FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

# Explainable Artificial Medical Intelligence for Automated Thoracic Pathology Screening

**Joana Maria Neves da Rocha**




Doctoral Program in Electrical and Computer Engineering

Supervisor: Prof. Ana Maria Mendonça, PhD, FEUP

Co-supervisor: Prof. Aurélio Campilho, PhD, FEUP

May 16, 2025

© Joana Rocha, 2025

# Explainable Artificial Medical Intelligence for Automated Thoracic Pathology Screening

**Joana Maria Neves da Rocha**

Doctoral Program in Electrical and Computer Engineering

May 16, 2025

# Abstract

Thoracic pathologies, such as lung cancer and pneumonia, have a high incidence worldwide and often lead to severe, life-threatening complications if not promptly diagnosed and treated. Chest X-Ray (CXR) images are crucial for the screening and management of these conditions, enabling clinicians to evaluate disease severity, monitor treatment progress, and identify complications early. Their non-invasive nature and accessibility make them an indispensable tool in both emergency and routine medical care. Given the complexity and subtle variations in CXR images, and consequently their time-consuming analysis, computer-aided diagnosis systems are being studied with growing interest as tools to help radiologists improve diagnostic accuracy and efficiency. Backed by more powerful computational resources and optimized training routines, deep learning models have proven unprecedented performance in extracting information from CXR data, making them highly suited for these applications. However, their lack of transparency is particularly problematic in healthcare because it limits clinicians' ability to understand, trust, and validate the decision-making process, potentially leading to errors in diagnosis and treatment.

With this in mind, this PhD thesis aims to develop a set of image analysis models to not only prioritize abnormal exams, but also accurately identify potential pathologies, while overcoming some of the inherent limitations of opacity in deep learning. The proposed models address the critical need for transparency in Medicine through eXplainable Artificial Intelligence (XAI), emphasizing not only predictive diagnostic accuracy but also providing clear justifications for the models' recommendations. State-of-the-art XAI methods typically use heat maps for that purpose, to highlight the most decisive regions for a given prediction. The thesis intends to enhance the quality of explanations by optimizing feature learning, refining the interpretation of current heat map-generating methods, and introducing a novel XAI approach grounded in a more inherently transparent classification framework. Each of these three topics is explored in a distinct study, described below.

The first one assesses the negative impact of image artifacts on a binary CXR classification task for detecting abnormal scans, since they introduce harmful biases that can significantly increase the rate of false positive results. To address this, an attention-driven approach, STERN, is proposed to automatically select the thoracic region of interest. By removing artifacts in the periphery of the images, STERN helps the model focus on essential anatomical features, enhancing both the classifier's performance and explainability outcomes in an end-to-end training framework. The results indicate that STERN achieves competitive performance with reduced computational costs and without requiring localization labels.

The second study is centered on a multi-label classification task targeting 14 radiological findings and on the primary limitations of current visual explanation techniques, such as the loss of detail due to heat map normalization and the inability to differentiate between reliable and unreliable interpretations. The proposed Confident-CAM method improves heat map interpretation by integrating the model's confidence in the diagnoses. This approach adds an extra validation step, using a confidence score to visually alert the users regarding prediction and/or explanation relia-

bility. Results indicate that the confidence score effectively distinguishes between likely accurate and inaccurate explanations.

At last, the final study presents CLARE-XR, a novel classification framework that eliminates the need for heat maps when diagnosing radiological findings. CLARE-XR identifies pathologies by mapping each input image to a 2D latent space and comparing the coordinates of that image to those of reference findings, obtained through label embedding. This approach enables the retrieval of similar cases, mimicking clinical practice. CLARE-XR provides both textual and visual justifications for each prediction, making it inherently transparent and interpretable. Results demonstrate that CLARE-XR offers an effective alternative to heat maps, outperforming typical classifiers and providing more transparent personalized explanations.

