒΙΟΕΠΙΣΤΗΜΗ ΙΑΤΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕΘΟΔΟΙ ΕΡΜΗΝΕΥΣΙΜΟΤΗΤΑΣ ΣΕ ΜΟΝΤΕΛΑ ΒΑΘΙΑΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗΝ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΑΚΤΙΝΟΓΡΑΦΙΩΝ ΘΩΡΑΚΟΣ

Θεοδώρα Χρυσούλα

Επιβλέποντες: **Δρ. Θεόδωρος Δαλαμάγκας,** Διευθυντής Ερευνών, Ινστιτούτο Πληροφοριακών Συστημάτων, Ερευνητικό Κέντρο «ΑΘΗΝΑ»
Δρ. Χρήστος Δίου
Επίκουρος Καθηγητής, Τμήμα Πληροφορικής και Τηλεματικής, Χαροκόπειο Πανεπιστήμιο

ΑΘΗΝΑ

ΦΕΒΡΟΥΑΡΙΟΣ 2023

MASTER THESIS

Explainable Artificial Intelligence for Deep Learning Methods in Chest X-Ray Classification

Theodora Chrysoula

A.M.: DS2200009

Supervisors:

Dr. Theodore Dalamagas
Research Director, Information Management Systems Institute,
ATHENA Research Center

Dr. Christos Diou
Assistant Professor, Department of Informatics and Telematics,
Harokopio University of Athens

EXAMINATION COMITEE:

Dr. Theodore Dalamagas, Research Director, Information Management Systems Institute, ATHENA Research Center

Dr. Christos Diou, Assistant Professor, Department of Informatics and Telematics, Harokopio University of Athens

Dr Manolis Koubarakis, Professor, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens

February 2023

ABSTRACT

Chest X-rays are a crucial tool for detecting abnormalities with the classification and localization of these diseases under intense research. The black box nature of deep learning algorithms necessitates the development of eXplainable Artificial Intelligence (XAI) methods. This study employs the VinBigData dataset, featuring 18,000 posterior-anterior (PA) images from Hospital 108 (H108) and Hanoi Medical University Hospital (HMUH) in Vietnam.

The focus of this study is on classifying six abnormalities ('Aortic Enlargement', 'Cardiomegaly', 'Lung Opacity', 'Pleural Effusion', 'Pleural Thickening' and 'Pulmonary Fibrosis') and a 'No-Finding' class which represents the absence of a disease. A pretrained ResNet50 on the ImageNet dataset is used, and Grad-Cam is the chosen XAI method. Evaluation of the XAI methods involves using the Intersection Over Union (IoU) metric to assess alignment between ground truth and predicted bounding boxes. Pixel importance analysis is also used for evaluation of the XAI method by replacing crucial pixels identified by Grad-Cam, with mean values in all three channels.

The model achieves a micro F1 score of 0.81, with 'No-Finding' obtaining the highest F1 score (0.96). 'Aortic Enlargement' and 'Cardiomegaly' show satisfactory F1 scores (0.86 and 0.83), while 'Lung Opacity' and 'Pulmonary Fibrosis' exhibit lower values (0.55 and 0.57). Examining Grad-Cam heatmaps reveals stable behaviour and localization for 'Aortic Enlargement' and 'Cardiomegaly'. However, other classes produce less reliable heatmaps, with 'Pleural Thickening' showing the least favourable results.

While this research provides encouraging outcomes, chest X-rays classification remains challenging, necessitating further research of XAI methods and evaluation processes.

SUBJECT AREA: Image Processing

KEYWORDS: Classification, Convolutional Neural Networks, chest X-rays, eXplainable Artificial Intelligence (XAI) methods, Grad-Cam

ΠΕΡΙΛΗΨΗ

Οι ακτινογραφίες θώρακα αποτελούν ένα σημαντικό εργαλείο για τον εντοπισμό διαφόρων παθολογιών στο θώρακα. Οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται για την κατηγοριοποίηση αυτών των ανωμαλιών χαρακτηρίζονται ως «μαύρα κουτιά» λόγω της αυξανόμενης πολυπλοκότητάς τους. Για την εφαρμογή αυτών των αλγορίθμων είναι απαραίτητη η χρήση και η ανάπτυξη μεθόδων ερμηνευσιμότητας (eXplainable Artificial Intelligence, XAI). Η παρούσα μελέτη χρησιμοποιεί το σύνολο δεδομένων VinBigData που περιλαμβάνει 18,000 οπισθο-πρόσθιας (PA) προβολής ακτινογραφίες θώρακα.

Στόχος αυτής της εργασίας είναι η ταξινόμηση και η δημιουργία χαρτών ερμηνείας έξι παθολογιών του θώρακα: ‘Aortic Enlargement’, ‘Cardiomegaly’, ‘Lung Opacity’, ‘Pleural Effusion’, ‘Pleural Thickening’, ‘Pulmonary Fibrosis’ και της κλάσης ‘No-Finding’ που αντιπροσωπεύει τις υγιείς ακτινογραφίες. Για την ταξινόμηση των ανωμαλιών χρησιμοποιείται ένα προ-εκπαιδευμένο ResNet50 μοντέλο στο σύνολο δεδομένων ImageNet και η μέθοδος ερμηνευσιμότητας είναι η Grad-Cam. Για την αξιολόγηση της Grad-Cam χρησιμοποιείται η μετρητική Intersection over Union (IoU) και η ανάλυση σημαντικότητας εικονοστοιχείων.

Το μοντέλο επιτυγχάνει ένα F1 score 0.81, με την κλάση ‘No-Finding’ να κατέχει την υψηλότερη τιμή. Οι κλάσεις ‘Aortic Enlargement’ και ‘Cardiomegaly’ παρουσιάζουν ικανοποιητικά αποτελέσματα, ενώ οι κλάσεις ‘Lung Opacity’ και ΄Pulmonary Fibrosis’ παρουσιάζουν τις χαμηλότερες τιμές. Από τους χάρτες ερμηνείας που προκύπτουν από την εφαρμογή της Grad-Cam, παρατηρείται η ικανότητα εντοπισμού των κλάσεων ‘Aortic Enlargement’ και ‘Cardiomegaly’, ενώ για τις υπόλοιπες κλάσεις παρουσιάζουν λιγότερα αξιόπιστα αποτελέσματα. Παρόλο, που αυτή η μελέτη καταλήγει σε ενθαρρυντικά αποτελέσματα, η ταξινόμηση των ακτινογραφιών θώρακα με μοντέλα βαθιάς μηχανικής μάθησης απαιτεί την περαιτέρω ανάπτυξη μεθόδων ερμηνευσιμότητας.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Επεξεργασία Εικόνας

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Κατηγοριοποίηση, Ερμηνεία Συνελικτικών Δικτύων, Ακτινογραφίες Θώρακα, Επεξηγηματική Τεχνητή Νοημοσύνη, Grad-Cam

ACKNOWLEDGMENTS

I would like to thank all the people who contributed to this research. First, I would like to thank professors Dr. Theodoros Dalamagas, Research Director of Information Management Systems Institute of the Athena Research Center and Dr Christos Diou, Assistant Professor at Department of Information and Telematics of Harokopio University of Athens for their guidance and their valuable advice throughout the entirety of this thesis. I would also like to thank Vasilis Gkolemis, Research Assistant of Athena Research Center, for the very good cooperation and the opportunity of discussing all the matters that came up in the thesis. Finally, special thanks to my friends and family for the support and courage during my studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

The master thesis “Explainable Artificial Intelligence in Chest X-rays” has been conducted at the ATHENA Research Center for the completion of the Postgraduate Program “Bioinformatics – Biomedical Data Science”, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece.

The first chapter introduces the need of eXplainable Artificial Intelligence (XAI) methods in medical imaging during the last years and the problem statement of this thesis.

In chapter 2, we provide a concise theoretical overview of the methods employed in this research, encompassing deep learning and neural networks, with a specific focus on convolutional neural networks widely utilized in computer vision techniques. Additionally, we touch upon common explainability methods.

Moving to chapter 3, we offer a more in-depth analytical description of the methods integral to our pipeline, placing particular emphasis on the detailed analysis of Grad-Cam and SmoothGrad. Subsequently, in Chapter 4, we present an overview of the evaluation methods used across various research studies and detail the methodology adopted in our research, including pixel importance analysis and the application of the Intersection over Union (IoU) method.

Chapter 5 delves into a detailed exploration of the dataset, our methodology, and the outcomes of our experiments. In chapter 6, we further expound upon the evaluation methods employed in our experiments. Finally, Chapter 7 initiates a comprehensive discussion of our results, accompanied by considerations of potential proposal or alternatives for future investigations.

1. Introduction

1.1 eXplainable Artificial Intelligence (XAI) Methods in Medical Images

In recent years, the number of Artificial Intelligence (AI) based applications for research and clinical care in medicine has increased dramatically, with medical imaging clearly being the focus of such developments. Specifically, deep learning techniques have been proven to be very useful tools in medical image analysis with tasks such as image classification, image segmentation or image detection. In deep learning, features such as edges or corners (low-level image properties) and higher-level image properties such as the spiculated border of a cancer are learned by a neural network to optimally give a result (or output) given an input. An example of a deep learning system could be the output ‘cancer’ given the input of an image showing a cancer.

Neural networks typically consist of many layers connected via many nonlinear intertwined relations. Even if one is to inspect all these layers and describe their relations, it is infeasible to fully comprehend how the neural network came to its decision. Therefore, deep learning is often considered a ‘black box’. Concern is mounting in various fields of application that these black boxes may be biased in some way, and that such bias goes unnoticed. Especially, in medical applications, this can have far-reaching consequences [1].

Past years’ research focused on implementing innovative and powerful system architectures, pursuing the goal of providing the best possible solution to several tasks. This led to increasingly opaque and complex systems. At the same time, explainability and interpretability suffered under this trend, resulting in increased difficulty in understanding the prediction process and inner workings of emerging solutions [2] .

As the number of parameters in machine learning models increases, it becomes more challenging for a human to understand the reasons behind a model’s decision, particular in critical domains like healthcare. A medical diagnosis system needs to be transparent, understandable, and explainable to gain the trust of physicians, regulators as well as the patients involved. Newer regulations like the European General Data Protection Regulation (GDPR) are making it harder for the use of black-box models in all business including

healthcare because retractability of the decisions is now a requirement [3]. In such cases, explainability becomes crucial for clinicians to comprehend the model's decisions and gain confidence in confirming diagnosis.

Research has focused on developing methods known as eXplainable Artificial Intelligence (XAI) to facilitate the explanation of a model's decisions. Various XAI approaches have emerged, categorized broadly into two types: model-agnostic and model-specific. Model-agnostic methods are designed to work with various machine learning models and do not depend on a specific architecture. On the other hand, model-specific methods are tailored to a particular type of model. By adopting these XAI methods, clinicians can gain specific insights into the results provided by the models and enhance their confidence. However, ensuring the credibility of these explanations necessitates the development of robust evaluation methodologies to assess their performance and determine their usefulness in practical applications.

1.2 Problem Statement

As mentioned above, XAI methods are of great importance in the healthcare domain. One of the most common medical images are Chest X-ray images (or CXR) that are widely used for the diagnosis of various chest pathologies. Distinguishing, between the different pathologies can be difficult due to the overlapping of the diseases or confusing abnormalities. Last years, many CXR datasets have been published, such as ChestX-ray8 and ChestX-ray14 [4], Padchest [5], CheXpert [6] and MIMIC-CXR [7] (**Error! Reference source not found.**). These datasets are not manually annotated (NLP tools or automated rule-based labeller that extract keywords from medical reports are used), posing significant issues related to the quality of the labels.

In 2020, VinBigData [8] (VINDr) Chest X-Ray Abnormalities Detection dataset was published, containing 18,000 posterior-anterior (PA) view CXR scans in DICOM format, which were di-identified to protect patient privacy. All images were labelled by 17 radiologists with at least 8 years of experience and they were manually annotated for the presence of 14 critical radiographic finding

visible in **Error! Reference source not found.**, along with the “no-finding” label indicating the absence of disease.

Table 1: An overview of existing public datasets for CXR interpretation.

Dataset	*Release Year*	*# findings*	*# samples*
ChestX-ray8 [4]	2017	8	108,948
ChestX-ray14 [4]	2017	14	112,120
CheXpert [6]	2019	14	224,316
PadChest [5]	2019	193	160,868
MIMIC-CXR [7]	2019	14	377,110
VinDr-CXR [8]	2020	28	18,000

VinBigData [9] group performed a classification task using EfficientNet [10] to distinguish six common lung diseases, including pneumonia, tuberculosis, lung tumor, pleural effusion, other diseases and no finding class with a mean F1-score of 0.631 and an object detection task using EfficientDet [11] to localize 14 critical findings from the CXR images, i.e. cardiomegaly, opacity, consolidation, atelectasis, pneumothorax, pleural effusion, aortic enlargement, interstitial lung disease (ILD), infiltration, nodule/mass, pulmonary fibrosis, pleural thickening, calcification and other lesions with the free-response receiver operating characteristic (FROC) analysis achieving a sensitivity of 80.2% at the rate of 1.0 false-positive lesion identified per scan.

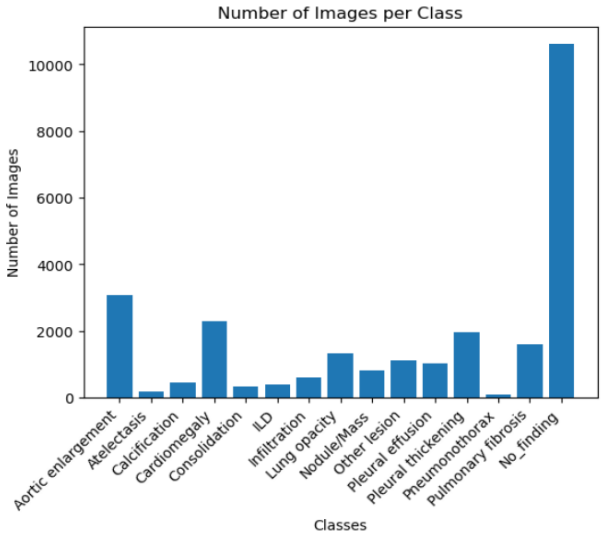


Figure 1: Distribution of the various pathologies in the VINDr dataset.

This research aims to develop an end-to-end method for the detection of CXR pathologies with a focus on employing eXplainable Artificial Intelligence (XAI) techniques and evaluating their efficacy. To this end, the VINDr dataset, sourced from kaggle, was chosen for its robustness, boasting high-quality labels provided by domain experts, along with detailed annotations highlighting critical findings in each CXR image. Building upon the groundwork laid by the VinBigData group [9] our study extends their efforts by undertaking a multilabel classification task. Specifically, we target six abnormalities identified within the VINDr dataset, in addition to categorizing healthy CXRs. This selection was made based on the prevalence of these abnormalities within the dataset, ensuring a representative sample of analysis. Impressively, our approach yields a noteworthy average micro F1 score of 0.81, signifying promising performance in pathology detection. Subsequently, we implement an explainability method to elucidate the salient features driving the classification decisions. Finally, thorough evaluation of the XAI method is conducted, encompassing comprehensive scrutiny of its effectiveness. In sum, this research encompasses the entirety of the classification process, the elucidation of crucial features via XAI, and meticulous evaluation, culminating in substantial findings poised to contribute significantly to the field of medical image analysis.

Our approach can be summarized as follows. The detection of the different pathologies is a multilabel task, where we input a chest X-ray image ang get seven labels (‘Aortic Enlargement’, ‘Cardiomegaly’, ‘Lung Opacity’, ‘Pleural Effusion’, ‘Pleural Thickening’, ‘Pulmonary Fibrosis’ and the ‘No-finding’ category indicating no disease) as output. These labels, denoted as y_i, are taking values between 0 and 1 indicating the existence or not, of each pathology. We used a pretrained ResNet50 [12] model trained on the ImageNet dataset. ResNet50 is advantageous because it incorporates skip connections or residual connections. These connections make it easier for information to move through the network, addressing issues like the vanishing gradient problem. This feature proves helpful, especially when training very deep networks.

To better understand the features involved to model’s decision, we adopted the Grad-Cam [13] explainability method which is widely used in medical imaging tasks and known for producing considerable results. In essence, Grad-Cam takes the gradient of the outputs with respect to the activation of the last convolution layer of the model, which is concerned to be the feature map with the more profound semantics. Subsequently, the heatmaps resulting from Grad-Cam are superimposed onto the original image to highlight important areas.

One of the main goals of this study is the evaluation of the XAI methods, as it is crucial to develop trustworthy methods to be able to assess the model’s decision and introduce this pipeline to medical domain. Two evaluation paths are employed:

1. The first method involves optical evaluation where the annotations of the different radiologists are compared with the respective heatmaps. The ground truth boxes of the diseases marked by the annotators are compared with the predicted bunding boxes generated form the heatmaps from Grad-Cam implementation. We computed the Intersection Over Union (IoU) metric to assess the overlapping between these boxes.
2. In the second approach, we identify the most important pixels based on the Grad-Cam results. These significant pixels are then subjected to

Gaussian blurring or replaced with the mean values across all three channels with varying percentages of these key pixels replaced in each iteration. Subsequently, we feed these modified blurred images to the model and observe its resultant prediction. This procedure aids in comprehending the specific contribution of these pixels to the model's final prediction. It serves as a means of assessing how changes in the pixels highlighted by Grad-Cam impact the prediction of each disease.

The primary contribution of this research lies in the development of an end-to-end pipeline for addressing the CXR classification problem. It offers insights into the application of eXpalinable Artificial Intelligence (XAI) methods in such tasks by proposing evaluation methodologies to assess their performance effectively. An overview of the pipeline is presented in Figure 2. In summary the key contributions include:

- Utilizing a pretrained ResNet50 model for a multilabel task achieving state-of-the-art performance in CXR datasets with an average micro F1 Score of 0.81 (ranging from 0.55 to 0.96) and AUC scores ranging from 0.93 to 0.99.
- Implementing the Grad-Cam explainability method to highlight important regions within CXR datasets.
- Developing two evaluation paths (Intersection over Union and Pixel Importance Analysis) to assess the performance of XAI method.

Results of the evaluation pipeline show interesting findings. A reduction in the prediction probability of each class is observed each time the crucial pixels of the image are replaced ranging from 0.1 to 0.7. IoU metric demonstrates alignment for the classes 'Aortic Enlargement' and 'Cardiomegaly' with values 0.201 and 0.237, respectively). However, it does not yield significant results for the remaining classes.

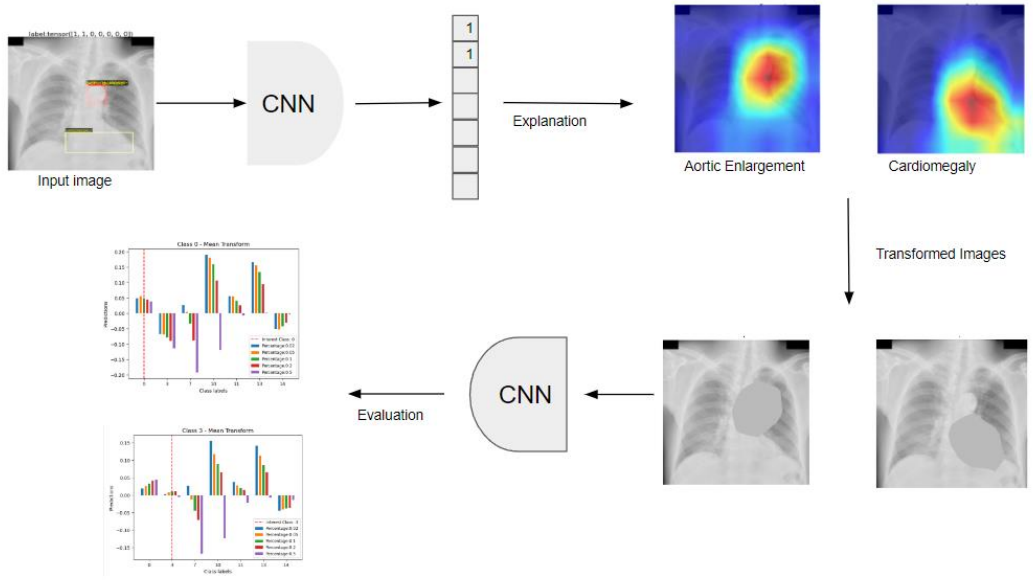


Figure 2: An overview of the pipeline: An input CXR image is fed into the model for multilabel classification to identify apparent diseases, such as Aortic Enlragement and Cardiomegaly in this example. Heatmaps generated from Grad-Cam highlight regions of abnormalities. The images undergo transformations based on the most important pixels indicated by Grad-Cam. These modified images are then fed back into the model, and plots illustrating the changes in predictions are obtained.

2. BACKGROUND AND RELATED WORK

2.1 Machine Learning Techniques, Dense Neural Networks, Convolution Neural Networks

In recent years, various statistical machine learning techniques classifiers such as support vector machines and decision trees have been employed for image recognition and classification. However, the emergence of deep neural networks (DNNs) has led to significant advancements in image classification, particularly in tasks involving complex pattern and motif recognition. One of the most prominent architectures in this domain is the Convolution Neural Network (CNN). The name “convolutional” arises from the mathematical operation called convolution, which is pivotal to its functioning.

CNNs have multiple layers, including convolutional layer, non-linearity activation layers (sigmoid, tanh, or ReLU), pooling layers and fully connected layers. While convolutional and fully connected layers feature learnable parameters, pooling and non-linear activations do not. This amalgamation of layers allows CNNs to exhibit exceptional performance in machine learning tasks [14].

- **Convolution Layer**

The convolution layer is the core building block of CNN. This layer performs a dot product between two matrices, where one matrix is the set of learnable parameters (kernel or filter) and the other matrix is the restricted portion of the receptive field form the input image. The kernel is spatially smaller than an image but is deeper in terms of dimensions. For instance, if the image is composed of three (RGB) channels, the kernel’s height and width will be spatially compact, while its depth extends up to all three channels.

During the forward pass, the kernel slides across the height and width of the image producing the image representation of each receptive region. This produces, two-dimensional representation of the image known as activation map or feature map depicting the kernel’s response at every spatial position. The sliding size of the kernel is called a stride.