# Resumo

As patologias torácicas, como o cancro do pulmão e a pneumonia, têm uma elevada incidência em todo o mundo e levam frequentemente a complicações graves e potencialmente fatais se não forem prontamente diagnosticadas e tratadas. As imagens de radiografia torácica são cruciais para o rastreio e gestão destas condições, permitindo aos médicos avaliar a gravidade da doença, monitorizar o progresso do tratamento e identificar complicações precocemente. A sua natureza não invasiva e acessibilidade tornam estas radiografias numa ferramenta indispensável, tanto nos cuidados médicos de emergência como de rotina. Dada a complexidade e variações subtis nas imagens de radiografia torácica e, consequentemente, a sua análise demorada, os sistemas de diagnóstico assistido por computador estão a ser estudados com crescente interesse como instrumentos para ajudar os radiologistas a melhorar a precisão e a eficiência do diagnóstico. Apoiados em recursos computacionais mais poderosos e rotinas de treino otimizadas, os modelos de aprendizagem profunda têm comprovado um desempenho sem precedentes na extração de informação de dados de radiografias torácicas, tornando-os altamente convenientes para estas aplicações. No entanto, a sua falta de transparência é particularmente problemática nos cuidados de saúde porque limita a capacidade dos médicos de compreender, confiar e validar o processo de tomada de decisão, potencialmente conduzindo a erros no diagnóstico e no tratamento.

Com isto em mente, esta tese de doutoramento visa desenvolver um conjunto de modelos de análise de imagem para não só priorizar exames anormais, como também identificar com precisão potenciais patologias, ao mesmo tempo que supera algumas das limitações inerentes à opacidade na aprendizagem profunda. Os modelos propostos abordam a necessidade crítica de transparência na Medicina através da Inteligência Artificial eXplicável (XAI), enfatizando não só a precisão diagnóstica preditiva, mas também fornecendo justificações claras para as recomendações dos modelos. Os métodos XAI de última geração utilizam normalmente mapas de calor para este fim, para destacar as regiões mais decisivas para uma determinada previsão. A tese pretende melhorar a qualidade das explicações através da otimização da aprendizagem de características, refinando a interpretação dos métodos atuais de geração de mapas de calor e introduzindo uma nova abordagem XAI baseada numa estrutura de classificação mais inerentemente transparente. Cada um destes três tópicos é explorado num estudo distinto, descrito a seguir.

O primeiro avalia o impacto negativo dos artefactos numa tarefa de classificação binária para deteção de anormalidade em imagens de radiografia torácica, uma vez que estes introduzem um viés prejudicial que pode aumentar significativamente os resultados falsos positivos. Neste sentido, é proposta uma abordagem orientada para a atenção, STERN, que tem como objetivo selecionar automaticamente a região torácica de interesse. Ao remover artefactos na periferia das imagens, a STERN ajuda o modelo a focar-se nas características anatómicas essenciais, melhorando o desempenho do classificador e os resultados de explicabilidade numa estrutura de treino extremo-a-extremo. Os resultados indicam que a STERN atinge um desempenho competitivo, com custos computacionais reduzidos e sem necessidade de anotações espaciais.

O segundo estudo centra-se numa tarefa de classificação *multi-label* visando 14 achados radiológicos e nas principais limitações das atuais técnicas de explicação visual, como a perda de detalhe devido à normalização dos mapas de calor e a incapacidade de diferenciar entre interpretações fiáveis e não fiáveis. O método proposto, Confident-CAM, melhora a interpretação dos mapas de calor, integrando a confiança do modelo nos seus diagnósticos. Esta abordagem adiciona uma etapa extra de validação, utilizando uma pontuação de confiança para alertar visualmente os utilizadores sobre a fiabilidade da previsão e/ou explicação. Os resultados indicam que a pontuação de confiança distingue eficazmente entre explicações provavelmente precisas e imprecisas.

Por fim, o último estudo apresenta o método CLARE-XR, uma estrutura de classificação inovadora que elimina a necessidade de mapas de calor aquando do diagnóstico de achados radiológicos. O CLARE-XR identifica patologias mapeando cada imagem de entrada num espaço latente 2D e comparando as coordenadas dessa imagem com as coordenadas de achados de referência, obtidas através da projeção de anotações. Esta abordagem permite a recuperação de casos semelhantes, mimetizando a prática clínica. O CLARE-XR fornece justificações textuais e visuais para cada previsão, tornando-se inerentemente transparente e interpretável. Os resultados demonstram que o CLARE-XR oferece uma alternativa eficaz aos mapas de calor, superando os classificadores típicos e fornecendo explicações personalizadas mais transparentes.

# Acknowledgments

I am deeply grateful for the past four years, and I finish this journey with a profound appreciation for how much I have grown, both professionally and personally. None of this would have been possible without the support of CBER's Biomedical Imaging Lab at INESC-TEC, and the valuable, caring, and rigorous supervision of Prof. Ana Maria Mendonça and Prof. Aurélio Campilho. I would also like to acknowledge Dr. Pedro Sousa, for his brief yet uplifting collaboration.

I feel incredibly fortunate to have shared this journey with an exceptional group of people, my lab colleagues, who helped me conquer this challenge day by day. A special thank you to Tânia Melo, Carlos Ferreira, and João Pedrosa for their warm welcome and steadfast support ever since I joined the group. It has been a true privilege to learn from you and to experience such a strong sense of unity. Moreover, my deepest gratitude to my favorite co-author, Sofia Cardoso Pereira, who came into my life at the very beginning of the PhD. I want to thank you for your unwavering support and encouragement throughout these four years, and for allowing me to experience the true essence of your friendship, even in the most stressful times. I cannot imagine anyone I would have rather shared this journey with. I wish to carry each of you with me as I embark on this next stage of my life.

Finally, I must acknowledge my mother, the most solid pillar of my life, for always believing in me and continuously inspiring me to pursue my dreams with confidence and determination. Thank you for your irreplaceable presence and guidance.

Joana Rocha was funded by the FCT - Portuguese Foundation for Science and Technology, I.P., grant contract *10.54499/2020.06595.BD*. This work was also partially financed by the ERDF - European Regional Development Fund, through the *Programa Operacional Regional do Norte* (NORTE 2020) and by National Funds through FCT within the scope of the CMU Portugal Program (*NORTE-01-0247-FEDER-045905*) and *LA/P/0063/2020*.



# Contents


**1 Introduction** **1**
- 1.1 Thesis Motivation . . . 1
- 1.2 Thesis Objectives and Contributions . . . 3
- 1.3 Thesis Outline . . . 5

**2 Chest X-Ray Image Analysis** **6**
- 2.1 Clinical Background . . . 6
- 2.2 Advances in Computer-Aided Diagnosis . . . 10
- 2.3 Public Chest Radiograph Datasets . . . 14
- 2.4 Concluding Remarks . . . 19

**3 Explainability in Deep Learning** **20**
- 3.1 Background . . . 20
- 3.2 Taxonomy . . . 21
- 3.3 Methodologies . . . 24
- 3.4 Risks and Evaluation . . . 27
- 3.5 Explainable Thoracic Pathology Screening . . . 28
- 3.6 Concluding Remarks . . . 40

**4 STERN: Attention-driven Spatial Transformer Network for Abnormality Detection in Chest X-ray Images** **42**
- 4.1 Introduction . . . 43
- 4.2 Related Work . . . 44
- 4.3 Method . . . 47
  - 4.3.1 STERN . . . 47
  - 4.3.2 Dataset Preprocessing . . . 49
  - 4.3.3 Experimental Setup . . . 50
  - 4.3.4 ROI Selection Evaluation Metrics . . . 52
- 4.4 Results and Discussion . . . 54
  - 4.4.1 Evaluation of ROI Selection . . . 55
  - 4.4.2 Classifier Performance . . . 59
- 4.5 Concluding Remarks . . . 63

**5 Confident-CAM: Improving Heat Map Interpretation in CXR Image Classification** **64**
- 5.1 Introduction . . . 65
- 5.2 Related Work . . . 66
- 5.3 Method . . . 67
  - 5.3.1 Goals and Premises . . . 67

5.3.2 Confident-CAM Algorithm . . . 69
5.3.3 Dataset Preprocessing . . . 70
5.3.4 Multi-label Classifier . . . 71
5.4 Results and Discussion . . . 71
5.4.1 Method Validation . . . 79
5.5 Concluding Remarks . . . 82

**6 CLARE-XR: Explainable Regression-based Classification of Chest Radiographs with Label Embeddings 84**
6.1 Introduction . . . 85
6.2 Related Work . . . 86
6.3 Method . . . 87
6.3.1 Dataset Preprocessing . . . 87
6.3.2 Framework . . . 89
6.4 Results and Discussion . . . 93
6.4.1 Baseline Classifier . . . 93
6.4.2 Label Autoencoder . . . 95
6.4.3 Embedding-based Regression . . . 96
6.4.4 Multi-label Classification . . . 97
6.4.5 PCA *vs.* Label Autoencoder . . . 103
6.4.6 Explainability . . . 107
6.4.7 Limitations . . . 109
6.5 Concluding Remarks . . . 110