If we have an input of size NxNxC and Dout number of kernels with a spatial size of F with stride S and amount of padding P, the output volume's dimensions are determined by the following formula:

$$\boldsymbol{Nout} = \frac{\boldsymbol{N - F + 2P}}{\boldsymbol{S}} + \mathbf{1} \qquad (\mathbf{1})$$

This will yield an output volume of size Nout x Nout x Dout.

CNNs possessed distinct characteristics that make them effective in pattern recognition problems compared to traditional artificial neural networks. A key trait is the reduction in learnable parameters of the network. In conventional neural networks, every output unit interacts with every input unit. In contrast, convolution neural networks exhibit sparse interaction due to their use of smaller kernels. For instance, an image can have millions or thousands of pixels, but while processing it using kernel, we can detect meaningful information that is of tens or hundreds of pixels. This means that we need to store fewer parameters that not only reduces the memory requirement of the model but also improves the statistical efficiency of the model.

If computing one feature at a spatial point (x_1, y_1) is useful, then it should also be useful at some other spatial point (x_2, y_2). It means that for a single two-dimensional slice i.e., for creating one activation map, neurons are constrained to use the same set of weights. In a traditional neural network, each element of the weight matrix is used once and then never revisited, while convolution network has shared parameters.

Due to parameter sharing, the layers of convolution neural network will have a property of equivariance to translation. This property signifies that if we change the input in a way, the output will also be modified in the same way.

- **Pooling Layer**

Pooling layers play a role in subsampling the network's output at specific locations by computing a summary statistic of nearby outputs. This helps in reducing the spatial dimensions of the representation, which decreases the required amount of computation individually. Common pooling functions include averaging over a rectangular neighborhood and max pooling, which retains the maximum output within a neighborhood.

- **Fully Connected Layer**

Neurons in this layer have fully connectivity with all neurons in the preceding and succeeding layer. This is why it can be computed as usual by a matrix multiplication followed by a bias effect. The fully connected layer helps to map the representation between the input and the output.

- **Non-linearity Layers**

Since convolution is a linear operation, non-linearity layers are often placed directed after the convolutional layer to introduce non-linearity to the activation map. There are several types of non-linear operations, among the more popular are sigmoid, tanh and ReLU, with ReLU being the most widely used activation function. It computes the function $f(x) = \max(x, 0)$, essentially applying a threshold at zero [15].

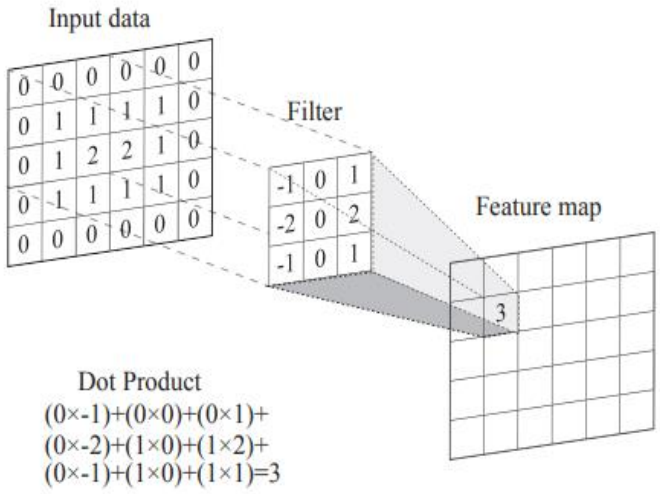


Figure 3: Calculation of a single dot product. The calculation of the dot product involves an element-wise multiplication between convolutional filter and the matching grid in the input data. The resulting values are summed obtaining a single number that is stored in the central pixel in the feature map [16].

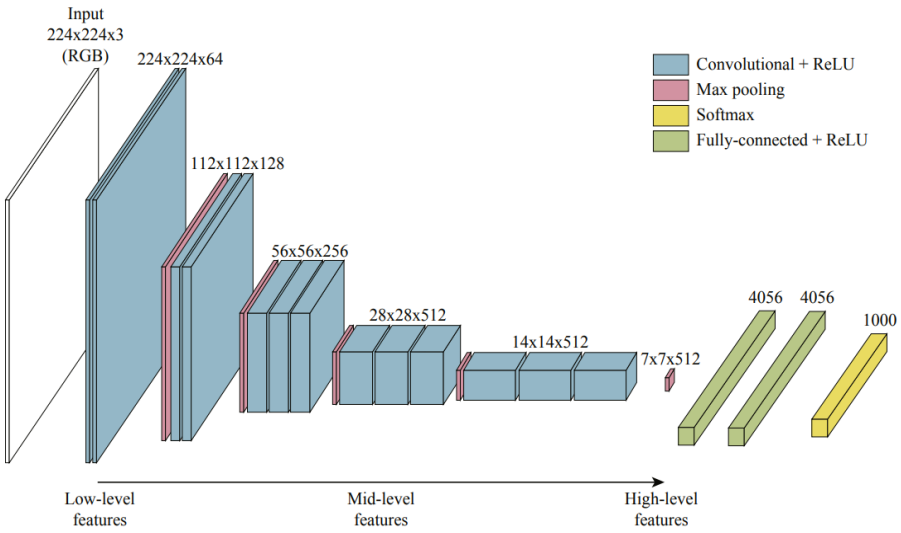


Figure 4: A schematic of the VGG16 [17].

2.2 Residual Networks (ResNets)

As mentioned before, Convolutional Neural Networks (CNNs) have made a significant breakthrough in image classification tasks. However, an important consideration is the extent to which we can optimize a model? Increasing the depth of a model can enhance its ability to recognize complex features and functions, potentially leading to improved accuracy results. This was the main idea behind VGG model architecture introduced by Karen Simonyan and Andrew Zisserman in 2015 [17].

Nevertheless, it has been observed, that the increasing the depth of a neural network does not necessarily conclude in better training accuracy. In fact, at a certain point the training error may start to increase instead of decreasing. This phenomenon is known as the degradation problem. It occurs when, as the network depth increases, the accuracy saturates and then begins to degrade rapidly if more layers are introduced (Figure 5).

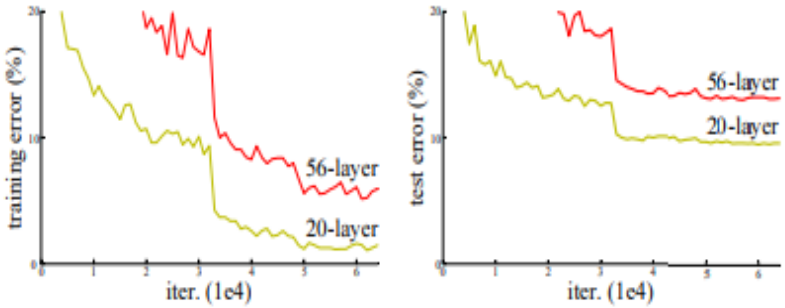


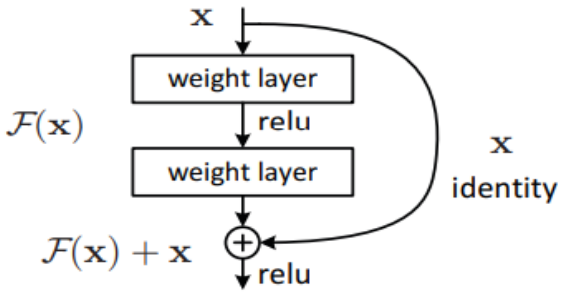
Figure 5: Training error (left) and test error (right) in CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper the network, the higher the training error, and thus the test error [12].

Researchers Kaiming He, Xiangyu Zhang and Shaoqing Ren [12] introduced in the paper Deep Residual Learning for Image Recognition the ResNets architecture where they address the degradation problem by introducing a deep residual framework. In ResNets, a technique called skip connections is used. Skip connections connects activations of a layer to further layers by skipping some layers in between. This forms a residual block. Instead of layers learning the underlying mapping, they allow the network to fit the residual mapping. Formally, denoting the desired underlying mapping as $H(x)$, they let the stacked nonlinear layer fit another mapping of $F(x) := H(x) - x$. The original mapping is recast into $F(x) + x$. They propose that is easier to optimize the residual mapping than to optimize the original.

The formulation of $F(x) + x$ has been realized by feedforward neural networks with shortcuts connections, which skip one or more layers. In case of ResNets shortcut connections perform identity mapping, and their outputs are added to the outputs of the stacked layers. They adopt residual learning to every few stacked layers. A building block is shown in Figure 6 and is defined as:

$$y = F(x, \{W_i\}) + x \quad (2)$$

Here x and y are the input and the output vectors of the layers considered. The function $F(x, \{W_i\})$ represents the residual mapping to be learned. For example,



in that has two layers, $F = W_2\sigma(W_1x)$ in which σ denotes ReLU. The operation $F(x) + x$ is performed by a shortcut connection and element-wise addition. The second nonlinearity is adopted after the addition $\sigma(y)$. Shortcut connections introduce neither extra parameter nor computation complexity, enabling the comparisons between plain and residual networks. It should be noted that the dimensions of x and F must be equal in equation 2.

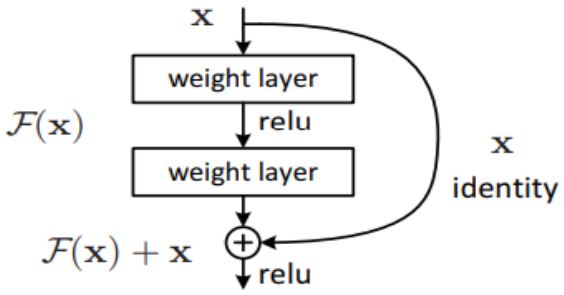


Figure 6: Residual learning: a building block [12].

Figure 7: Example network architectures for ImageNet. Left: the VGG-19 model (19.6 billion FLOPs) as a reference. Middle: a plain network with 34 parameters layers (3.6 billion FLOPs) as a reference. Right: a residual network with 34 parameters layers (3.6 billion FLOPs). The dotted shortcuts increase dimension so a stride of 2 is used [12].

2.3 Overview of XAI Methods

2.3.1 Taxonomy of XAI Methods

In general, there exist various classifications of the different explainability methods, depending on the method's characteristics. These methods can be concurrently grouped into many over-lapping or non-overlapping categories [3]. We present some of the criteria employed in the categorization of the XAI methods, and a summarizing flowchart is shown in Figure 8.

- **Model Specific or Model Agnostic**

Model-specific interpretation methods are based on the parameters inherent to individual models. In contrast model-agnostic methods are independent of a model's internal parameters they are applicable in post-hoc analysis. Consequently, these methods can be used on any machine learning model.

- **Local or global Method**

Local interpretable methods are applicable to a single prediction of the model. This can be done by designing methods that can explain the reason for a particular prediction focusing on specific features and their attributes. On the contrary, global methods concentrate on the inside of a model by exploiting the overall knowledge about the model, its training, and the associated data. It tries to explain the behaviour of the model in general. Feature importance is a representative example of this method, which tries to figure out the features which are in general responsible for better performance of the model among all different features.

- **Surrogate Methods or Visualization Methods**

Surrogate methods consist of different models as an ensemble which are used to analyse other black-box models. The black box models can be understood better by interpreting surrogate model's decision. The decision tree is an example of surrogate methods. The visualization methods are not a different model, but it helps to explain some parts of the models by visual understanding like activation maps [3].

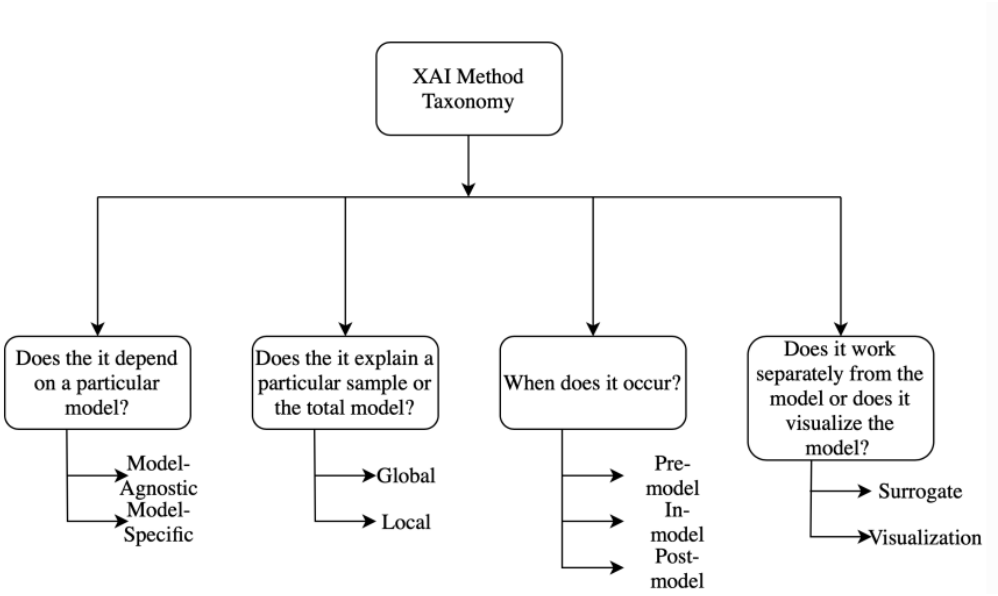


Figure 8: Taxonomy of XAI methods [3].

2.3.2 XAI Methods

Pixel attribution methods highlight the pixels that were relevant for a certain image classification by a neural network.

2.3.2.1 Occlusion – or perturbation-based

Methods like SHAP and LIME manipulate parts of the image to generate explanations. These methods investigate properties of DNNs by perturbating the input of a model. One of the most popular model-agnostic method is LIME (Local Interpretable Model-Agnostic Explanations) proposed by Ribeiro et al in 2016 (18). In this paper, the authors propose a concrete implementation of local surrogate models. Surrogate models are trained to approximate the predictions of the underlying black box model. The main idea is that LIME generates a new dataset consisting of perturbed samples and the corresponding predictions of the black box model. On this new dataset LIME trains an interpretable model, which is weighted by the proximity of the sampled instances to instance of interest.

Another famous method is SHAP (Shapley Additive exPLanations) by Lundeberg and Lee (2017) [19]. The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction. The SHAP explanation method computes Shapley values from coalitional game theory. The feature values of a data instance act as players in a coalition.

Shapley values implies how to fairly distribute the prediction among the features. A player could be an individual feature (tabular data) or a group of features – superpixels (image data) [20].

2.3.2.2 Gradient based methods [20].

Many methods compute the gradient of the prediction (or classification score) with respect to the input features. The gradient-based methods mostly differ in how the gradient is computed. Some of the most widely used gradient-based methods are Vanilla Gradient (Saliency Map), DeconvNet, Grad-Cam, Guided Grad-Cam, Smooth-Grad, and Layer Wise Propagation (LRP).

- **Vanilla Gradient (Saliency Map)**

One of the first pixel attribution methods is Vanilla Gradient introduced be Simonyan et al. (2013) [21]. The idea behind this method is to calculate the gradient of the loss function for the class in interest with respect to the input pixels. This provides us with a map of the size of the input features with negative to positive values. The recipe for this approach is:

1. Perform a forward pass of the image of interest.
2. Compute the gradient of class score of interest with respect to the input pixels:

$$E_{grad}(I_0) = \left.\frac{\delta S_c}{\delta I}\right|_{I = I_0} \quad (3)$$

   Here we set all other classes to zero.

3. Visualize the gradients. You can either show the absolute values or highlight negative and positive contributions separately.

Vanilla Gradient has a saturation problem, when ReLU is used, and the activation at the previous layer is below zero, then the activation is capped at zero and does not change anymore.

- **Deconvolution Network (DeconvNet)**

In 2014 DeconvNet was proposed by Zeiler and Fergus [22]. The goal of DeconvNet is to reverse a neural network and the paper proposed operations that are reversals of the filtering, pooling, and activation layers. DeconvNet is

equivalent to Vanilla Gradient, but it makes a different choice for backpropagating the gradient through ReLU:

$$R_n = R_{n+1} I(R_{n+1} > 0) \quad (4)$$

where Rn and R_{n+1} are the layer reconstructions and I the indicator function. When backpropagating from layer n+1 to layer n, DeconvNet remembers which of the activations in layer n+1 was set to zero in the forward pass and sets them to zero in layer n. Activations with negative value in layer n+1 are set to zero in layer n.

- **Layer Wise Propagation (LRP)**

Layer Wise Propagation (LRP) is an XAI method proposed by Gregoire Montavon and Sebastian Blach [23] and is based on deep Taylor decomposition. The basic idea is to associate to every pixel (p) of the input image a relevance score $R_p(x)$, that indicates for an image x to what extent the pixel p contributes to explaining the classification decision f(x). The relevance of each pixel can be stored in a heatmap denoted by $R(x) = \{R_p(x)\}$ of same dimensions as the image x. The heatmap can therefore also be visualized as an image. The method begins from the last layer by taking the relevance score of the output and redistributes in the proceeding layers, till the input image. The equation behind this process is the following:

$$R_i = \sum_j \frac{w_{ij}^2}{\sum_{i'} w_{i'j}^2} R_j \quad (5)$$

Where R_i and R_j are the relevance scores at layers i and j respectively and w_{ij} the weights that connect these layers.

- **Grad-Cam (Gradient-weighted Class Activation Map)**

Grad-Cam (Gradient-weighted Class Activation Map) described in the paper [13] by Reamparasaath R. Selavaraju et al. uses the gradients of any target concept (say a dog in a classification network or a sequence of words in captioning network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. The difference with the previous methods is that here

we do not backpropagate all the way back to the image, but (usually) to the last convolution layer. More details in Grad-Cam will be given in chapter 3.1.

- **Guided Grad-CAM**

Since Grad Cam uses the last convolution layer, one can suggests that the convolution feature maps have a much coarser resolution compared to the input image. In contrast, other attribution methods backpropagate all the way to the input pixels. They are therefore much more detailed and can show individual edges or spots that contributed most to a prediction. A fusion of both method is called Guided Grad-CAM. The idea is that one can compute for an image both the Grad Cam explanation and the explanation from another attribution method, such as Vanilla Gradient. The Grad Cam output is then unsampled with bilinear interpolation, then both maps are multiplied element-wise Grad Cam works like a lens that focuses on specific parts of the pixel-wise attribution map.

- **SmoothGrad**

The idea of SmoothGrad by Smilkov et al. [24] is to make gradient-based explanations less noisy by adding noise and averaging over these artificially noisy gradients. SmoothGrad is not a standard explanation method, but an extension to any gradient-based method.

In Figure 9 presented below, various instances depict the application of distinct Explainable Artificial Intelligence (XAI) methods trained with VGG-16 [17] on the ImageNet dataset. The corresponding images are accompanied by the classification scores assigned by the neural network. Upon closer examination of these examples, it becomes clear that assessing the reliability of an explanation method poses a considerable challenge. Certain explanations resonate with human intuition, such as the clear highlighting of an octopus by SmoothGrad and vanilla methods, or the accurate identification of a dog in the initial image. However, it is notable that in some instances, Grad-Cam appears to produce results that may be perceived as less coherent.

A notable challenge inherent in most XAI methods is the absence of a definitive ground truth for the explanations they provide. Consequently, at this juncture, our evaluation is limited to accepting or rejecting an explanation based on its interpretability and alignment with human understanding. The inherent

ambiguity underscores the complexity of evaluating XAI methods and emphasizes the need for careful consideration in accepting or dismissing explanations within the given context [20].

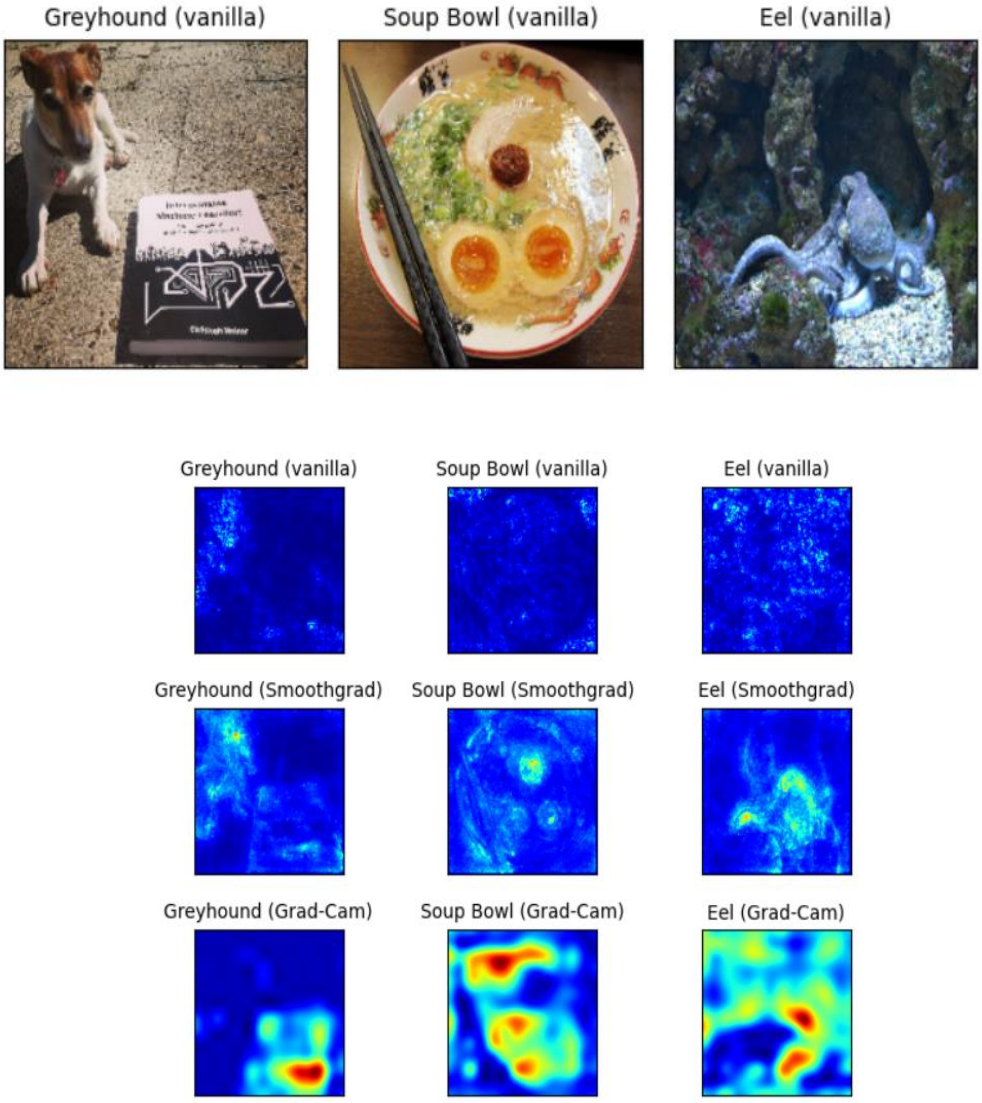


Figure 9: a) Images of a dog classified as greyhound (35%), a ramen soup classified as soup bowl (50%) and octopus classified as eel (70%). b) Pixel Attributions or saliency maps for the Vanilla Gradient method, Vanilla Gradient + SmoothGrad and Grad-Cam [20].