**7 Conclusions and Future Work 111**

**Publications 114**

**References 116**

**A Evaluation Metrics 136**

**B Supplementary Material of CLARE-XR 138**

# List of Figures


1.1 A typical XAI framework analyzes each CXR image to generate a final diagnosis, such as cardiomegaly and effusion in this example, and provides heat map-based explanations that highlight regions with higher activation values, indicating areas of the image most likely associated with the identified findings. Each prediction has an associated probability ($P$). . . . . 3
1.2 Impact of XAI implementation on clinician end-users. Adapted from Lunkad (2023). 3
1.3 The main tasks and respective contributions of the present PhD work, contextualized in a typical CAD workflow. . . . . 5

2.1 Illustration of the thoracic cavity and its constituents. OpenStax (2024). . . . . 7
2.2 Schematic of a digital radiography system. Adapted from Sprawls Jr (1993). . . . 9
2.3 CXR image with (a) frontal and lateral views, and (b) posterior-anterior and anterior-posterior projections, respectively. St Vincent's University Hospital (2024); Algarve Hospital and University Centre (2018). . . . . 9
2.4 Eight common thoracic conditions represented in CXR images. Wang et al. (2017a). 10
2.5 Typical CAD workflow for medical image analysis. Adapted from Nillmani et al. (2022). . . . . 11
2.6 Number of publications per year, presenting chest radiography-focused DL applications. Data collected from *Scopus*, using the search criteria `(deep learning) AND (chest OR thoracic) AND (x-ray OR radiography)`. Results for 2024 are only recorded up to September, 2024. . . . . 12

3.1 Taxonomy of XAI methods. Singh et al. (2020). . . . . 22
3.2 The three stages of explainability: pre-model, in-model, and post-model explainability. Khaleghi (2019). . . . . 23
3.3 Gradient-based saliency maps for image classification. Reyes et al. (2020). . . . . 26
3.4 The predicted class score is mapped back to the previous convolutional layer to generate the CAMs, thus visualizing the class-specific discriminative regions. The $w_i$ values are the learned weights associated with each feature map after the GAP layer. Zhou et al. (2016). . . . . 26
3.5 Number of publications per year, presenting general and chest radiography-focused XAI applications, respectively. Data collected from *Scopus*, using the search criteria `explainable AND (ai OR intelligence)` and `(explainable OR xai) AND (chest OR thoracic) AND (x-ray OR radiography)`, respectively. Results for 2024 are recorded up to September, 2024. . . . . 29
3.6 Overview of (a) explanation types, and (b) their average end user rating. Herm et al. (2023). . . . . 30
3.7 Eight key system designs for XAI in thoracic pathology screening. Xie et al. (2020). 31

3.8 Infiltration case from the ChestX-ray14 dataset, with the spatial annotation marked in a red box. The ResNet50 model's Grad-CAM is displayed to its right. The explanation's resulting mask on the far right selects the same number of high-intensity pixels as the ones present in the BB. . . . . 32
3.9 Effusion case from the ChestX-ray14 dataset, with the spatial annotation marked in a red box. The ResNet50 model's Grad-CAM explanation is displayed to its right. The EBPG calculations are derived from the point-wise multiplication of the annotation and the heat map, shown on the far right. . . . . 32
3.10 Explainability maps obtained by the ResNet50 model for a cardiomegaly instance from the ChestX-ray14 dataset, using high-resolution XAI methods: integrated gradients (*IGr*), saliency (*Sal*), and input $\times$ gradient ($I \times G$). The matrix on the right shows the average cosine similarity across the entire dataset for the same methods. . . . . 34
3.11 Explainability maps obtained for eight instances of the independent test set of ChestX-ray14, using the ResNet50 model. . . . . 36
3.12 Explainability maps obtained for the same eight instances of the independent test set of ChestX-ray14 as Figure 3.11, using the DenseNet121 model. . . . . 36
3.13 Average cosine similarity across the entire dataset for the DenseNet121 model, using high-resolution XAI methods: integrated gradients (*IGr*), saliency (*Sal*), and input $\times$ gradient ($I \times G$). . . . . 37
3.14 Explainability maps obtained for three instances of the ChestX-ray14 dataset, using high-resolution XAI methods (integrated gradients, saliency, and input × gradient, respectively) on the ResNet50 model (left) and DenseNet model (right). The cosine similarity (*CSim*) between the three maps of that particular instance and model is displayed on top of the original radiograph. . . . . 40