2.3.2.3 Path-Attribution Methods / Integrated Gradients (IG)

There are also path-attribution methods which compare the current image to a reference image, which can be an artificial “zero” image such as a completely grey image. The difference in actual and baseline prediction is divided among

the pixels. Integrated Gradients (IG) [25] is a widely used model-specific method of this category. It computes the gradient of the model's prediction output to its input features and requires no modification to the original deep neural network.

For the explanation of Integrated Gradients, suppose we have a function $F:R^n \rightarrow [0,1]$ that represents a deep network. Specifically, let $x \in R^n$ be the input at hand and $x' \in R^n$ be the baseline input. For image networks, the baseline could be a black image. We consider the straightline path (in R^n) from the baseline x' to the input x and compute the gradients at all points along the path. Integrated gradients are obtained by cumulating these gradients. Specifically, integrated gradients are defined as the path integral of the gradients along the straightline path from the baseline x' to the input x.

Consequently, the gradients are computed to measure the relation between the change in one feature and the change in the output of the model. Gradients inform us which pixel has the more powerful impact to the predicted class of the model.

The integrated gradient along the i^{th} dimension for an input x and baseline x' is defined as follows:

$$Integrated\ Gradients_i(x) ::= (x_i - x_i') \times \int_{a=0}^{1} \frac{\theta F(x' + a \times (x - x'))}{\theta x_i} da \quad (6)$$

Here, $\frac{\theta F(x)}{\theta x_i}$ is the gradient of the F(x) along the i^{th} dimension.

2.4 Background

Several studies have been published with the use of convolution neural networks for classification tasks in the medical domain, with impressive accuracy rates. The classification tasks in medical imaging could be Magnetic Resonance Imaging (MRI) (brain tumour detection [26], [27]), chest X-Rays (binary or multi class classification [28], [29], [4]) images. As mentioned before the use of machine learning techniques in the medical domain is of great importance. To make this happen, the black box nature of neural networks

should be overcome. Because of this, the use of XAI methods in these classification tasks has been of critical importance. For example, Windisch et al. [26] used Grad Cam to show which areas of brain MRI made the classifier decide on the presence of tumor and Böhle et al. [30] used LRP for identifying regions responsible for Alzheimer's disease from brain MRI images.

Chest X-ray (CXR) images are very important in the diagnosis about a patient's condition, consequently the correct classification and interpretability of such images should be considered. The distinguish between the various pulmonary diseases is challenging due to their high-inter class similarities and low inner-class variant abnormalities, especially given the complex nature of radiographs and the complex anatomy of chest [28].

With the outbreak of COVID-19 pandemic, numerous classification studies have been published, focusing on distinguishing COVID-19 disease, pneumonia, and healthy lung conditions. In paper [31], the authors employed a VGG-16 architecture to analyze 6,523 chest X-Rays collected from various medical institutes. Their model achieved a notable 96% accuracy in discerning between healthy chest X-rays and those indicative of pulmonary diseases. Moreover, the classification performance was impressive, with a 98% accuracy in correctly identifying images associated with COVID-19 or other health conditions. Additionally, the authors employed the Grad-Cam XAI method to elucidate critical regions in the X-rays, enhancing the interpretability of their model for COVID-19 prediction, illustrated in Figure 10.

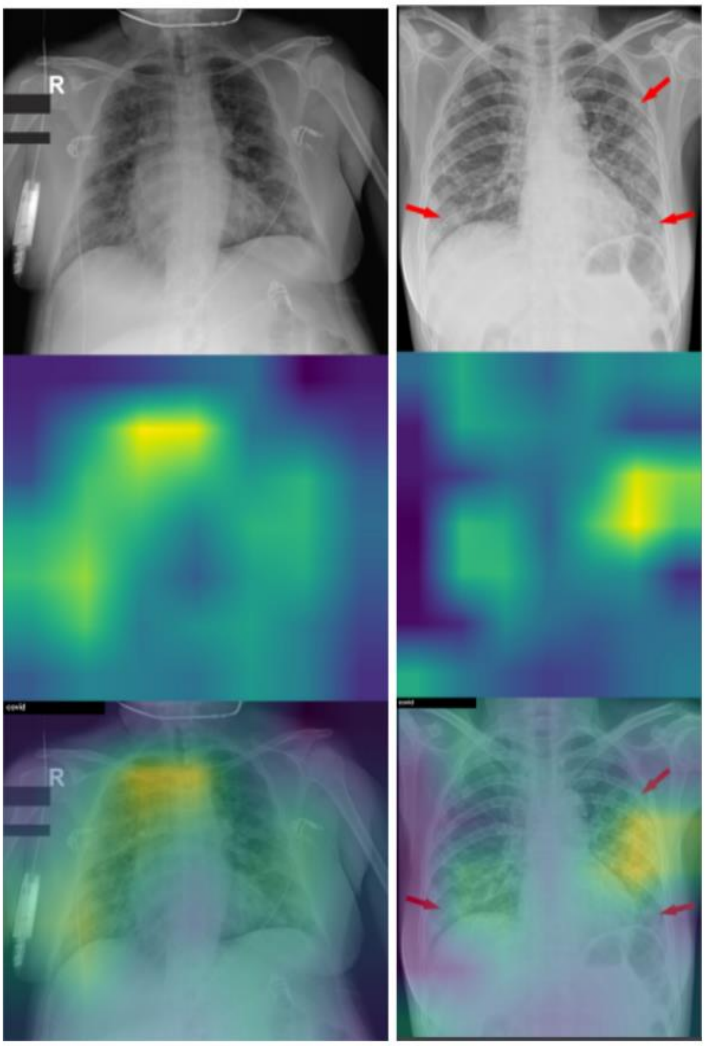


Figure 10: Examples of COVID-19 model activation maps [31].

Many state-of-the-art methods have made efforts for the multilabel classification problem. For instance, Wang et al. [4] developed an ChestX-Ray 8 dataset, which comprises 108,948 frontal-view X-ray images of 32,717 unique patients with text-mined eight disease image labels (where each image can have multi-labels), from the associated radiological reports using natural language processing Figure 11. They utilized the ImageNet pretrained deep CNN models, i.e., AlexNet [32], VGGNet [17], GoogleNet [33] and ResNet [12], to perform multi-label thoracic disease classification, which led to mass enthusiasm for the automated CXR analysis task. Their model achieved high AUC scores for class ‘Cardiomegaly’, while facing challenges with classes such as ‘mass’ due to its huge within-class appearance variation or ‘Pneumonia’,

possibly due to the limited instances in their patient population (less than 1% of total instances).

In their evaluation, Wang et al. utilized Intersection Over Union (IoU) metric to estimate the overlap between the predicted and the ground truth bounding boxes. Additionally, they considered the Intersection over the detected B-Box area ratio (BBIoU) which focused on the predicted area, unlike IoU which considers the union, BBIoU becomes relevant when there is no high overlap between the predicted bounding box and the ground truth area but still captures a considerable portion. The results, as illustrated in Figure 12, demonstrate better agreement, particularly in the case of 'Cardiomegaly' class.

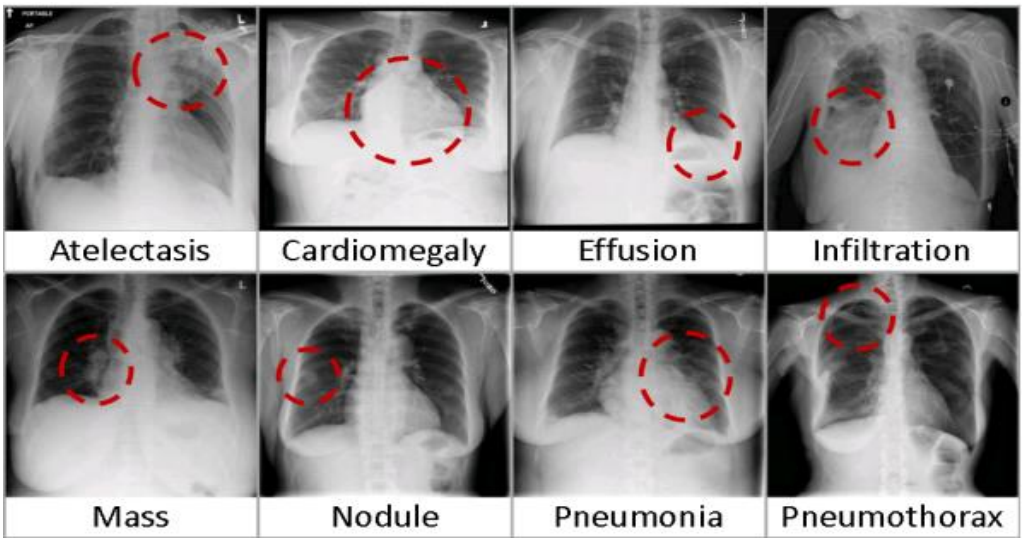


Figure 11: Eight common thoracic diseases observed in chest X-rays that validate a challenging task of fully automated diagnosis [4].

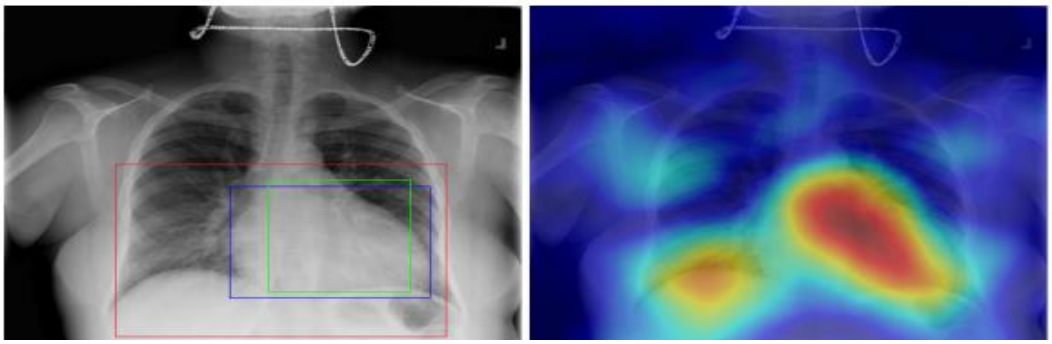


Figure 12: Localization result from 'Cardiomegaly' class. Correct bounding box (in green), false positive (in red) and the ground truth (in blue) are plotted over the original image [4].

Rajpurkar et al. [29] developed an algorithm CheXNet for the detection of lung-pneumonia on the ChestX-ray14 dataset with 112,120 frontal-view X-ray images with 14 diseases, where the test dataset is annotated by four radiologists. CheXNet is a 121-layer Dense Convolutional Network [34] that inputs a chest X-ray image and outputs the probability of pneumonia along with a heatmap provided by cam algorithm localizing the areas of the image more indicative of pneumonia (Figure 13). They compared f1 metric of their model with that of the annotators and they realized that their performance exceeds average radiologist performance on the f1 metric. They even extend CheXNet to detect all 14 diseases in ChestX-ray14 with very promising results.

In 2019, CheXpert [6] dataset was published consisting of 224,316 chest radiographs of 65,240 patients from Stanford Hospital, where the presence of 14 observations in radiology reports was detected, capturing uncertainties inherent in radiograph interpretation. Different uncertainty policies for the training of the convolutional neural networks were investigated that led to useful results in different pathologies. The test dataset consists of 500 chestX-rays annotated by a consensus of 5 board-certified radiologists. The authors compared the performance of their model with that of 3 additional radiologists in the detection of 5 selected pathologies.

The training procedure compared different models and concluded that DenseNet-121 [34] produced the best results, so they performed all the experiments with DenseNet. As they use different uncertainty policies the choose the best 10 checkpoints per run using the average AUC across the competition tasks. They run each model three times and take the ensemble of the 30 generated checkpoints by computing the mean of the output probabilities over the 30 models. They manage impressive AUC scores (the best is 0.97 for Pleural Effusion and the worst is 0.85 for Atelectasis). The model achieves higher results performance than the 3 radiologists on the test set in most of the cases. Finally, they visualize the areas of the radiograph which the model predicts to be most indicative of each observation using Grad-Cam (Figure 14).

In 2021 Hieu H. Pham et al. [9] developed an explainable deep learning system called VinDr-CXR, that can classify a CXR scan into multiple thoracic diseases, and at the same time, localize most types of critical findings on the image.

VinDr-CXR was trained on 51,485 CXR scans [8] with radiologist-provided bounding box annotations. The model's performance was validated on a separate set of 3,000 CXR scans, resulting in an F1 score of 0.631. The core of the VinDr-CXR system is based on DL-networks EfficientNet [10] and EfficientDet [11]. To assess robustness of their system, the researchers conducted evaluations on different datasets, including CheXpert [6] and CheXphoto [35].

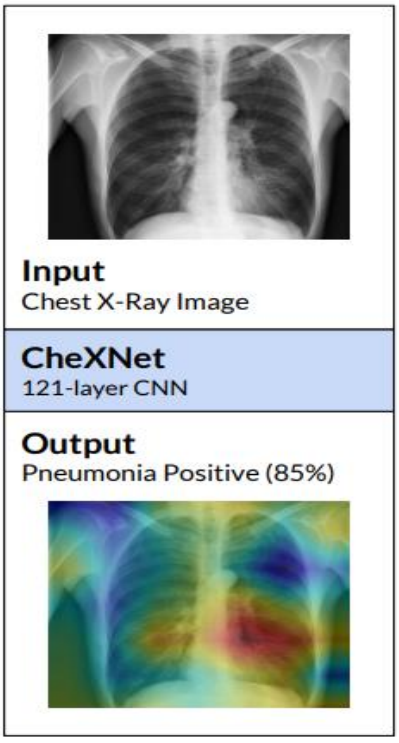


Figure 13: CheXNet is a 121-layer convolutional neural network that takes a chest X-ray image as input, and outputs the probability of a pathology. On this example, CheXNet correctly detects pneumonia and localizes areas in the image indicative of the pathology [29].

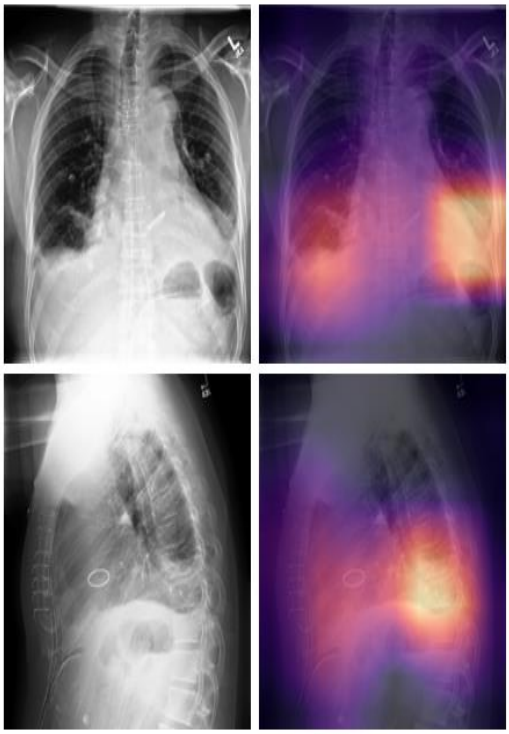


Figure 14: Frontal and lateral radiographs of the chest in a patient with bilateral pleural effusions on both the frontal (top) and the lateral (bottom) views, with predicted probabilities p = 0.936 and p = 0.939 in the frontal and lateral views respectively [6].

3. XAI METHODS AND EVALUATION TECHNIQUES

3.1 Grad Cam

The main method employed in this research is Grad-Cam. According to the research cited in [2], Grad-Cam has gained popularity among various explainability methods in the medical domain over the last few years, as shown in **Error! Reference source not found.**. Grad-Cam was introduced by researchers in 2017 in paper 'Grad Cam: Visual Explanations from Deep Networks via Gradient-based Localization' [13]. Grad-Cam, like other pixel attribution techniques, assigns each neuron a relevance score for the decision of interest. This decision of interest could be the class prediction (output layer), but theoretically it could be any other layer in the neural network. It can be used with different CNNs including fully connected layers, for structured output such as captioning and in multi-task outputs and for reinforcement learning [20].

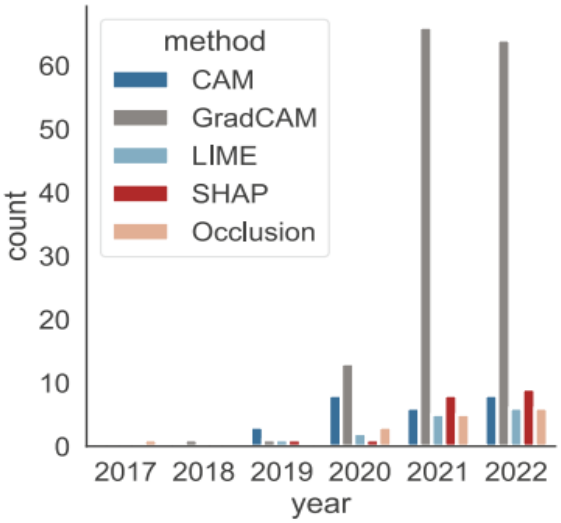


Figure 15: Annual development of Top 5 saliency-based XAI methods applied in medical image analysis based on the total number of citations [2].

Several previous works have asserted that deeper representations in a CNN capture higher-level visual constructs. Furthermore, convolutional layers naturally retain spatial information, which is lost in fully connected layers, so we can expect the last convolutional layers to have the best compromise between high-level semantics and detailed spatial information. The neurons in these layers look for semantic class-specific information in the image like object parts. Grad-Cam uses the gradient information flowing into the last convolution layer

of the CNN to assign importance values to each neuron for a particular decision of interest [13].

The goal of Grad-Cam is to understand which parts of an image, a convolution layer searches for a certain classification. To understand how CNN makes decisions, Grad-Cam analyzes which regions are activated in the feature maps of the last convolutional layers. There are k feature maps in the last convolutional layer, which are noted as A_1, A_2, …., A_k. Grad-Cam is interested in deciding which of the feature k maps is important for our class of interest c, by setting all the other classes to zero. As shown in Figure 16, in order to obtain class-discriminative localization map Grad Cam, $L^c_{Grad-CAM} \in \mathbb{R}^{u\ x\ v}$ of width u and height v for any class c, we first compute the gradient score for class c, y^c (before the softmax), with respect to feature map activations A^k of a convolution layer, i.e. $\frac{\partial y^c}{\partial A^k_{ij}}$. These gradients flowing back are global-average-pooled over the width and height dimensions (indexed by i and j respectively) to obtain the neuron importance weights a^c_k.

$$a^c_k = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{global\ average\ pooling} \underbrace{\frac{\partial y^c}{\partial A^k_{ij}}}_{gradients\ via\ backprop} \qquad (6)$$

where $A^k_{i,j}$ is a neuron positioned at (i,j) in the (u,v) feature map A_k and $Z = u\ x\ v$. During computation of a^c_k while backpropagating gradients with respect to activations, the exact computation amounts to successive matrix products of the weight matrices and the gradient with respect to the activation functions till the final convolution layer that the gradients are being propagated to. Hence, this weight a^c_k represents a partial linearization of the deep network downstream from A, and captures the 'importance' of feature map k for a target class c.

We perform a weighted combination of forward activations maps, and follow it by a ReLU to obtain,

$$L^c_{Grad=CAM} = ReLU \underbrace{\left(\sum_k a^c_k A^k\right)}_{linear\ combination} \qquad (7)$$

The resulting heatmap will have the same dimensions as the convolutional feature map, which in the case of ResNet50 is 7 x 7. We apply a ReLU to the linear combination of maps because we are only interested in the features that have a positive influence on the class of interest, i.e. pixels whose intensity should be increased to increase y^c. Negative pixels are likely to belong to other categories in the image. As expected, without this ReLU, localization maps sometimes highlight more than just the desired class and perform worse at localization.

Practitioners can use the CAM family of the methods to determine, given an input and a class, what is the information in the input that gives evidence for that class. Based on this information the practitioner can determine to what extent model predictions can be interpreted and assess for which classes consistent model predictions can expected. For example, if we have two models where both models have the same accuracy score, a model that produces heatmaps consistent with human experience is often considered more trustworthy compared to one where the heatmaps correspond poorly to human experience. Practitioners can also use the CAM family of methods to determine if there is an unfavourable class bias that the model is picking up on e.g., skin colour.

To sum up, the steps that are followed in Grad-Cam to obtain the localization map, $L^c_{Grad-CAM} \in \mathbb{R}^{u\ x\ v}$, are the following:

1. Forward-propagate the input image through the convolutional neural network.
2. Obtain the raw score for the class of interest, meaning the activation of the neuron before the softmax layer.
3. Set all other class activations to zero.
4. Back-propagate the gradient of the class of interest to the last convolution layer before the fully connected layers: $\frac{\partial y^c}{\partial A^k}$.
5. Weight each feature map “pixel” by the gradient for the class. Indices i and j refer to the width and height dimensions.
6. Calculate an average of the feature maps, weighted per pixel by the gradient.

7. Apply ReLU to the averaged feature map.
8. For visualization: Scale values to the interval 0 and 1. Upscale the image and overlay it over the original image.

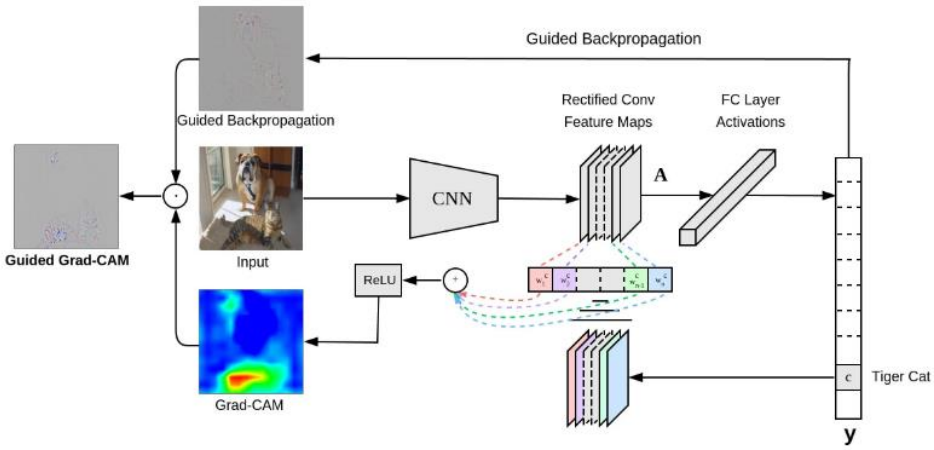


Figure 16: An overview of the Grad-Cam. Given an image and a class of interest (e.g. 'tiger cat' as input, we forward propagate the image through the CNN part of the model and through task-specific computations to obtain a raw score for the category. The gradients are set to zero for all classes except the desired class ('tiger cat'), which is set to 1. This signal is then backpropagated to the rectified convolutional feature maps of interest, which we combine to compute the coarse Grad-Cam localization (blue heatmap) which represents where the model has to look to make the particular decision [13].

3.2 SmoothGrad [20]

We also use SmoothGrad in combination with Grad Cam to examine if we have more located explanations. The main idea of SmoothGrad is to add noise to multiple samples and average the resulting heatmaps. SmoothGrad works in the following way:

1. Generate multiple versions of the image of interest by adding noise to it.
2. Create pixel attribution maps for all images.
3. Average the pixel attribution maps.

It is quite a simple idea, that according to the theory it should work as the derivative fluctuates greatly at small scales. Neural networks have no incentive during training to keep the gradients smooth, their goal is to classify images correctly. Averaging over multiple maps "smooths out" these fluctuations:

$$R_{sg}(x) = \frac{1}{N}\sum_{i=1}^{n} R\,(x + g_i) \quad (8)$$

here, $g_i \sim N(0, \sigma^2)$ are noise vectors sampled from the Gaussian distribution. The most appropriate noise level depends on the input image and the network. The authors suggest noise level of 10% - 20%, which means $\frac{\sigma}{x_{max} - x_{min}}$ should be between 0.1 and 0.2, as this level balances sharpness and structure of the image. The limits x_{min} and x_{max} refer to minimum and maximum pixel values of the image. The other parameter is the number of samples, denoted as n, for which it is suggested to use n=50, since there are diminishing returns above that.

3.3 Evaluation of XAI methods

Given that we have the output of a model, which is likely a prediction in a classification task, our primary concern is to understand the rationale before the model's decision. However, in addition to this, we also require an evaluation metric to assess the quality of this explanation. Practitioners need to know whether they can trust the explanation that may be returned, as there are cases where explanations may lead to misinterpretation. As mentioned earlier, there is a challenge with XAI methods, as there is no ground truth available for these explanations. In addition, the favourable evaluation metric may vary lot according to the specific evaluation goal and oriented user groups [36]. Consequently, several suggestions have been put forth to enable the evaluation of XAI methods. Some of these suggestions will be described in the following paragraphs.

According to Gabriëlle Ras et al. Guide [36], there are two main approaches for evaluating explanations. The first is to devise an objective metric or benchmark to evaluate the explanations without human intervention. This approach has the benefit of being able to compare numerous explanations with each other. Given that visualization methods that produce heatmaps are a popular intuitive type of explanation method, it has gained most of the attention in the subfield of evaluation explanation. The second approach involves involving a human in the assessment of the explanation. By using professionals, such as radiologists for

a medical image analysis, to evaluate an explanation, we can investigate to what extent the explanation aligns with the human's potential decision.

3.3.1 Evaluating heatmaps based on human experts.

One way to assess explanations using a human baseline is to examine the degree of alignment between model-generated explanations and those provided by humans. One approach could be to let domain experts test the explanation based on their expertise. An alternative method involves comparing a radiologist's manually annotated tumor with the size and location of the region highlighted by the explainability method using a specific metric. However, this can be challenging in the context of medical image analysis, as obtaining manual annotations is a resource-intensive process. This approach offers the advantage of determining the interpretability and utility of the specific setting and explanation method for individuals who will be using them.

3.3.2 Evaluating heatmaps based on experiments.

A common approach is to introduce perturbation-based method for evaluating the quality of heatmaps. Using this method, one can replace the region of the image that corresponds to the highlighted area by the heatmap with randomly uniform data and check how much the classification score changes. According to this metric, the more the classification score changes the better the heatmap corresponds to class-discriminative features. This is a method that is represented in numerous papers as evaluating XAI methods [37], [38].

However, Hook et al. [39] argues that the perturbation-based method violates the assumption that the training and evaluation data come from the same distribution. In response, they propose a benchmark for evaluating feature importance estimates in DNNs. Their benchmark is called ROAR: RemOve and Retrain. The goal of ROAR is to determine whether the removal of important information caused classification degradation or whether the introduction of the so-called uninformative information caused the modified images to go out of distribution, thereby causing classification degradation. It replaces the fraction of pixels deemed important according to some heatmap with the channel mean, like perturbation-based methods. The difference with their method is that they

remove the same percentage of important pixels both in the training and test data with the channel mean of the image. Finally, the train separate models on the modified data and evaluates the classification accuracy. If the accuracy of the re-trained model goes down, we can conclude more safely that the removed information caused the classification degradation.

Methods like [40] investigate the reliability of heatmaps by modifying the input with information that does not change the classification result and checking how the heatmaps change as a result. They find that various visualization methods are vulnerable to input modification and return incorrect heatmaps as a result. The main conclusion is that many visualization methods are unreliable because they do not satisfy input invariance.

In contrast to the previous methods, Vu et al. [41] suggests a metric to evaluate heatmaps based on perturbing regions that are not indicate as important by the xai method. Their metric is called c-Eval, where c is a number that indicates how robust the classifier is to perturbations in regions indicated as important. This method indirectly measures how accurate the heatmaps are: the larger the c, the more robust the classifier, the more accurate the explanation method is at identifying class-discriminating features. Using c-Eval they compare various explanation methods and find that there is a significant difference in the quality of heatmaps produced by black-box models (e.g. SHAP, LIME) compared to back-propagation based methods (e.g., LRP).

Adebayo et al. [42] proposes two sanity checks for evaluating the quality of heatmaps. The first is the model parameter randomization test, and it compares heatmaps generated by a trained model with heatmaps generated by a randomly initialized model. If the output is similar, the explanation method is insensitive to model properties such as the weights. The second sanity check is the data randomization test, and it compares heatmaps generated by a model trained on the original dataset with heatmaps generated by a model trained on a version of the dataset where all the labels have been randomly permuted. If the heatmaps are similar, it indicates that the explanation method is not dependent on the relationship between the data and the labels that exist in the original data. The distance between the heatmaps is measured using various similarity metrics.

3.4 Evaluation in this research

In this research, we adopted two primary approaches to evaluate our results. The first approach involves comparing the resulting heatmaps with the annotated images of our dataset and utilizing the Intersection Over Union (IoU) metric. The second approach entails replacing a varying percentage of the most important pixels and examine the classification scores for each class.

3.4.1 Intersection over Union (IoU)

Intersection over Union is a common evaluation metric in the field of computer vision and image segmentation. It is useful for assessing the performance of object detection and image segmentation algorithms. It expresses the degree of overlap between a ground truth box and a predicted box. IoU is calculated as the ratio of the intersection area to the union area, where intersection is the region where the predicted object and the actual object coincide, and the union is the combined area of the predicted and the actual object. It is calculated as:

$$\text{IoU} = \frac{\text{Intersection Area}}{\text{Union Area}} = \frac{|\text{A} \cap \text{B}|}{|\text{A} \cup \text{B}|} \quad (8)$$

where A and B are the prediction and the ground truth area respectively. The IoU value typically ranges from 0 to 1, where:

- IoU = 0 means no overlap between the predicted and ground truth regions.
- IoU = 1 indicates a perfect match, where the predicted region precisely matches the ground truth region.

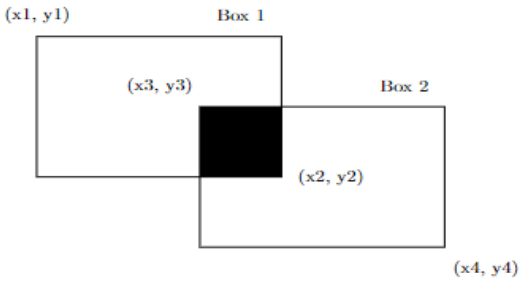


Figure 17: Intersection of two boxes.

The steps to calculate IoU are:

1. Calculate the top left corner of the intersection, we compare the top left corners of each of the boxes.

   X_intersection_left = max(x1, x3)

   Y_intersection_left = max(y1, y3)

   Calculate the bottom right corner of the intersection, we compare the bottom right corners of each of the boxes.

   X_intersection_right = max(x2, x4)

   Y_intersection_right = max(y2, y4)

2. Check if there is an actual intersection by ensuring that x_intersection_right and y_intersection_right is greater than zero. If either of them is less than or equal to zero, there is no intersection.
3. Calculate the area of the intersection:

   Width_inter = (x_intersection_right - x_intersection_left)

   Height_inter = (y_intersection_right - y_intersection_left)

   Area_intersection = width_inter x height_inter

4. Calculate the area of the bounding box:

   Area_box1 = (x2-x1) x (y2-y1)

   Area_box2 = (x4-x3) x (y4-y3)

5. Calculate the area of the union:

   Area_union = area_box1 + area_box2

6. Calculate IoU:

   IoU = area_intersection / area_union

Given that our dataset contains annotated ground truth boxes for the different pathologies each delineated by different radiologists, we decided to evaluate our results by generating bounding boxes with varying proportions of the most significant pixels found in the resulting heatmaps. We then employed Intersection over Union (IoU) metric to measure the degree of alignment between these newly created bounding boxes and the ground truth boxes only for the classes that are predicted correctly from our model.

Since the ground truth boxes originate from multiple radiologists and thus exhibit variation, we perform comparisons between our bounding boxes and all the radiologists' boxes for each pathology class. Subsequently, we calculate the mean IoU across these comparisons to provide comprehensive evaluation.

3.4.2 Pixel Importance Analysis

In the second approach used for evaluation, by using the resulted heatmaps by Grad-Cam implementation, we identify crucial pixels influencing predictions across the different classes. The identified pixel important thresholds include 0.02, 0.05, 0.1, 0.2 and 0.5. Subsequently, we conduct an extraction of these pivotal pixels, replacing them either with the mean values across all three channels or through Gaussian blurring.

The transformed images are then subjected to our pretrained model, and predictions are analysed for each specific class. A comparative examination is performed between the modified predictions and the original model predictions on the unaltered images. This methodology aims to enhance the interpretability of the model by elucidating the impact of the individual pixels on predictions and understanding how image transformations affect the classification outcomes.

4. MACHINE LEARNING PIPELINE FOR XAI METHODS IN CHEST X-RAY CLASSIFICATION TASK

4.1 Dataset

As mentioned before, the dataset used in this research is the VinDr-CXR dataset created by Vingroup Big Data Institute (VinBigData, [8]) with more than 100,000 images in DICOM format that were retrospectively collected from the Hospital 108 (H108) and Hanoi Medical University Hospital (HMUH), two of the largest hospitals in Vietnam. The published dataset consists of 18,000 postero-anterior (PA) view CXR scans that come with both the localization of critical findings and the classification of common thoracic diseases. The images were annotated by a group of 17 radiologists with at least 8 years of experience for the presence of 22 critical findings (local labels) and 6 diagnoses (global labels); each finding is localized with a bounding box. The local and global labels correspond to the “Findings” and “Impressions” sections, respectively, of a standard radiology report. Subsequently they divide the dataset into two parts: the training set of 15,000 scans and the test set of 3,000 scans. Each image on the training set was independently labeled by 3 radiologists. The labeling process was performed via an in-house system called VinDrLab, which was built in top of a Picture Archiving and Communication System (PACs).

In this research, a slightly modified version of this dataset is used published in Kaggle platform (https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection/). We use only the training set of the dataset which we split into three parts: training set, validation set, and test set with a ratio of 0.8:0.1:0.1 respectively. All the images are transformed to jpg format and resized to 512x512 while the boxes coordinates were rescaled to this size.

As can be seen in **Error! Reference source not found.**, there is an imbalance of the dataset, where the healthy images are much more than the images annotated with a thoracic disease. For this reason, six classes were finally chosen for the classification problem which are Aortic Enlargement, Cardiomegaly, Pleural Effusion (PE), Lung Opacity (LO), Pleural Thickening (PT), Pulmonary Fibrosis (PF) and healthy images labelled “No Finding”.

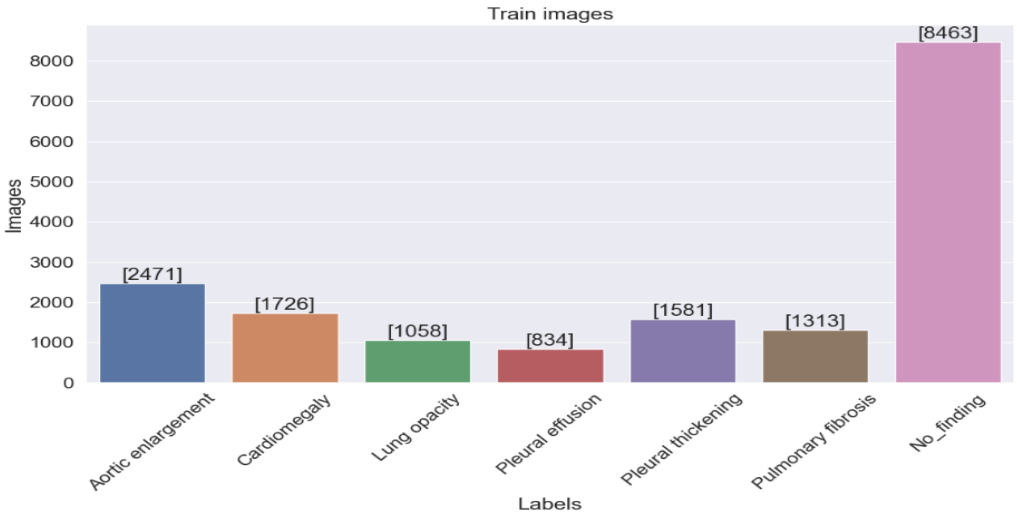
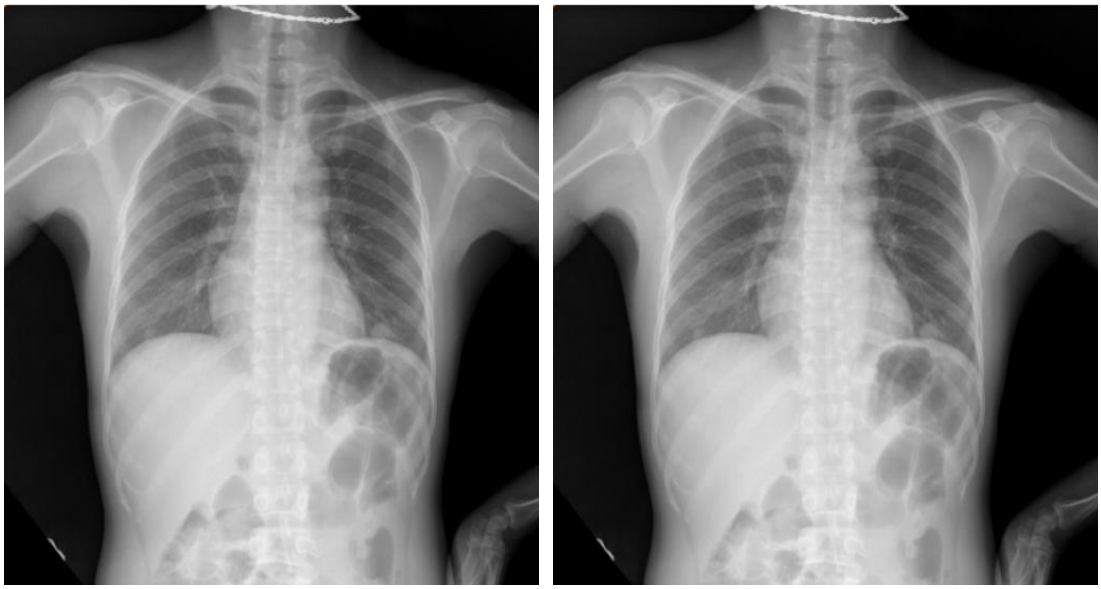


Figure 18: Distribution of the classes of the train dataset.



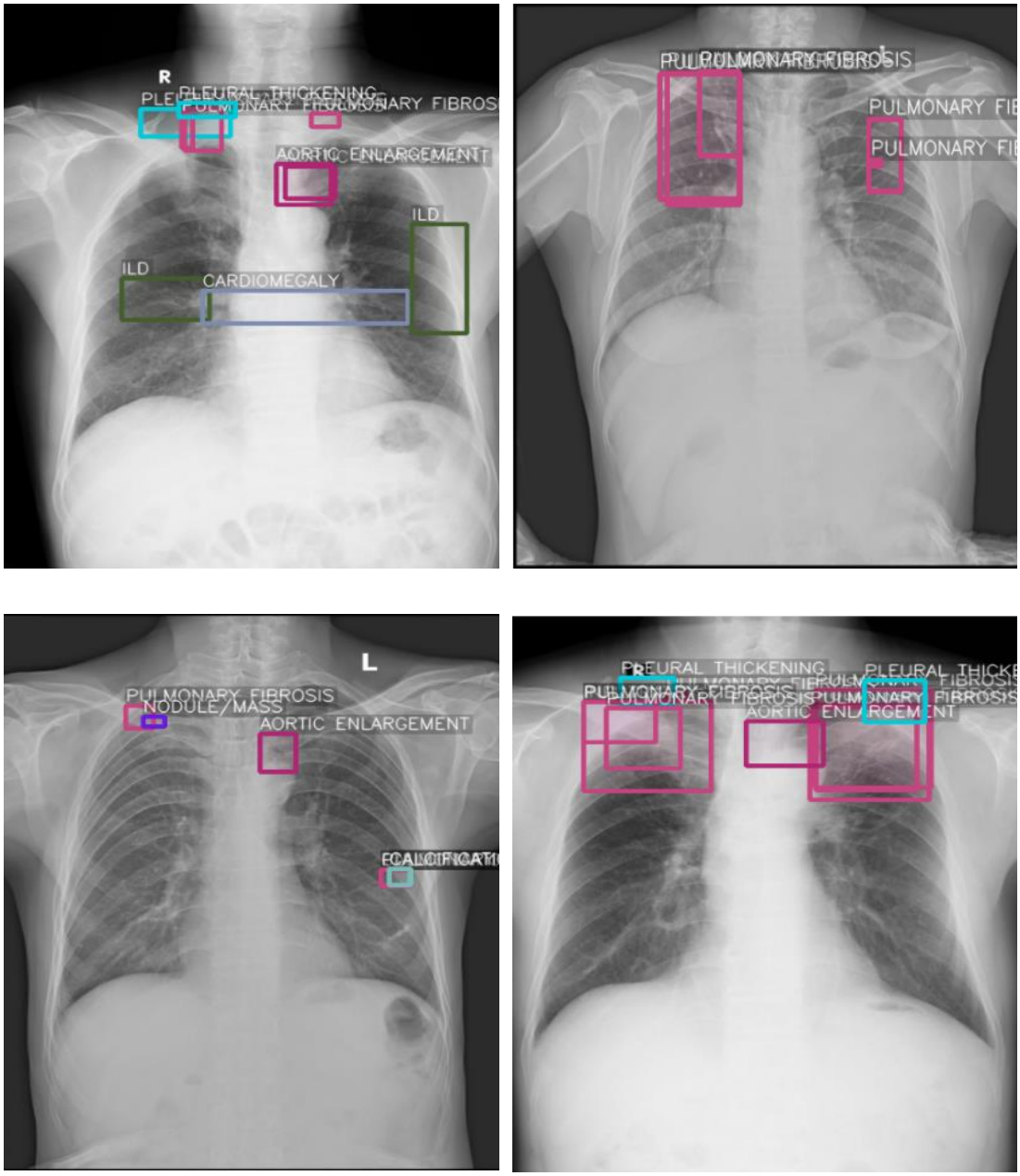


Figure 19: Examples of chest X-rays with their corresponding labels.

Table 2: Final Dataset

Train	11,836
Valid	1.480
Test	1,480

Aortic Enlargement	3067
Cardiomegaly	2168
Lung Opacity	1322
Pleural Effusion	1032
Pleural Thickening	1981
Pulmonary Fibrosis	1617
No Finding	10606

4.2 Technical Characteristics

This master thesis employs the PyTorch framework for training neural network models. PyTorch is chosen for its flexibility and ease of use in constructing and training models. The experimentation is carried out in Jupyter Notebooks, providing an interactive environment for prototyping and analysis. The implementation is done in Python, known for its simplicity and readability. To handle the computational demands, an NVIDIA GeForce GTX 1660 Ti with Max-Q Design GPU is utilized, accelerating the training process, and optimizing model performance. For the execution of the experiments some more useful and popular libraries are used such as Numpy, Matplotlib and OpenCV.