4.1 Example of an image before and after cropping, in which the outer frame of the image is discarded, consequentially removing the artifacts outside that region. . . 44
4.2 Schematic of the STERN components. (a) ST module and classifier within the STN. Class 0 is considered a normal instance, while class 1 is an abnormal instance. (b) Components of the ST module. The localization network predicts the transformation parameters $\theta$, the grid generator uses an inverse transform $T_\theta$ to calculate the corresponding position in the input image, and the sampler samples the pixel value from that input image using bilinear interpolation to produce the transformed output. Rocha et al. (2022). . . . . 47
4.3 Experimental setup baselines. Class 0 is considered a normal instance, while class 1 is an abnormal instance. . . . . 51
4.4 Example of ROI selected by the YOLO (red line) and the STERN (blue dashed line) models in an original radiograph before the cropping and transformation, respectively. The four horizontal (H) and vertical (V) margins of each approach are marked using the same color scheme, and the green cross marks the center of the image. . . . . 53
4.5 Example of vertical projections for an original CXR, where the horizontal axis corresponds to the column index and the vertical axis corresponds to the average intensity per image column. . . . . 53
4.6 Example of ROI BBs before and after transforming an image with the STERN approach. The BBs and corresponding centers are marked in red, and the center of the image is marked with a green cross. . . . . 53

4.7 Visual comparison between the ROI selected by the proposed method and the two baselines. Each row displays an example of an original, cropped (YOLO), and transformed (STERN) images, respectively. The proposed method focuses on the vertical delimitation of the ROI. . . . . 56
4.8 Example of original radiographs with regions selected by the YOLO (red line) and STERN (blue dashed line) models before cropping and transformation. The green cross marks the center of the image. . . . . 57
4.9 Vertical projections of both approaches, representing the mean image intensity (vertical axis) per image column (horizontal axis). Sectors I and III correspond to the positioning of the lungs, Sector II corresponds to the positioning of the spine and sternum, while the black arrows indicate the extremities of the original image. 57
4.10 Example of original and transformed (STERN) images, respectively, with the corresponding ROI BBs as reference. These BBs and their centers are marked in red, and the center of the image is marked with a green cross. . . . . 58
4.11 STERN transformations before and after the fine-tuning stage. . . . . 60
4.12 True positive and false negative rates per method and finding. . . . . 61
4.13 Grad-CAM examples of cropped (YOLO) and transformed (STERN) images, respectively. . . . . 62

5.1 Example of two TP instances for atelectasis, in which the model activates on the correct and incorrect regions, respectively. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability. . . . . 68
5.2 Confidence score calculation based on the rescaling of the predicted probability (*y_pred*), minding the hypothetical class-specific threshold. . . . . 69
5.3 Confident-CAM algorithm workflow. Example for a FP cardiomegaly prediction (*y_pred*). . . . . 70
5.4 Number of instances per pathology in the ChestX-ray14 dataset (left), and their respective pathology correlation matrix (right, Chen et al., 2019a). . . . . 71
5.5 Schematic of the three possible Grad-CAM scenarios and corresponding "rectified" maps, named Confident-CAMs. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability. . . . . 75
5.6 Confident-CAM results for several FP predictions with unsatisfactory Grad-CAMs. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability. . . . . 76
5.7 Confident-CAM results for several TP predictions with satisfactory Grad-CAMs. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability. . . . . 77
5.8 Confident-CAM results for several TP predictions with unsatisfactory Grad-CAMs. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability. . . . . 78
5.9 Inadequate Confident-CAM results for a FP and a TP prediction. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability. . . . 79
5.10 Example of TP instances with low and high confidence scores, (a-d) and (e-h) respectively, and their corresponding Grad-CAMs. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability. . . . . 80
5.11 Inadequate results for TP instances. High confidence scores with low overlap in (a-b), and vice versa in (c-d). The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability. . . . . 81