4.3 Model

The dataset preprocessing pipeline in this study incorporates a series of transformations aimed at enhancing the diversity and quality of the input images. Initially, the images are converted to PIL format using ‘transforms.toPILImage()’. Subsequently, a resizing operation is applied to

standardize the dimensions to (224,224) through ‘transforms.Resize((224,224)). To introduce variability and aid in model generation, a random horizontal flip with a probability of 0.5 is employed via transforms.RandomhHorizontalFlip(p=0.5)’. Additionally, random rotations up to 45 degrees are applied using ‘transforms.RandomRotation(degrees=45)’. The transformed images are then converted to tensors with ‘transforms.ToTensor()’. Finally, the pixel values are normalize using the mean and standard deviation of the ImageNet dataset, aligning with the pre-training statistics commonly used in deep learning models (‘transforms.Normalize(mean=[0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225])’). This comprehensive set of transformations contributes to the augmentation and standardization of the input data, crucial for the training robust and effective neural network models.

In this study, the model employed for the training process is ResNet50 architecture, initially trained on the ImageNet dataset. To adapt the model for the specific classification problem at hand, the final dense layer is replaced with an ‘nn.Linear(num_ftrs, num_classes)’ layer, where num_classes corresponds to the number of classes involved in the classification task. The optimization is carried out using Stochastic Gradient Descent (SGD). Given the nature of the classification task as a multilabel problem, the chosen loss function is Sigmoid (‘nn.BCEWithLogitsLoss(pos_weight=pos_weight)’). This loss function treats each class as a binary task, providing probabilities for the presence or absence of a specific disease in an image. To address the dataset’s class imbalance, with a significantly larger number of healthy images compared to diseased ones, a ‘pos_weight’ is introduced. The ‘pos_weight’ is calculated as the ratio of negative counts to positive counts for each class (‘pos_weight = num_negatives/num_positives’). This approach aims to mitigate the model’s bias toward predicting the majority class, ensuring a more balanced and accurate learning process. The learning rate of the training process starts at 0.01 and we utilize an LR Scheduler through ‘lr_scheduler.LinearLR(optimizer, start_factor=1.0, end_factor = 0.3 and total_iters = 20)’. We train the model for 20 epochs, and we save the model in the minimum loss in the validation loss

as can be seen in the Figure 20. We use a batch size of 32 based on the available hardware and the size of the dataset.

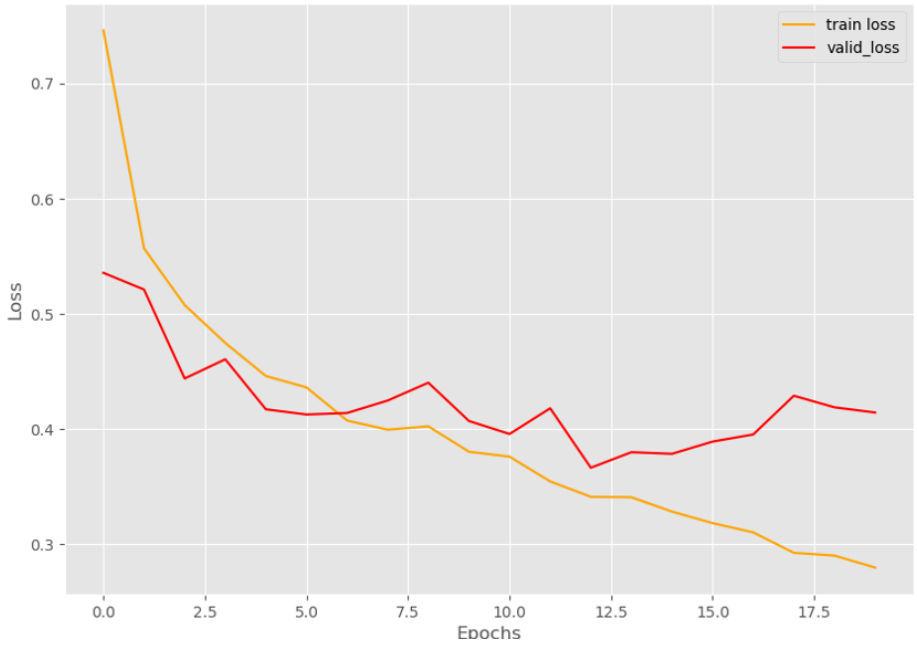


Figure 20: Training loss and validation loss for 7 classes.

4.4 Results

We conducted a series of experiments to assess the classification scores for each class in our study. The model was trained across different scenarios, including configurations for 3 classes (Aortic Enlargement, Cardiomegaly, No-finding), 5 classes (Aortic Enlargement, Cardiomegaly, Pleural Thickening and No-finding), 7 classes (Aortic Enalrgement, Cardiomegaly, Lung Opacity, Pleural Effusion, Pleural Thickening, Pulmonary Fibrosis and No-finding). These configurations were chosen based on the distribution of diseases images. Additionally, we conducted training on the entire dataset.

The subsequent Table 3 presents the results of these models on the test dataset. Following an analysis of the outcomes, we opted to implement Grad-Cam specifically for the 7-class scenario due to achieving a satisfactory micro F1 score.

Table 3: F1 Scores for different experiments

F1-score: 3 classes	
Label	F1-score
Aortic Enlargement	0.89
Cardiomegaly	0.85
No-Finding	0.95
Micro avg	**0.90**

F1 score: 5 classes	
Label	F1-score
Aortic Enlargement	0.81
Cardiomegaly	0.82
Pleural Thickening	0.58
No Finding	0.91
Micro avg	**0.76**

F1-score: 7 classes	
Label	F1-score
Aortic Enlargement	0.86
Cardiomegaly	0.83
Lung Opacity	0.55
Pleural Effusion	0.69
Pleural Thickening	0.65
Pulmonary Fibrosis	0.57
No Finding	0.96
Micro avg	**0.81**

F1-score: whole dataset	
Label	F1-Score
Aortic Enlargement	0.81
Atelectasis	0.20
Calcification	0.17
Cardiomegaly	0.80
Consolidation	0.39
ILD	0.24
Infiltration	0.31
Lung Opacity	0.47
Nodule/Mass	0.32
Other Lesion	0.38
Pleural Effusion	0.48
Pleural Thickening	0.57
Pneumothorax	0.09
Pulmonary Fibrosis	0.51
No Finding	0.96
Micro avg	**0.60**

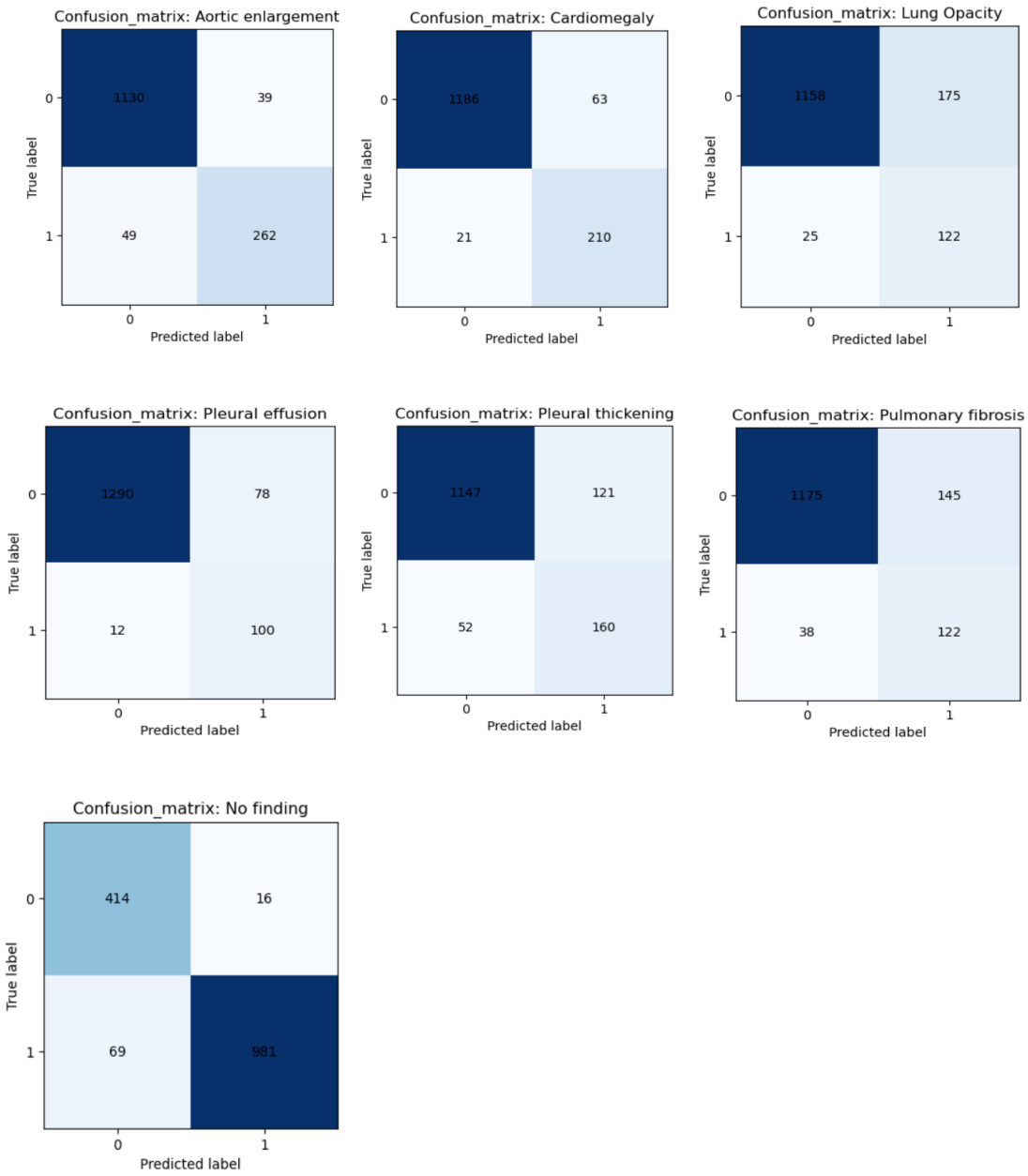


Figure 21: Confusion matrices for each label, configuration of 7 classes.

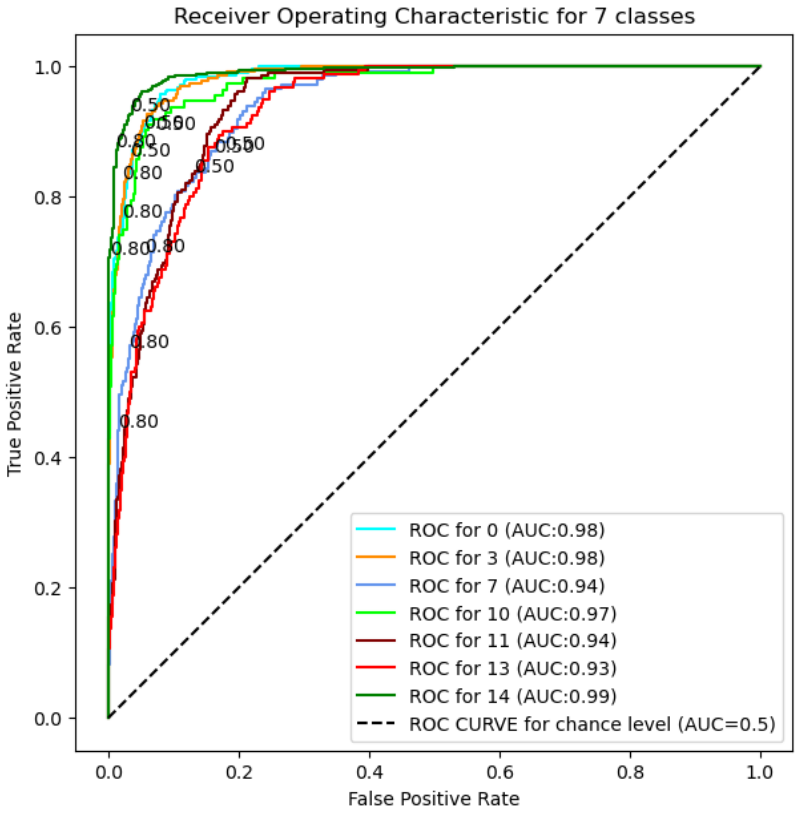


Figure 22: ROC Curves. 0: 'Aortic Enlargement', 3: 'Cardiomegaly', 7: 'Lung Opacity', 10: 'Pleural Effusion', 11: 'Pleural Thickening', 13: 'Plumonary Fibrosis', 14: 'No-Finding'.

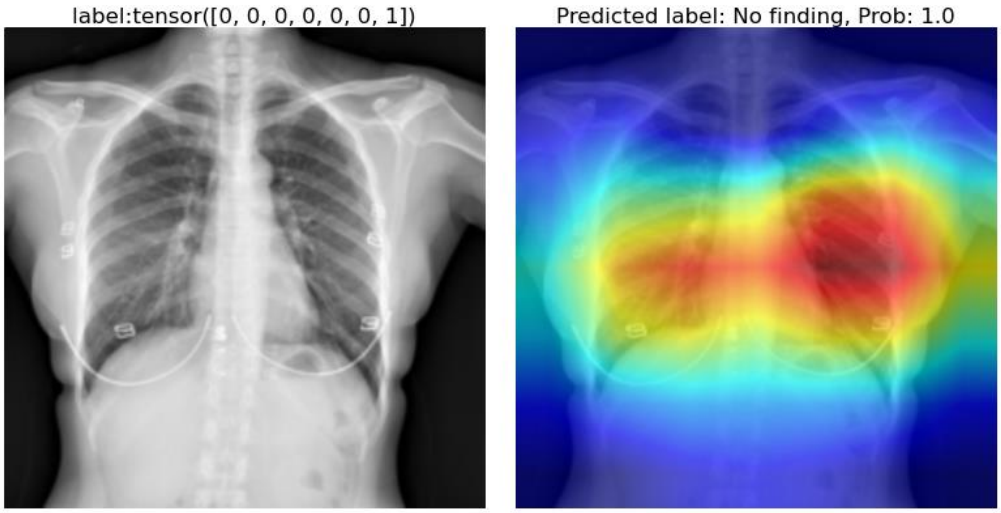
4.5 Grad-Cam Results

In this study, we employed Grad-Cam as the XAI method, particularly pertinent in medical domain. The methodology involved partitioning the ResNet50 model into two segments, delineated by the last convolutional layer generating activation maps of size 7x7. To implement Grad-Cam, we conducted a forward pass of input images through the network, subsequently computing gradients of the output concerning the chosen convolutional layer. Given the multilabel nature of our task, outputs with a probability of 0.7 or higher were selected for Grad-Cam analysis. For each selected output, we performed Grad-Cam, followed by computing the average pooling of gradients to derive importance weights for each channel. The final step involved calculating the weighted sum of activation maps using the obtained importance weights.

To visualize the resultant heatmaps, normalization was executed by dividing with the maximum value, and the heatmaps were upscaled to match the original dimensions of the image. This comprehensive process facilitated the

production of interpretable and visually accessible heatmaps, contributing to a deeper understanding of model decisions in our medical classification task. In the following examples, images are represented with annotated bounding boxes delineating the specific regions indicative of the disease. Additionally, the corresponding heatmaps are provided, emphasizing the salient areas within the images that significantly contribute to the classification.

At first, images belonging to the “No-finding’ class are presented along with the corresponding heatmaps, revealing the regions where the model focuses its attention to make a decision. The heatmaps consistently indicate attention in both lungs or just below the letter indicating the side (R: Right or L: Left) of the chest X-Ray. This trend is observed across the majority of heatmaps generated by the model. It’s noteworthy that ‘No-Finding’ is the predominant class with the highest F1 score of 0.96.



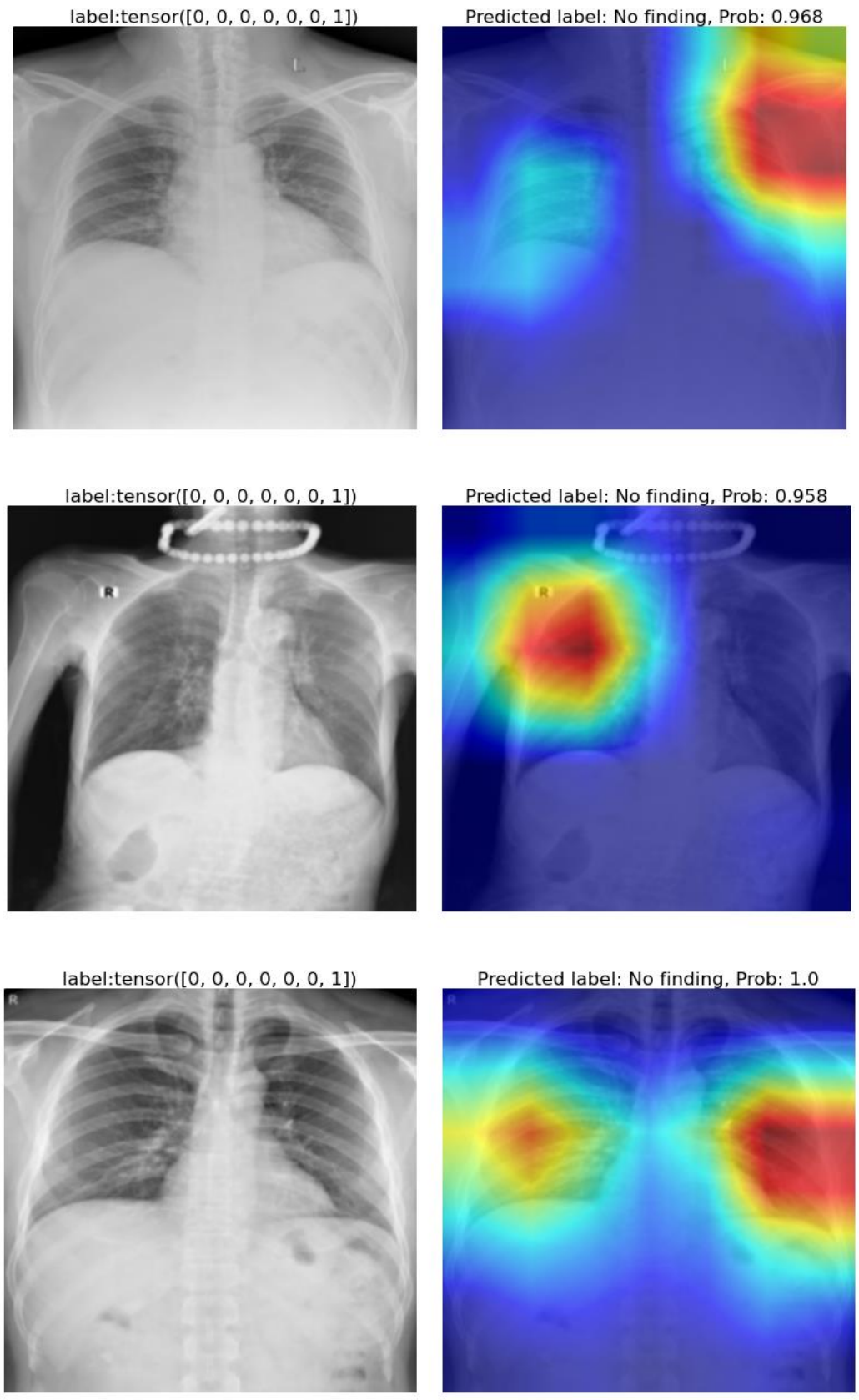
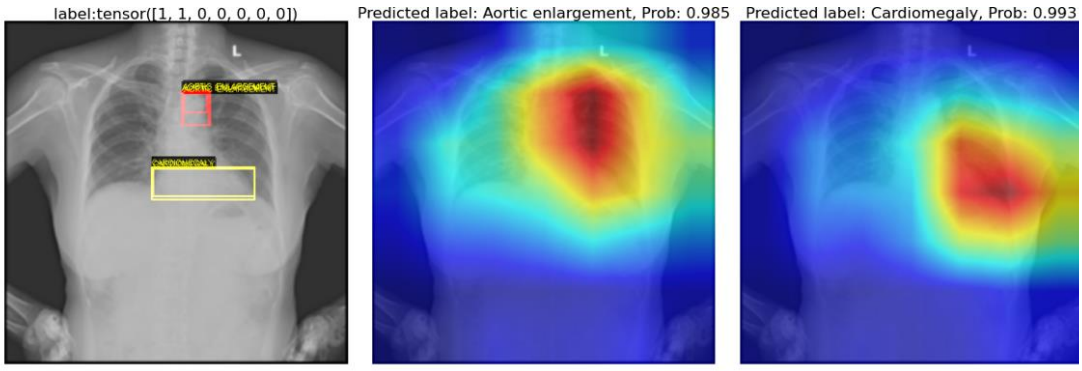


Figure 23: Images of healthy chest X-rays along with the corresponding heatmaps and the probabilities for the 'No-Finding' class displayed above each image.

Subsequently, images of the classes ‘Aortic Enlargement’ and ‘Cardiomegaly’ are presented together due to their frequent co-occurrence in a distinct region of chest X-rays. Unlike other diseases, these conditions are consistently localized within the chest X-ray, making their identification more standardized. They both exhibit high F1 scores of 0.86 and 0.83 respectively. From the heatmaps in Figure 24, we can observe that the model generally succeeds in localizing these diseases to a considerable extent. However, as shown in Figure 25, there are instances that the model exhibits unexpected behavior and localize these diseases inaccurately. Despite occasional misestimations, the model tends to consistently identify and localize these diseases within the specified region in most of the cases.



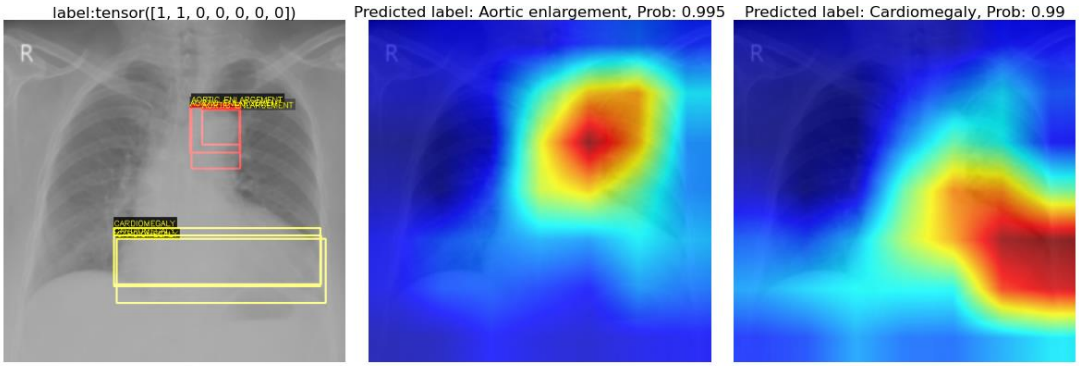


Figure 24: Images both annotated with the classes 'Aortic Enlargement' and 'Cardiomegaly' along with the corresponding heatmaps and the probabilities of each class displayed above each heatmap. 'Aortic Enlargement is annotated with red color and 'Cardiomegaly' is annotated with yellow.

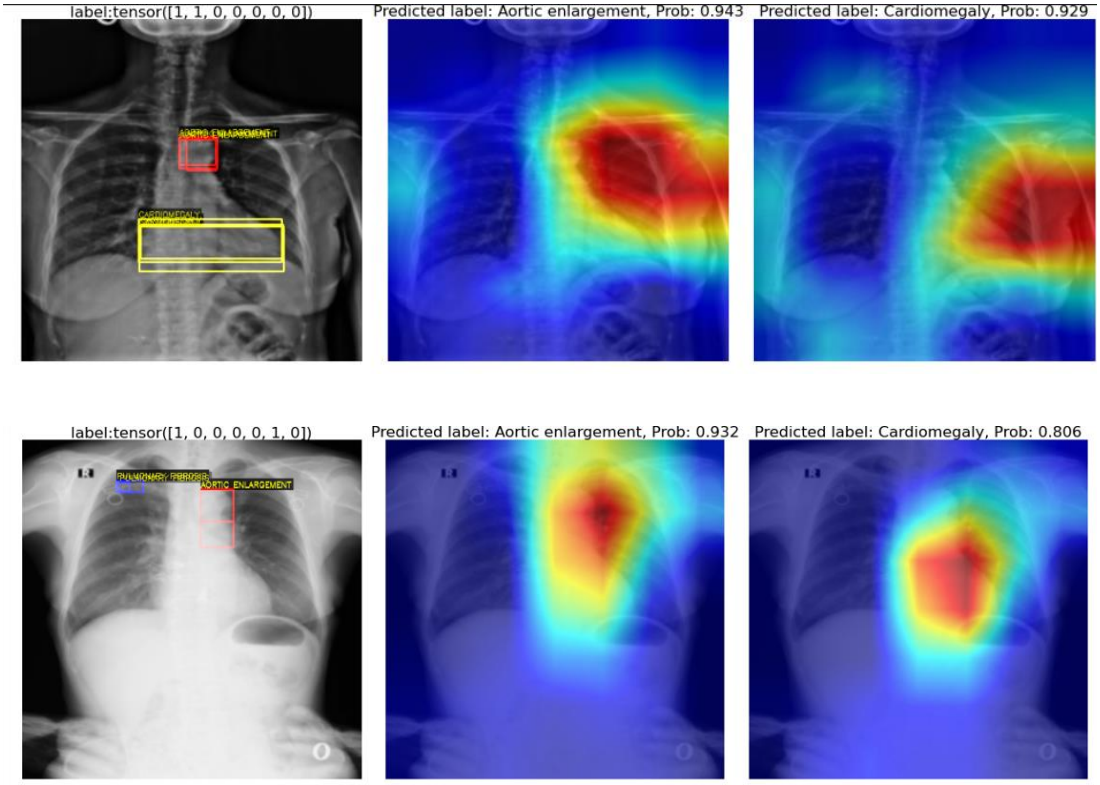


Figure 25: top – the model has correctly classified the two diseases 'Aortic Enlargement' and 'Cardiomegaly', but it is observed an unexpected behavior regarding the localization of the diseases. Bottom – Although the model misclassifies 'Cardiomegaly' class, it still localizes it in a region of interest for this disease.

In the provided set of Figure 26, several noteworthy characteristics can be observed for the rest of the classes. Figure (26a) exemplifies a typical instance of confusion, especially apparent in the presence of nearly all classes. The

model exhibits challenge in localizing ‘Lung Opacity’, achieves partial localization of ‘Pleural Effusion’, and demonstrates a typical behaviour for the ‘Pleural Thickening’ class, capturing a broad range on the pleural side. Surprisingly, it achieves a relatively accurate localization of ‘Pulmonary Fibrosis’.

In Figure (26b), the model correctly classifies the classes ‘Lung Opacity’, ‘Pleural Effusion and ‘Pleural Thickening’, although it fails in completely localizing ‘Lung Opacity’ class. Additionally, its complete mislocalization of the ‘Pleural Thickening’ indicates a limitation in its ability to precisely pinpoint the affected region, which is a common characteristic of this class.

Figure (26c) showcases the model’s ability to identify the general area of pathologies in the chest X-ray, yet it struggles to distinguish between them, resulting in the misclassification of ‘Lung Opacity’ - a common challenge in its prediction.

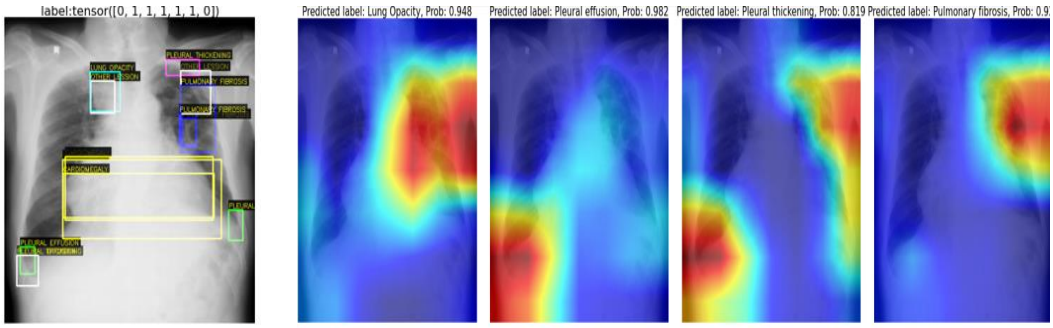
Figure (26d) reveals the model’s capability to identify an issue on the right side of the chest X-ray but fails to predict all the diseases. This suggests a limitation in disease recognition despite the localization ability.

In figure (26e), the model successfully predicts and localizes ‘Lung Opacity’ and ‘Pleural Effusion’ with high probabilities (0.994 and 0.998 respectively). However, it misclassifies ‘Pulmonary Fibrosis’ and localizes it in the general area of abnormalities.

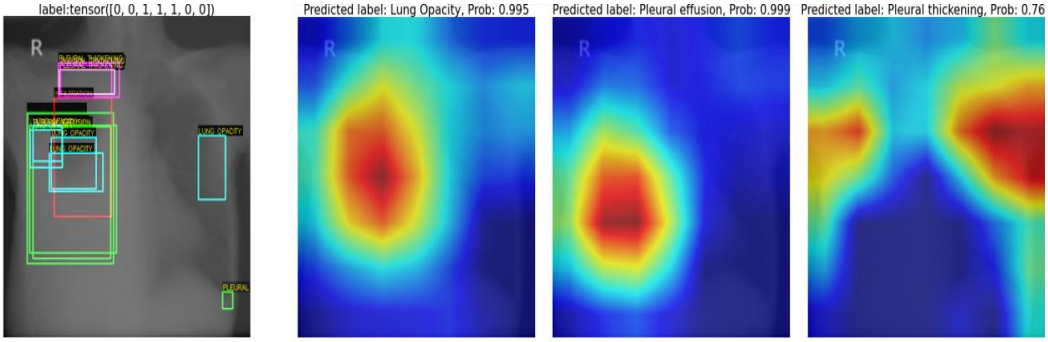
Figure (26f) demonstrates another instance where the model seems to comprehend the presence and location of pathologies, successfully localizing them. Notably, even when ‘Cardiomegaly’ is misclassified, the model correctly identifies the expected region for examination.

It is crucial to acknowledge that the annotated pathologies result from diverse radiologists, leading to variations in labelling and bounding boxes annotations. Instances, where one radiologist annotates a disease while another does not underscore the inherent difficulty and subjectively in analyzing chest X-ray images, emphasizing the complex of drawing definitive conclusions.

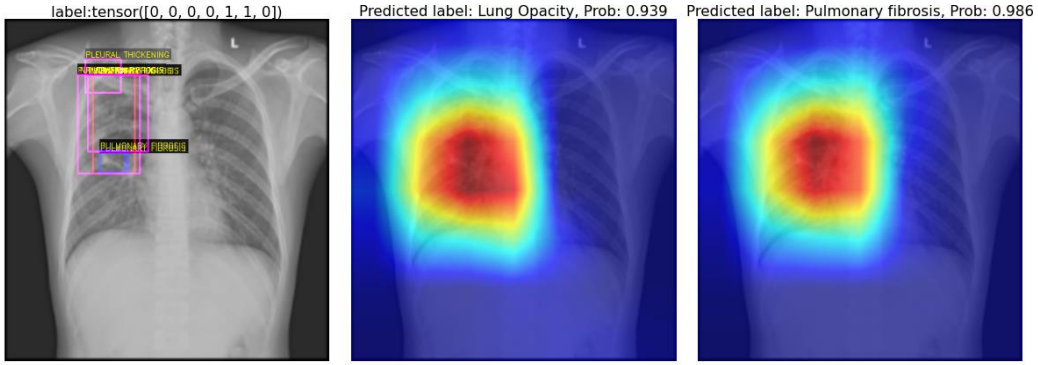
a


label:tensor([0, 1, 1, 1, 1, 1, 0])
Predicted label: Lung Opacity, Prob: 0.948
Predicted label: Pleural effusion, Prob: 0.982
Predicted label: Pleural thickening, Prob: 0.819
Predicted label: Pulmonary fibrosis, Prob: 0.9

b


label:tensor([0, 0, 1, 1, 1, 0, 0])
Predicted label: Lung Opacity, Prob: 0.995
Predicted label: Pleural effusion, Prob: 0.999
Predicted label: Pleural thickening, Prob: 0.76

c


label:tensor([0, 0, 0, 0, 1, 1, 0])
Predicted label: Lung Opacity, Prob: 0.939
Predicted label: Pulmonary fibrosis, Prob: 0.986

d


label:tensor([0, 0, 1, 1, 1, 1, 0])
Predicted label: Lung Opacity, Prob: 0.83
Predicted label: Pleural effusion, Prob: 0.725
DA DIEU TRI LAO
DA DIEU TRI LAO
DA DIEU TRI LAO

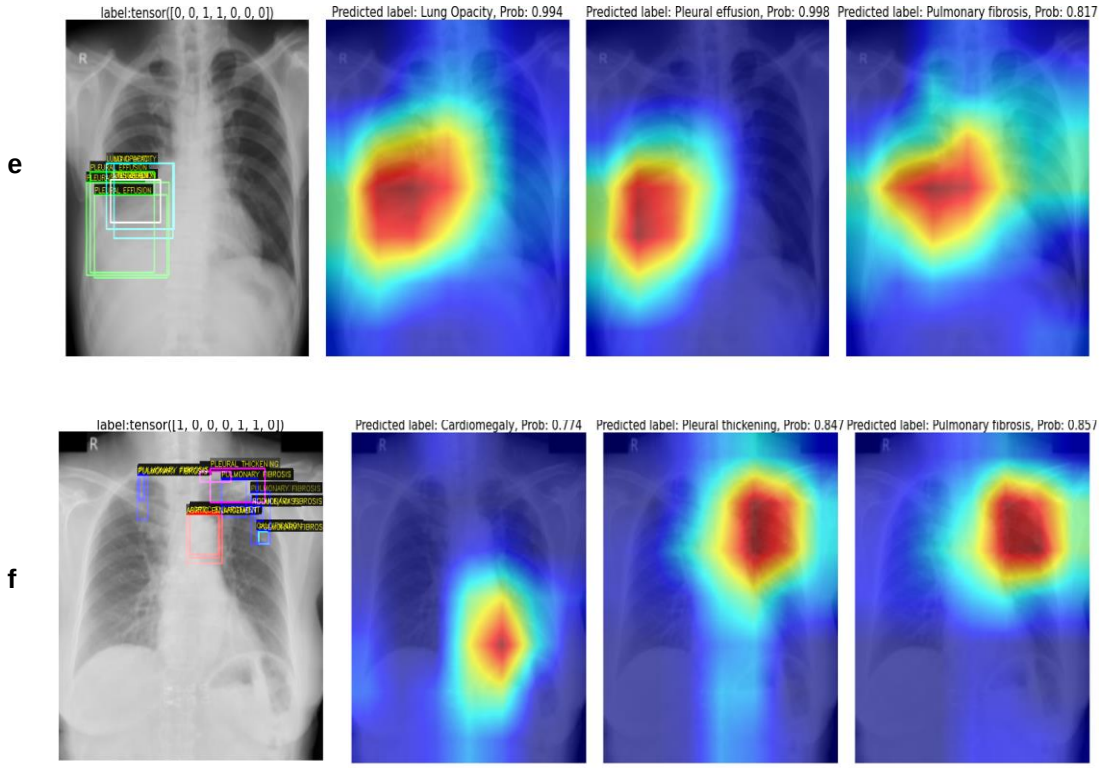


Figure 26: Examples of the abnormalities 'Lung Opacity', 'Pleural Effusion', 'Pleural Thickening' and 'Pulmonary Fibrosis' with the corresponding heatmaps.

4.6 Smooth Grad

Subsequently, in our research we implement Smooth Grad, a method that introduces perturbations to the input data and observes the model's sensitivity to these variations, providing additional information beyond what may be captured by methods solely relying on gradients. This is achieved by adding gaussian noise to the original image creating a number of samples. Subsequently we put these samples as an input to our model and implement Grad-Cam to take a heatmap for each sample. Finally, we average over the samples to take the final heatmap. The suitable noise proposed by (24) are between 0.1 and 0.2 to balance the sharpness of the heatmap and maintain the original structure of the image. The number of samples proposed by the authors, after several experiments is 50 samples. In the following set of figures, some examples of SmoothGrad are provided.

			Noise 0.1	Noise 0.2
Original Image	GradCam - Aortic enlargement	SmoothGrad - Aortic enlargement	SmoothGrad - Aortic enlargement	
Original Image	GradCam - Cardiomegaly	SmoothGrad - Cardiomegaly	SmoothGrad - Cardiomegaly	
Original Image	GradCam - Lung Opacity	SmoothGrad - Lung Opacity	SmoothGrad - Lung Opacity	

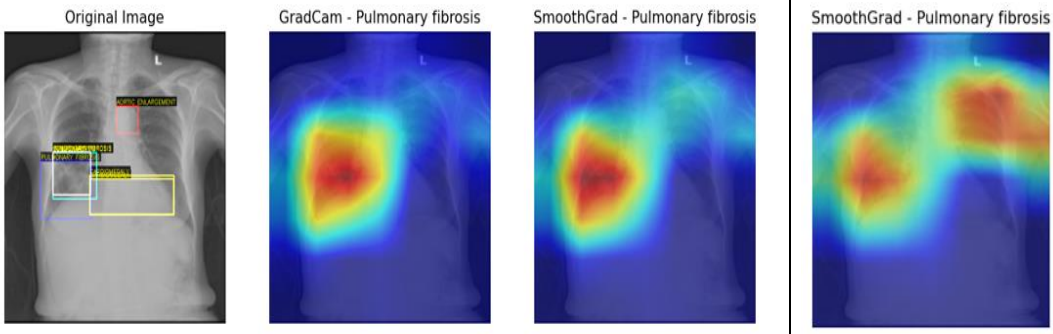


Figure 27: Implementation of Grad-Cam and SmoothGrad in the same chest X-ray both with 0.1 and 0.2 noise.

We can conclude in some general results from the implementation of both Grad-Cam and SmoothGrad with the help of the above set of figures and the different chest X-rays that we examined. In instances of 'Aortic Enlargement' and 'Cardiomegaly', both Grad-Cam and SmoothGrad consistently produces sharper heatmaps with intensified pixel highlighting. This behaviour indicates the robustness of specific pixels, which remain invariant to noise. Differences between Grad-Cam and SmoothGrad, appear in some Cardiomegaly instances and may signify areas of uncertainty or lower model confidence.

Regarding 'Lung Opacity' and 'Pulmonary Fibrosis', most cases revealed consistent behaviour between Grad-Cam and SmootGrad in 0.1 level of noise. Notably, by scenarios of increased noise, more pixels are highlighted in different regions. These pixels sometimes showcase improved localization and sometimes they give a more confused heatmap.

Challenges persists in the interpretation of 'Pleural Thickening', with Grad-Cam generating noisy maps capturing a broad region of the chest X-ray. SmoothGrad mitigates noise to some extent, resulting in clearer heatmaps, however precise localization remains elusive for this class.

In general, SmoothGrad consistently highlights specific pixels across multiple perturbed versions of an input for 'Aortic Enlargement' and 'Cardiomegaly', emphasizing the robustness of these regions in contributing to the model's

decision-making. These two classes exhibit well-localized heatmaps, while behaviours vary for the other classes, in agreement with Grad-Cam results.

5. EVALUATION RESULTS

5.1 Intersection over Union (IoU)

In the first approach to heatmap evaluation, we utilize the Intersection over Union (IoU) metric to gauge the alignment between the true bounding boxes and those deduced by the heatmaps. The process of extracting bounding boxes involves distinguishing critical pixels through the creation of binary masks from the heatmaps. This is achieved by applying varied thresholds percentages (0.01, 0.02, 0.03, 0.05, 0.08, 0.1 and 0.2).

Table 4: Part of code for the extraction of the predicted bounding boxes.

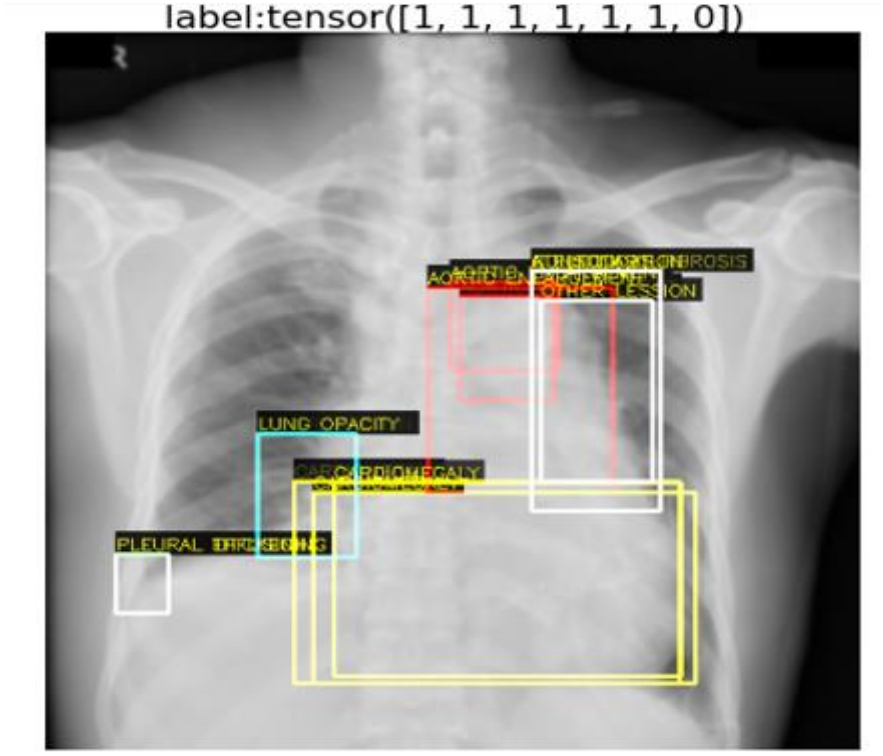
```
def create_bounding_boxes(self, mask):
    mask = mask.detach().cpu().numpy().astype('uint8') # 224, 224, max: 255, min: 0
    mask = cv2.resize(mask, (512,512)) # shape: 512,512
    contours,_ = cv2.findContours(mask, cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE) # list of points defining the contour's shape
    for contour in contours:
        # Find the bounding box
        x,y,w,h = cv2.boundingRect(contour)
        # Draw the bounding box in an image for visualization
```

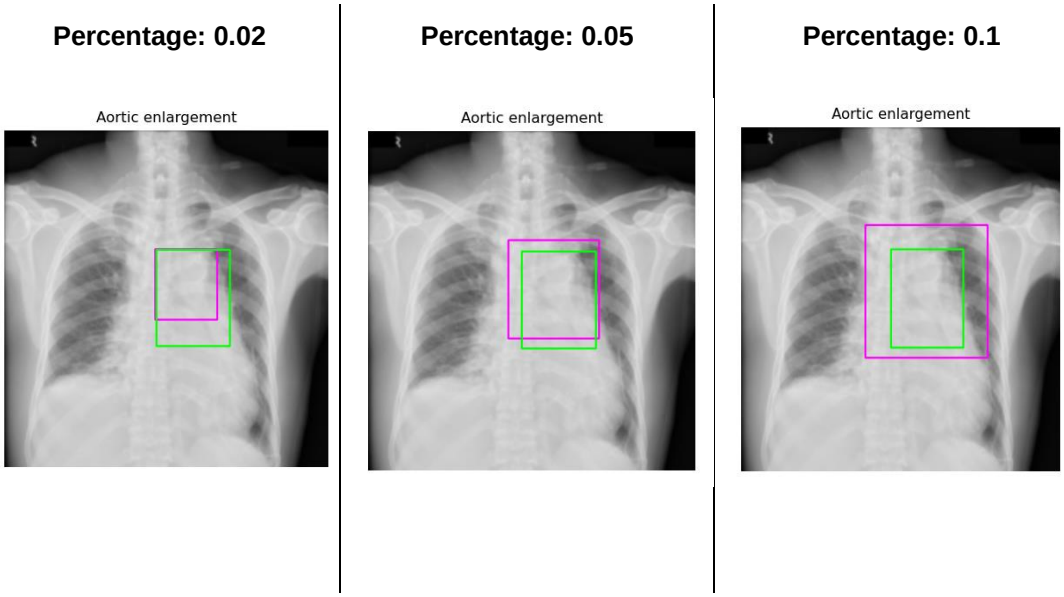
The binary masks are created by isolating pixels based on their significance and subsequently, contours are extracted from these binary masks using the ‘RETR_EXTERNAL’ mode provided by cv2 library. This mode selectively retains only the outermost contours, ensuring that nested contours not overshadow each other.