6.1 Frequency of 0, 1, or more co-occurring pathologies in the original distribution of the ChestX-ray14 dataset. . . . . 87
6.2 Framework of the CLARE-XR methodology. . . . . 90
6.3 Schematic of the proposed autoencoder, exemplified for an input vector $L$ representative of a pleural lesion. The encoder yields the embedding $[x,y]$ and the decoder the reconstructed vector $\hat{L}$. . . . . 91
6.4 Reference label embeddings obtained by the Phase I encoder. The *No Finding* and single-class GT coordinates are plotted in colored dots, while the remaining multi-label combination coordinates are marked with black crosses. . . . . 96
6.5 Predicted embedding coordinates obtained by the Phase II regressor for different combinations. The GT references are marked with black crosses. All predicted coordinates whose GT matches the combination in the title are marked with red dots. 97
6.6 Inference image examples and their respective predicted coordinates in the reference space (red plus sign). The *No Finding* and single-class references are plotted in colored dots, while the remaining multi-label combination coordinates are marked with black crosses. The corresponding GT reference is highlighted in red. 100
6.7 Frequency of instances per error category. . . . . 102
6.8 Reference label embeddings obtained by the Phase I PCA. The *No Finding* and single-class GT coordinates are plotted in colored dots, while the remaining multi-label combination coordinates are marked with black crosses. . . . . 103
6.9 Inference image examples and their respective predicted coordinates in the reference space obtained with PCA (red plus sign). The *No Finding* and single-class references are plotted in colored dots, while the remaining multi-label combination coordinates are marked with black crosses. The corresponding GT reference is highlighted in red. . . . . 106
6.10 Inference image and its respective predicted coordinate (red plus sign) in the reference space obtained with the autoencoder and PCA, respectively. The *No Finding* and single-class references are plotted in colored dots, while the remaining multi-label combination coordinates are marked with black crosses. The corresponding GT reference is highlighted in red. . . . . 107
6.11 Embedding coordinates obtained by the Phase II regressor for the training instances annotated with *CaPl* (red dots). The GT reference is marked with the black cross, and the predicted coordinate for the current inference image is marked with a black plus sign. . . . . 108
6.12 Visual explanations for the inference image examples in Figure 6.6, represented on the left. . . . . 109

A.1 Examples of ROC and PR curves, respectively. . . . . 137

B.1 ROC curves of the multi-label baseline classifier. . . . . 140
B.2 PR curves of the multi-label baseline classifier. . . . . 140
B.3 Predicted embedding coordinates obtained by the Phase II regressor for under-represented combinations. The GT references are marked with black crosses. All predicted coordinates whose GT matches the combination in the title are marked with red dots. . . . . 141
B.4 Predicted embedding coordinates obtained by the Phase II regressor for intermediately-represented combinations. The GT references are marked with black crosses. All predicted coordinates whose GT matches the combination in the title are marked with red dots. . . . . 142

B.5 Predicted embedding coordinates obtained by the Phase II regressor for highly-represented combinations. The GT references are marked with black crosses. All predicted coordinates whose GT matches the combination in the title are marked with red dots. . . . . 143
B.6 Frequency of instances per error category for under-represented combinations. . . 144
B.7 Frequency of instances per error category for intermediately-represented combinations. . . . . 145
B.8 Frequency of instances per error category for highly-represented combinations. . 146

# List of Tables


2.1 Description of multiple-pathology CXR datasets. In the CXR view column, *F* stands for Frontal and *L* stands for Lateral. In the text-mined labels column, *Y* stands for Yes and *N* stands for No, indicating if the majority of the dataset's annotations were obtained using NLP on radiology reports. . . . . 16
2.2 Advantages of the ChestX-ray14 and CheXpert datasets. . . . . 17
2.3 Comparison of the ChestX-ray8, ChestX-ray14, and CheXpert annotations, with the respective number of samples in each class. Adapted from Wang et al. (2017a) and Irvin et al. (2019). . . . . 18