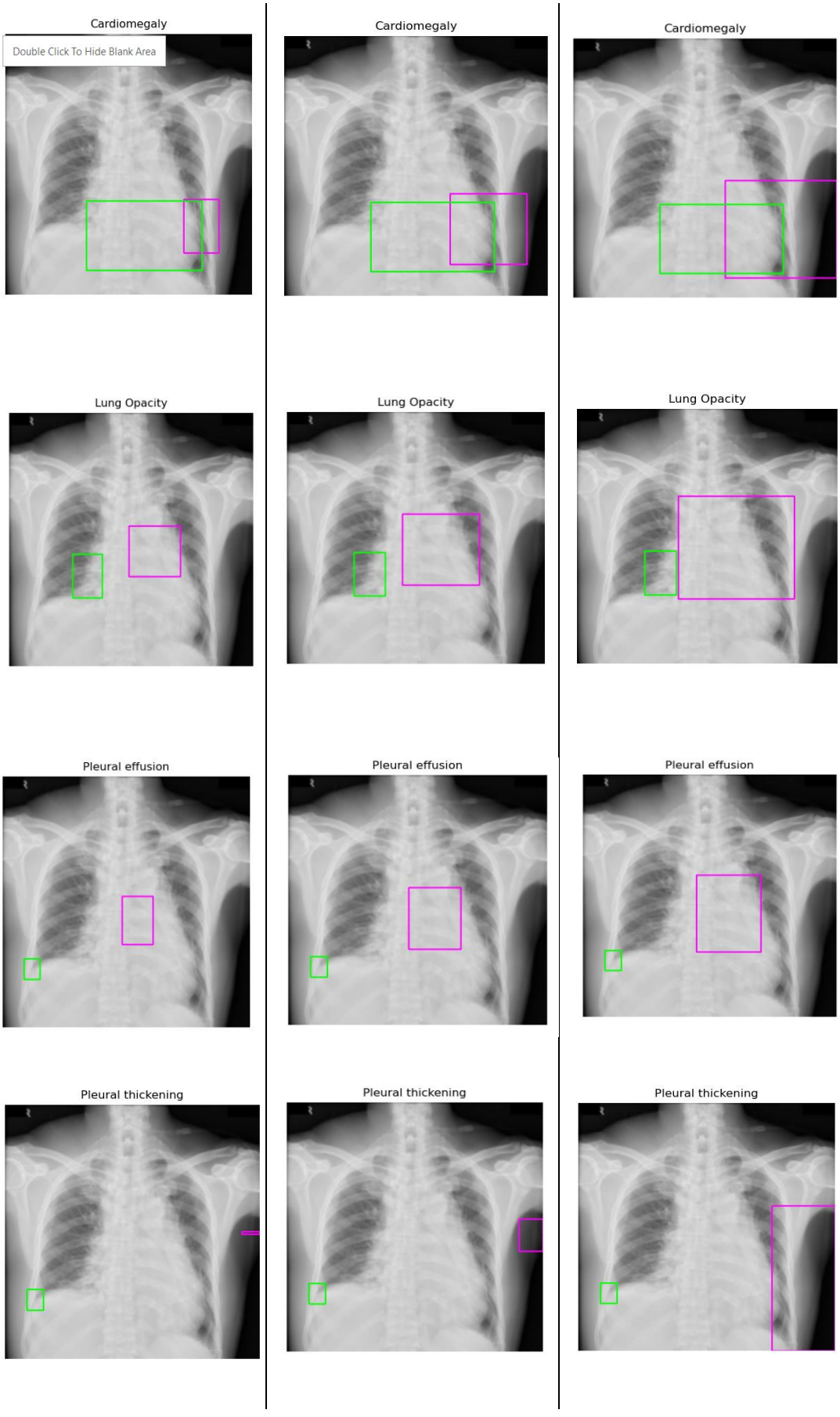
Moreover, the ‘CHAIIN_APPROX_SIMPLE’ contour approximation method is employed, which streamlines horizontal, diagonal and vertical segments, storing only their endpoints. This approach aids in memory usage. Within the loop, the cv2.boundingRect(contour)’ function is applied to determine the bounding rectangle for each contour. This bounding rectangle is the smallest

rectangular area encompassing the contour and provides the coordinates (x,y) of the top-left corner, along with its width (w) and height (h).

IoU metric is a metric that is applied only in the instances where the abnormality has been detected correct and therefore the ground truth bounding boxes of the abnormalities are available. In Figure 28, examples are presented to examine the results of the evaluation and discuss some of the difficulties associated with this method.






Cardiomegaly
Double Click To Hide Blank Area
Cardiomegaly
Cardiomegaly
Lung Opacity
Lung Opacity
Lung Opacity
Pleural effusion
Pleural effusion
Pleural effusion
Pleural thickening
Pleural thickening
Pleural thickening

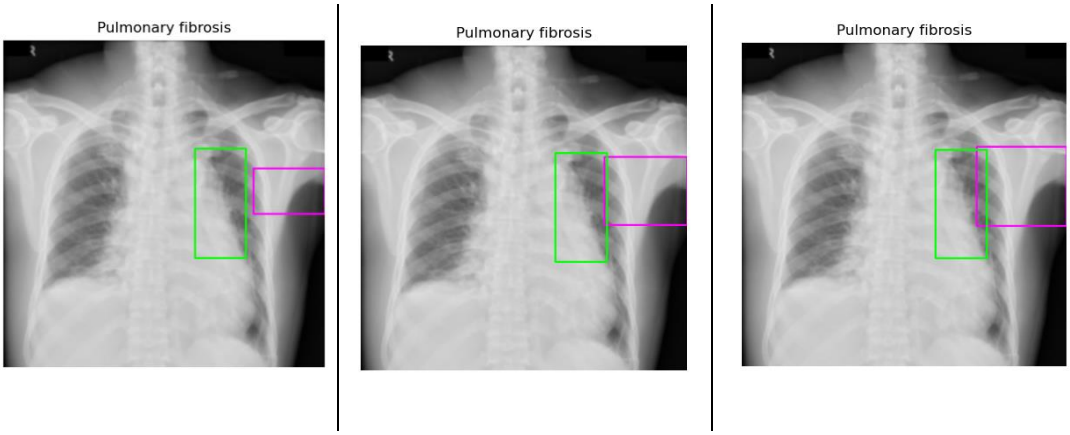


Figure 28: Top figure: Original image annotated with all the pathologies and the Ground Truth boxes. Bottom figures: Bounding boxes for each label (percentages: 0.02, 0.05 and 0.1). Green: GT Box, Purple: Predicted Bounding Box.

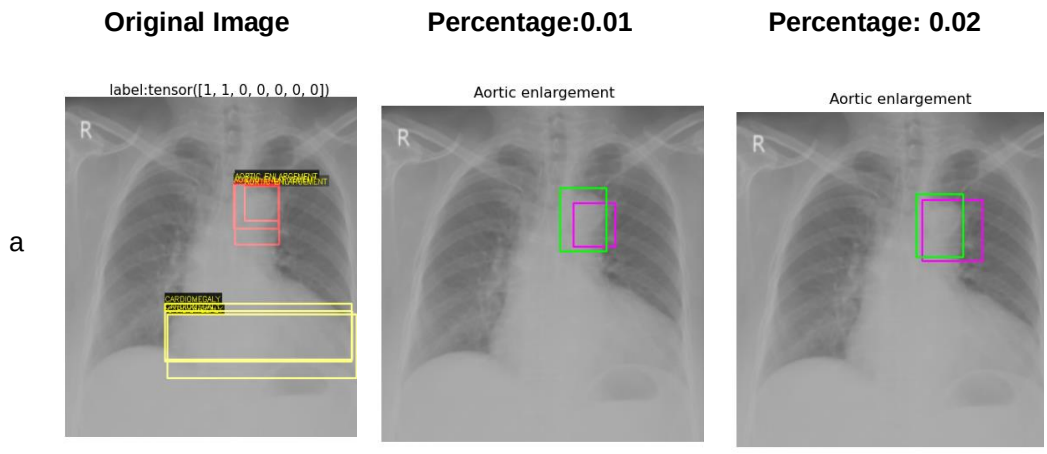
Table 5: Intersection over Union (IoU) for each class in different percentages for Figure 28.

Pcg	Aortic Enlargement	Cardiomegaly	LO	PE	PT	PF
0.01	0.202	0.056	0.0	0.0	0.0	0.0
0.02	**0.463**	0.089	0.0	0.0	0.0	0.0
0.03	0.425	0.145	0.0	0.0	0.0	0.0
0.05	0.36	0.235	0.0	0.0	0.0	0.0
0.8	0.268	0.249	0.0	0.0	0.0	0.0
0.1	0.221	0.25	0.0	0.0	0.0	0.065
0.2	0.092	**0.268**	0.04	0.0	0.0	0.149

In the selected chest X-ray depicted in Figure 28 , the model successfully predicts all existing abnormalities. Images representing each class are accompanied by corresponding bounding boxes for both true and predicted labels, annotated in green and purple, respectively, at three percentages: 0.02, 0.05 and 0.1. The IoU metric in the accompanying Table 5 details the overlap accuracy for all the classes and various percentages. Notably, ‘Aortic Enlargement’ and ‘Cardiomegaly’ exhibit high localization accuracy, with IoU

values of 0.463 and 0.268 respectively. Strikingly, for other classes, the IoU is zero, indicating a lack of overlap between the bounding boxes from the heatmaps and the true labels. The additional figure of the original image with all the annotated boxes, is included to illustrate that certain classes, such as ‘Aortic Enlargement’ and ‘Cardiomegaly’, are annotated by multiple radiologists, while others have sole annotations by a single expert. This diversity in annotations, stemming from different radiologists who may identify varying abnormalities, introduces a challenging aspect to the analysis of chest X-rays in this dataset.

In the provided Figure 29, additional examples are presented to further investigate this phenomenon. For instance, in Figure 29a, it is evident that the class ‘Aortic Enlargement’ has been annotated by three radiologists using bounding boxes with slightly different sizes. Figure 29b illustrates a similar scenario with the class ‘Pleural Thickening’, where three experts have labelled the abnormality, each delineating a distinct region. These instances underscore the challenge faced by the model in precisely localizing the pathology, particularly noteworthy in the case of ‘Pleural Thickening’, where the regions identified by each radiologist are entirely disparate. It is also noteworthy to mention cases where only one radiologist has labelled a specific abnormality.



b

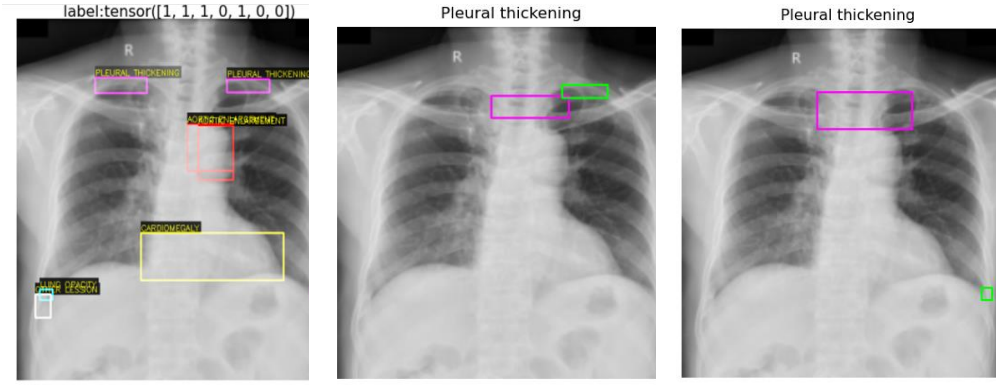


Figure 29: Chest X-rays labelled by different annotators.

Subsequently, IoU metric was calculated for all the chest X-ray images, specifically for the correctly predicted classes. The mean IoU for each class was computed at various percentages, and the results are presented in both Figure 30 and Table 6. Notably, the model exhibits a complete failure in localizing 'Pleural Thickening' and 'Pulmonary Fibrosis', despite accurately predicting the correct class in the chest X-ray image. Once again, it is evident that 'Aortic Enlargement' and 'Cardiomegaly' are the most accurately localized classes.

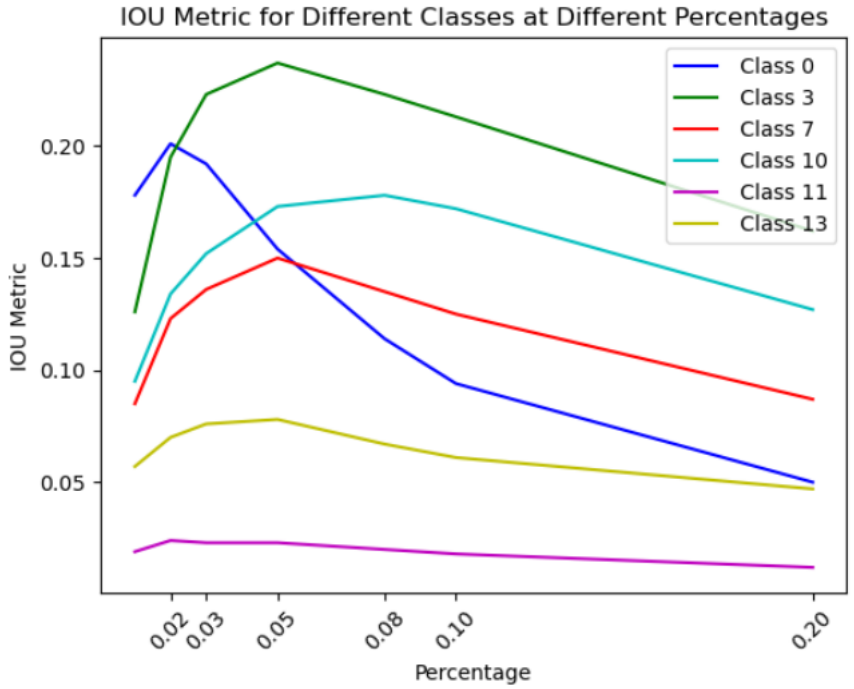


Figure 30: Mean IoU for each class at percentages: 0.01, 0.02, 0.03, 0.05, 0.08, 0.1, 0.2. Classes: 0: 'Aortic Enlargement', 3: 'Cardiomegaly', 7: 'Lung Opacity', 10: 'Pleural Effusion', 11: 'Pleural Thickening', 13: 'Pulmonary Fibrosis', 14: 'No-Finding'.

Table 6: IoU Metric for each class in different percentages.

Pcg	Aortic Enlargement	Cardiomegaly	LO	PE	PT	PF
0.01	0.178	0.126	0.085	0.095	0.019	0.057
0.02	**0.201**	0.195	0.123	0.134	0.024	0.07
0.03	0.192	0.223	0.136	0.152	0.023	0.076
0.05	0.154	**0.237**	**0.15**	0.173	0.023	0.078
0.08	0.114	0.223	0.135	**0.178**	0.02	0.067
0.1	0.094	0.213	0.125	0.172	0.018	0.061
0.2	0.05	0.162	0.087	0.127	0.012	0.047

5.2 Pixel Importance Analysis

In the second evaluation method we employ, a specified percentage of the most significant pixel withing the regions identified by the heatmaps undergo various transformations. The initial transformation involves computing the mean value across all three colour channels and substituting the regions of interest with this calculated mean value. Subsequently, a Gaussian blur is applied to the regions of interest using kernel sizes of (11,11) or (21,21). Gaussian blurring is employed to mitigate noise and reduce image detail, with the degree of blurring contingent upon the standard deviation of the Gaussian kernel. This standard deviation is indirectly influenced by the chosen ‘kernel_size’, where larger kernel_sizes lead to more pronounced blurring effects.

Subsequently, we procced to generate new predictions using our model, where the input to the model is the transformed images. It is anticipated that these predictions will exhibit variations, given that the pixels identified as crucial by Grad-Cam have undergone transformations. This iterative process is applied across all classes. For each image in the test dataset, we extract the corresponding heatmaps and predictions. The image is modified exclusively for the classes predicted by the model for that specific image, and fresh predictions

are generated for all classes. The subsequent step involves calculating the difference between the original prediction from the initial image and the prediction from the transformed image (prediction_original - prediction_tansformed). Ultimately, the mean values of the prediction differences are computed for each class across various percentages (0.02, 0.05, 0.1, 0.2 and 0.5). Figure 31 showcases an example of a transformed image for illustration.

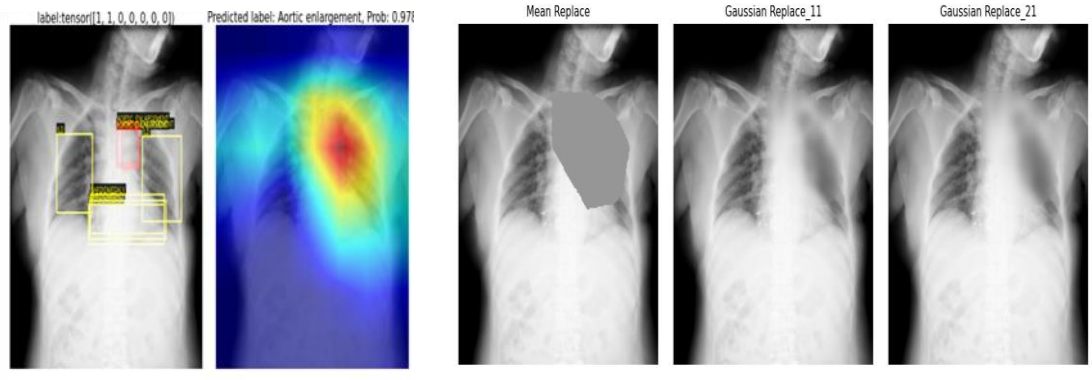
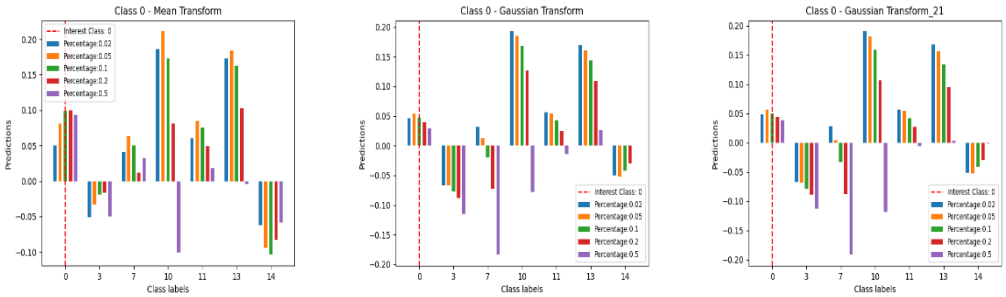


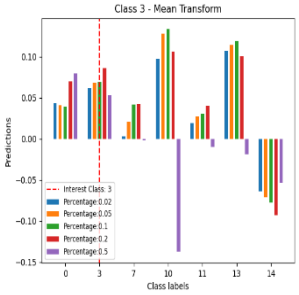
Figure 31: Mean and Gaussian Transformed Images with 10% replacement of the most important pixels.

In the set of Figure 32 presented below, the plots depict the resultant variations in predictions for the transformed images at each percentage. Each figure showcases the transformation of images based on the class indicated in the title. For instance, in Figure 32a, the transformed images correspond to class 0 (‘Aortic Enlargement’), signifying that the images have been transformed based on the heatmaps derived from this class. Additionally, the predictions for the new images are displayed for various percentages.

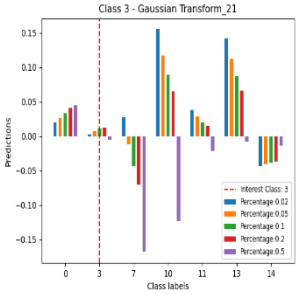
a



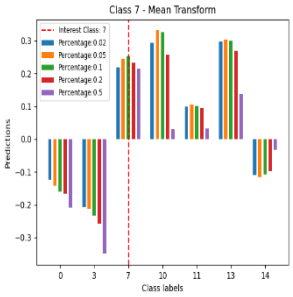
b

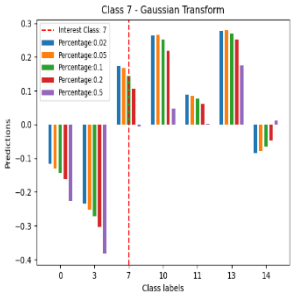

Class 3 - Mean Transform
Predictions
Class labels
Interest Class: 3
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5

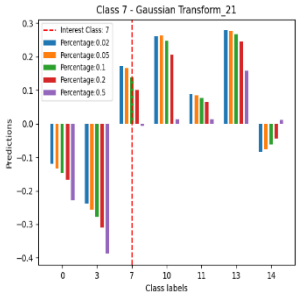

Class 3 - Gaussian Transform
Predictions
Class labels
Interest Class: 3
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5


Class 3 - Gaussian Transform_21
Predictions
Class labels
Interest Class: 3
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5

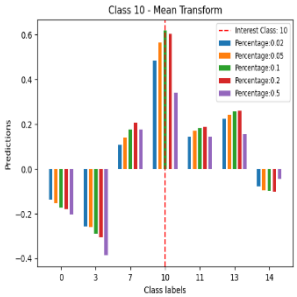
c


Class 7 - Mean Transform
Predictions
Class labels
Interest Class: 7
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5

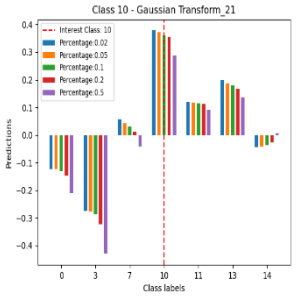

Class 7 - Gaussian Transform
Predictions
Class labels
Interest Class: 7
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5


Class 7 - Gaussian Transform_21
Predictions
Class labels
Interest Class: 7
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5

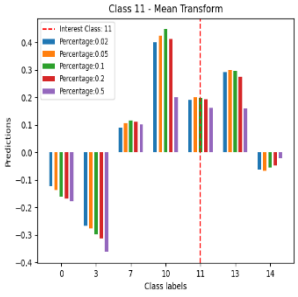
d

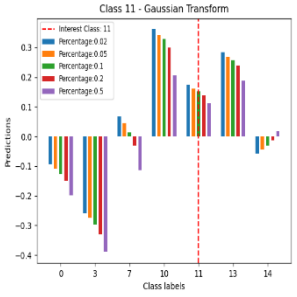

Class 10 - Mean Transform
Predictions
Class labels
Interest Class: 10
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5

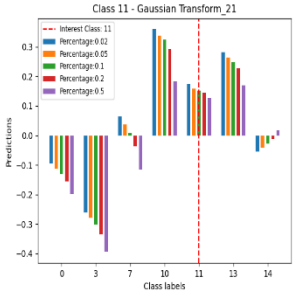

Class 10 - Gaussian Transform
Predictions
Class labels
Interest Class: 10
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5


Class 10 - Gaussian Transform_21
Predictions
Class labels
Interest Class: 10
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5

e


Class 11 - Mean Transform
Predictions
Class labels
Interest Class: 11
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5


Class 11 - Gaussian Transform
Predictions
Class labels
Interest Class: 11
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5


Class 11 - Gaussian Transform_21
Predictions
Class labels
Interest Class: 11
Percentage:0.02
Percentage:0.05
Percentage:0.1
Percentage:0.2
Percentage:0.5

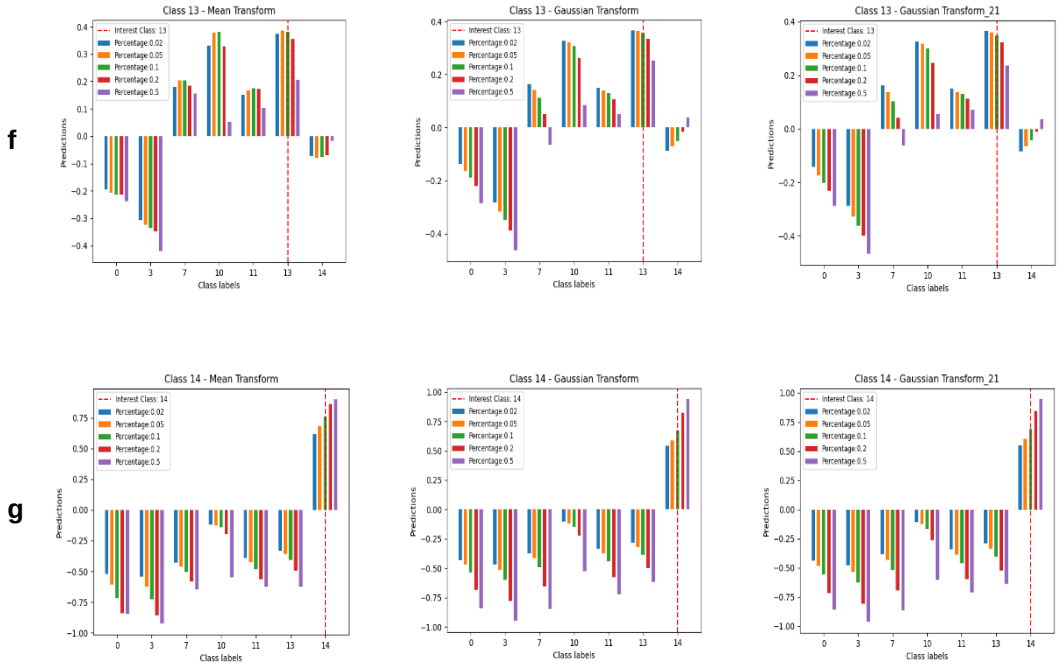


Figure 32: Transformed Images predictions in various percentages. Transformations: Mean, Gaussian Blur (kernel_size = (11,11), Gaussian Blur (kernel_size = (21,21). Each letter (a-g) gives a transformation based on the heatmaps of following classes: (a) 0 – 'Aortic Enlargement', (b) 3 – 'Cardiomegaly', (c) 7 – 'Lung Opacity', (d) 10 – 'Pleural Effusion', (e) 11 – 'Pleural Thickening', (f) 13 – 'Pulmonary Fibrosis', (g) 14 – 'No-Finding'. In the y-axis of the above plots the difference in the prediction is presented ($p_{original} - p_{replaced}$). The more the increase in the prediction the more the reduction of $p_{replaced}$).

In the plots shown in Figure 32 a to g, several noteworthy observations can be made. Firstly, examining Figure 32a, reveals a slight decrease (~0.1) in the prediction of the 'Aortic Enlargement' class following pixel replacement, observed in both mean and Gaussian transformations. Similarly, modest reductions in prediction are observed in the 'Pleural Effusion' and 'Pulmonary Fibrosis' classes. Notably, the classes 'Cardiomegaly', 'Lung Opacity' and 'Pleural Thickening' appear unaffected by this transformation. Lastly, it is important to highlight that the 'No Finding' class demonstrates an increase in prediction.

Regarding 'Cardiomegaly', we note a slight reduction in prediction (0.05 - 0.1) with the maximum decrease reaching in the percentage of 0.2 in mean transformation. Additionally, in Gaussian transform, the reduction in prediction

is even smaller. Similar patterns are observed for the other classes, as mentioned earlier.

In the Figure 32c, where the images are transformed based on the heatmaps generated from the ‘Lung Opacity’ class, a more substantial reduction (0.1 - 0.2) is observed for this class. Simultaneously, ‘Aortic Enlargement’ and ‘Cardiomegaly’ classes show an increase in their prediction. The remaining classes exhibit consistent behaviour with the previous cases.

For ‘Pleural Effusion’, the predictions align with expectations as we increment the pixel replacement percentage. A proportional rise in reduction of prediction probability is evident (mean: 0.4 - 0.6, Gaussian: 0.3 - 0.4). Conversely, ‘Aortic Enlargement’ and ‘Cardiomegaly’ exhibit an increase in their probabilities, notably with ‘Cardiomegaly’ experiencing a 0.4 increment. Other classes appear to be less significantly impacted by this transformation.

For ‘Pleural Thickening’, a marginal reduction in prediction (~0.1) is noted, while other classes exhibit the same behavior as observed in previous cases. A similar pattern is observed when adjusting the image based on the ‘Pulmonary Fibrosis’ class, with a more substantial reduction (~0.3). In both instances, there appears to be an inverse proportional relationship between the percentage increase and the probability of reduction.

In the case of the ‘No-Finding’ class, the observed behavior aligns precisely with expectations. A substantial reduction in the prediction of this class (0.5-1) is evident, demonstrating a proportional relationship with the percentage increase. Concurrently, the other classes display a noteworthy increase in their probabilities (0.5-1), also exhibiting a proportional relationship with the percentage rise. This suggests that the model, upon detecting features indicative of non-normal chest X-rays, shifts its focus towards assessing the likelihood of various diseases.

In conclusion, we can infer that for ‘Aortic Enlargement’ and ‘Cardiomegaly’, a slightly larger increase in the reduction was anticipated, particularly for ‘Cardiomegaly’, which appears to be the class least impacted by the transformations. Interestingly, these two classes demonstrate an increase in predictions when images are transformed based on the heatmaps generated

by the remaining classes. Conversely, ‘Pleural Thickening’ and ‘Lung Opacity’ emerge as the two classes least affected by transformations based on the heatmaps of other classes. Both ‘Pleural Effusion’ and Pulmonary Fibrosis’ exhibit a reduction in predictions, demonstrating consistent behavior when images based on other classes are modified, they keep reducing their prediction. The behavior of ‘Non-Finding’ aligns with expectations, representing the most anticipated response among all the classes.

It is important to consider that chest X-ray pathologies may overlap, meaning that a reduction in the prediction of one class could lead to a corresponding decrease in another class or vice versa. It should also be mentioned that class ‘Pleural Thickening’, as observed in the heatmaps earlier, is not well-localized even when predicted accurately, encompassing a broad region of the image. Additionally, it is worth mentioning that employing a larger kernel size, which increases noise in Gaussian blurring, does not significantly alter the behavior of the model.

6. CONCLUSION

Chest X-rays play a crucial role in diagnosing chest abnormalities and serve as an essential tool for this purpose. In recent years, extensive research has been conducted to create datasets and develop pipelines for the accurate classification and localization of these abnormalities. Distinguishing between different abnormalities is a challenging task due to the overlapping nature of these diseases. Moreover, to create an effective tool to assist medical professionals in disease diagnosis, it is imperative to understand the reasons behind a model's decision. Therefore, it is crucial to experiment with various explainability methods and integrate them into clinical routines. This ensures that healthcare professionals have access to these tools, providing valuable insights into the model's decision-making process.

In this research, we utilized the VinDr [8] dataset, encompassing a range of abnormalities, to conduct experiments focused on classifying distinct abnormalities. We employed Grad-Cam as an explainability method to elucidate the reasons behind the model's outcomes. Subsequently, we conducted various experiments to evaluate the effectiveness of this method and gauge the robustness of the model based on the results obtained from the explainability analysis.

From the obtained results, it is evident that the model does not achieve a satisfactory F1 score for all classes. Notably, classes such as 'Aortic Enlargement' and 'Cardiomegaly' exhibit a respectable F1 score (0.86 and 0.83 respectively). These classes demonstrate greater stability as they appear in the same region in the chest X-ray with consistent shapes and sizes. The model effectively predicts these classes, and the resulting heatmaps show a satisfactory alignment with the bounding boxes provided by the annotators (IoU: 0.201 and 0.237 respectively).

The remaining classes exhibit lower F1 scores, with 'Lung Opacity' and 'Pulmonary Fibrosis' showing the lowest values. The resulting heatmaps generated by Grad-Cam are confusing. The model appears to face challenges, particularly in localizing the class 'Pleural Thickening', placing it across a broad region of chest X-rays. Heatmaps for 'Lung Opacity' and 'Pleural Effusion'

exhibit more consistent alignments with the disease regions when predicted correctly. While minimal overlap is observed for ‘Lung Opacity’ and ‘Pleural Effusion’, IoU metrics are disappointing for ‘Pleural Thickening’ and ‘Pulmonary Fibrosis’. It is crucial to note that these diseases are annotated by different radiologists. There are instances where only one radiologist labels the chest X-ray with the disease, while others do not, or different radiologists identify diverse regions for the disease, particularly in the case of ‘Pleural Thickening’. This diversity in annotation practices adds complexity, making it more challenging for the model to accurately predict these abnormalities.

Finally, it should be mentioned that healthy chest X-rays exhibit the highest F1 scores (0,96), and the corresponding heatmaps present a stable behaviour. In most instances, both lungs are highlighted, or specific regions below the letters indicating the right or left side of the chest X-ray. Furthermore, it is observed from the heatmaps that the model successfully localizes the diseased regions of chest X-rays, but encounters difficulty in distinguishing between the various abnormalities.

As for the importance of pixel analysis, the expectation is that by replacing part of the most crucial pixels, as identified by the Grad-Cam generated heatmaps, the probability of prediction for each class should decrease. From the conducted experiments, we observe a slight reduction in the prediction probability for classes ‘Aortic Enlargement’ and ‘Cardiomegaly’, although not as significant as anticipated. Conversely, classes such as ‘Pleural Effusion’, ‘Lung Opacity’ and ‘Pulmonary Fibrosis’ exhibit a more pronounced reduction in prediction probability. ‘Pleural Thickening’ appears to be the least affected by the altered figures. The class exhibiting the anticipated behaviour is ‘No-Finding’, where a proportional increase in reduction to the prediction probability is observed with the rising percentage of replaced pixels, while simultaneously, the remaining classes show an increase in their prediction.

As evident from the above results, it is once again underscored that chest X-ray classification is a challenging task and further research should be done. First, addressing the need for a more balanced dataset is imperative, as numerous classes are excluded due to dataset imbalance, a challenge persisting in our research too. Despite the recent publication of several

datasets, achieving balance remains difficult. It could also be possible to check the robustness of the model in different datasets. Conducting additional experiments is essential for enhancing accuracy and F1 scores, instilling greater confidence in the results obtained from explainability methods.

The utilization of explainability methods is crucial and exploring additional methods or combinations of explainable artificial intelligence (XAI) techniques is recommended. Optimal evaluation by experts is a pivotal initial step for an XAI method, offering insights into its effectiveness. However, more rigorous sanity checks are warranted to assess these methods thoroughly. One proposed sanity check involves training a model with the transformed images from the pixel importance analysis and evaluating the new predictions derived from these images, as suggested Hook et al. [39].

In conclusion, this research underscored the utility of Grad-Cam in chest X-ray analysis, leading to valuable insights into various abnormalities and highlighting challenges encountered, including the nuanced nature of certain diseases and variations in annotations form different radiologists. XAI methods are essential for introducing Deep Learning algorithms in medical domain, consequently evaluation techniques should be developed to assess the performance of these methods. This research performed different experiments for the evaluation with some interesting findings. These findings suggest that further investigation and additional experiments are imperative to deepen our understanding and refine the methodologies employed.

ABBREVATIONS

XAIs	eXplainable Artificial Intelligence
DNNs	Deep Neural Networks
CNNs	Convolutional Neural Networks
ReLU	Rectified Linear Unit
RGB	Red-Green-Blue
ResNets	Residual Networks
CAM	Class Activation Map
Grad-CAM	Gradient-weighted Class Activation Map
IG	Integrated Gradients
LRP	Layer-wise Propagation
DeconvNet	Deconvolution Network
jpg	Joint Photographic Experts Group
PIL	Python Imaging Library
SGD	Stochastic Gradient Descent
AUC	Area Under the Curve
IoU	Intersection over Union
CXR	Chest X-rays
MRI	Magnetic Resonance Image
NLP	Natural Language Processing

REFERENCES

[1] *Explainable artificial intelligence (XAI) in deep learing-based medical image analysis.* **Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, Max A. Viergever.** s.l. : Medical Image Analysis, 2022, Vol. 79.

[2] *Explainable AI in medical imaging: An overview for clinical practioners-Saliency-based XAI approaches.* **Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Kramer, Christoph M. Friedrich, Felix Nensa.** s.l. : European Journal of Radiology, 2023, Vol. 162.

[3] *Explainble Deep Learning Models in Medical Image Analysis.* **Amitojdeep Singh, Sourya Sengupta and Vasudevan Lakshminarayanan.** s.l. : Journal of Imaging, 2020.

[4] *ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervises Classification and Localization of Common Thorax Diseases.* **Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers.** 2017. IEEE: Conference on Computer Vision and Pattern Recognition (CVPR).

[5] *Padchest: A large chest X-ray image dataset with multi-label annotated reports.* **Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, Maria de la Iglesia-Vaya.** s.l. : Medical Image Analysis, 2019, Vol. 66.

[6] *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison.* **Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P. and Ng,.** 2019. Proceedings of the AAAI Conference in Artificial Intelligence.

[7] *MIMIC-CXR, a de-identified publicly available database of chest ragiographs with free text reports.* **Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J. et al.** 317, s.l. : Sci Data, 2019, Vol. 6.

[8] **al., Ha Q. Nguyen et.** *VinDr-CXR: An open dataset of chest X-rays with radiologist's anntations.* [arXiv:2012.15029v3] 2022.

[9] **Vu, Hieu H. Pham and Ha Q. Nguyen and Khanh Lam and Linh T. Le and Dung B. Nguyen and Hieu T. Nguyen and Tung T. Le and Thang V. Nguyen and Minh Dao and Van.** *An Accurate and Explainable Deep Learning System Improves Interobserver Agreement in the Interpretation of Chest Radiograph.* 2021.

[10] *EfficientNet: Rethinking model scaling for convolutional neural networks.* **V.Le, Mingxing Quoc V.Le.** 2019. In International Conference on Machine Learning (ICML).

[11] *EfficientDet: Scalable and efficient object detectioin.* **Mingxing Tan, Ruoming Pang and Quoc V.Le.** 2020. Institute of Electrical and Electronics Engineers/Computer Vision Foundation IEEE/CVF conference on CVPR.

[12] **Sun, Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian.** *Deep Residual Learning for Image Recognition.* [arxiV:1512.03385] 2015.

[13] *Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.* **Selvaraju, Ramprasaath R. and Cogswell, Michael and Das, Abhishek and Vedantam, Ramakrishna and Parikh, Devi and Batra, Dhruv.** s.l. : Springer Science and Business Media LLC, Oct 2019, International Journal of Computer Vision, Vol. 128, pp. 336-359.

[14] **Albawi, Saad, Abed Mohammed, Tareq, ALZAWI, Saad.** *Understanding of a Convolutional Neural Network.* 2017.

[15] **Mishra, Mayank.** Towards Data Science. [Online] 2020. https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939.

[16] **Walter Hugo Lopez Pinaya, Sandra Vieira, Rafael Garcia-Dias, Andrea Mechelli.** Convolutional Neural Networks. [ed.] Sandra

Vieira and Andrea Mechelli. *Machine Learning.* s.l. : Academic Press, 2020, 10, pp. 173-191.

[17] **Zisserman, Karen Simonyan and Andrew.** *Very Deep Convolutional Networks for Large-Scale Image Recognition.* [arXiv:1409.1556] 2015.

[18] *Why Should I Trust You? Explaining the Predictions of Any Classifier.* **Ribeiro, Marco Tulio, Singh, Sameer and Guestrin, Carlos.** New York : s.n., 2016. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135-1144.

[19] *A Unified Approach to Interpreting Model Predictions.* **Lundberg, Scott M. and Lee, Su-In.** 2017. Advances in Neural Information Processing Systems (NeurIPS).

[20] **Molnar, C.** *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd edition).* 2022.

[21] **Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman.** *Deep inside convolutional networks: Visualising image classification models and saliency maps.* [arXiv:1312.6034] 2013.

[22] *Visualising and understanding convolutional networks.* **Zeiler, Matthew D., and Rob Fergus.** 2014. European Conference on Computer Vision.

[23] *Explaining nonlinear classification decisions with deep Taylor decomposition.* **Blach, Layer Wise Propagation (LRP) is an XAI method proposed by Gregoire Montavon and Sebastian.** s.l. : Elsevier BV, May 2017, Pattern Recognition, Vol. 65, pp. 211-222.

[24] **Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, Martin Wattenberg.** *SmoothGrad: removing noise by adding noise.* [arXiv:1706.03825] 2017.

[25] **Yan, Mukund Sundararajan and Ankur Taly and Qiqi.** *Axiomatic Attribution for Deep Networks.* [arXiv: 1703.01365] 2017.

[26] *Implementation of model explainability for a basic brain tumor detecting using convolutional neural networks on MRI slices.* **Windisch P, Weber P, Fürweger C, Ehret F, Kufeld M, Zwahlen D, Muacevic A.** s.l. : Epub, 2020, Neuroradiology, Vol. 62, pp. 1515-1518.

[27] *Explainability of deep neural networks for MRI analysis of brain tumors.* **Zeineldin, R.A., Karar, M.E., Elshaer, Z. et al.** 2022, Int J CARS, Vol. 17, pp. 1673-1683.

[28] *OView-AI Supporter for Classifying Pneumonia, Pneumothorax, Tuberculosis, Lung Cancer Chest X-ray Images Using Multi-Stage Superpixels Classification.* **Oh J, Park C, Lee H, Rim B, Kim Y, Hong M, Lyu J, Han S, Choi S.** 2023, Diagnostics, Vol. 13, p. 1519.

[29] **Ng, Pranav Rajpurkar and Jeremy Irvin and Kaylie Zhu and Brandon Yang and Hershel Mehta and Tony Duan and Daisy Ding and Aarti Bagul and Curtis Langlotz and Katie Shpanskaya and Matthew P. Lungren and Andrew Y.** *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning.* [arXiv:1711.05225] 2017.

[30] *Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification.* **Böhle M, Eitel F, Weygandt M, Ritter K.** Jul 2019, Front Aging Neurosci, Vol. 11, p. 194.

[31] *Explainable Deep Learing for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays.* **Brunese L, Mercaldo F, Reginelli A, Santone A.** 2020, Compute Methods Programs Biomed, Vol. 196.

[32] *ImageNet Classification with Deep Convolutional Neural Networks.* **Krizhevsky, Alex and Sutskever, Ilya and Hinton, Geoffrey E.** [ed.] F. Pereira and C.J. Burges and L.Bottou and K.Q. Weinberger. s.l. : Curran Associates, Inc, 2012. Advances in Neural Information Processing Systems.

[33] *Going Deeper with Convolutions.* **Rabinovich, Christian Szegedy and Wei Liu and Yangqing Jia and Pierre Sermanet and Scott Reed and Dragomir Anguelov and Dumitru Erhan and Vincent Vanhoucke and Andrew.** 2014. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[34] **Weinberger, Gao Huang and Zhuang Liu and Laurens van der Maaten and Kilian Q.** *Densely Connected Convolutional Networks.* [arXiv:1608.06993] 2018.

[35] **Phillips, N. A. et al.** *CheXphoto:10,000+ smartphone photos and synthetic photographic transformations of chest X-rays for benchmarking deep learning robustness.* [arXiv:2007.06199] 2020.

[36] **Doran, Gabrielle Ras and Ning Xie and Marcel van Gerven and Derek.** *Explainable Deep Learining: A Field Guide for the Uninitiated.* [arXiv:2004.14545] 2021.

[37] *On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation.* **Bach, Sebastian and Binder, Alexander and Montavon, Grégoire and Klauschen, Frederick and Müller, Klaus-Robert and Samek, Wojciech.** 7, 2015, PLOS ONE, Vol. 10.

[38] **Zisserman, Karen Simonyan and Andrea Vedaldi and Andrew.** *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.* [arXiv:1312.6034] 2014.

[39] *A Benchmark for Interpretability Methods in Deep Neural Networks.* **Hooker, Sara and Erhan, Dumitru and Kindermans, Pieter-Jan and Kim, Been.** s.l. : Curran Associates, Inc, 2019. Advances in Neural Information Processing Systems (NeurIPS). Vol. 32.

[40] **Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B.** The (un)reliability of saliency methods. *Explainable AI: Interpreting, Explaining and Visualizind Deep Learning.* s.l. : Springer, 2019, pp. 267-280.

[41] **Vu, M. N., Nguyen, T. D., Phan, N., Gera, R., and Thai, M. T.** *c-Eval: A unified metric to evaluate feature-based explanations via perturbation .* [arXiv:abs/1906.02032] 2019.

[42] *Sanity checks for saliency maps.* **Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B.** 2018. Advances in Neural Information Processing Systems (NeurIPS).

[43] **al., Ha Q. Nguyen et.** *VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations.* [arXiv:2012.15029v3] 2022.