3.1 Classification results of the ResNet50 and DenseNet121 models per pathology class (mean and standard deviation in the last row). . . . . 35
3.2 Mean overlap between Grad-CAM activations and lesion locations, for all instances of the independent test set predicted as positive. . . . . 37

4.1 Radiographic projection and view distribution of the CheXpert frontal subset. The 17 images falsely annotated as left or right lateral were also annotated as frontal radiographs, and for that reason were kept in this study's set. . . . . 50
4.2 Label distribution of the CheXpert frontal subset. . . . . 50
4.3 Summary of the proposed criteria to evaluate the relevance of the STERN-selected ROI. . . . . 54
4.4 Mean length of the margins of the selected regions in the original radiographs, before cropping and transforming. Measured in pixels. Each margin is defined in Figure 4.4. . . . . 58
4.5 Mean thoracic BB center coordinates before and after the STERN transformation. The image center coordinates are (128,128) in both cases. Measured in pixels. . . 59
4.7 Computational cost analysis of the three methodologies. Inference time is calculated per batch of 8 images, in an Nvidia GeForce GTX 1080 GPU (8 GB). . . . 60
4.6 Classification results of the baselines and STERN method. The best result for each metric is shown in bold. . . . . 62

5.1 Classification results of the ResNet50 model. . . . . 72
5.2 Comparison of AUC scores for state-of-the-art methods evaluated on the ChestX-ray14 dataset. The reported scores were obtained minding different evaluation strategies/conditions across studies. . . . . 73
5.3 Mean confidence score for instances predicted as positive during inference. . . . 82
5.4 Confidence score per class for the TP instances of the independent test set, containing a total of 848 BBs (TPs). All the available BBs of the ChestX-ray14 dataset are described in the last column. The final row provides the mean and STD values across these test samples. . . . . 82

6.1 Number of instances per class, annotated with one, two, or three types of pathology (C1, C2, and C3, respectively), in the selected ChestX-ray14 subset. . . . . 88
6.2 Number of instances per combination in the selected ChestX-ray14 subset. . . . . 89
6.3 Classification results of the multi-label baseline classifier. Results obtained for all instances predicted as positive for a given class, regardless of whether they are *No Finding*, single-pathology, or multi-label. The precision, recall, and F1-score values were calculated using class-specific optimal thresholds, optimized based on the ROC curve (*B_ROC*) or the PR curve (*B_PR*). . . . . 94
6.4 Comparison with the results of a state-of-the-art multi-label classification method that discloses the precision, recall, and F1-score. This method (*) is trained and tested on CheXpert, while the proposed baseline and CLARE-XR methods are trained and tested on the ChestX-ray14 dataset. *N/A* stands for Not Available. . . 95
6.5 Classification results of the CLARE-XR method. Results obtained for all instances predicted as positive for a given class, regardless of whether they are *No Finding*, single-pathology, or multi-label.The precision, recall, and F1-score metrics were calculated based on the predictions obtained by each rule (*R*). . . . . 98
6.6 Classification results of the baseline, fixing the recall or precision obtained by the best setting of the CLARE-XR methodology (Rule 4, *R4*). . . . . 98
6.7 GT labels and final decision based on the proposed rules, for multiple inference images. . . . . 101
6.8 Classification results of the CLARE-XR method per combination, ordered by their frequency and using Rule 4. *N/A* stands for Not Available. . . . . 102
6.9 Classification results of the CLARE-XR method, using the PCA-generated references. Results obtained for all instances predicted as positive for a given class, regardless of whether they are *No Finding*, single-pathology, or multi-label. *R* stands for Rule. . . . . 104
6.10 GT labels and final decision based on the proposed rules and the PCA encoding, for multiple inference images. . . . . 107

# Abbreviations

| | |
|---|---|
| 2D | Two-Dimensional |
| AI | Artificial Intelligence |
| AP | Anterior-Posterior |
| AUC | Area Under the ROC Curve |
| BB | Bounding Box |
| BCE | Binary Cross Entropy |
| CAD | Computer-Aided Diagnosis |
| CAM | Class Activation Map |
| CNN | Convolutional Neural Network |
| CXR | Chest X-Ray |
| CT | Computed Tomography |
| DL | Deep Learning |
| EBPG | Energy-Based Pointing Game |
| FN | False Negative |
| FP | False Positive |
| GAP | Global Average Pooling |
| GCN | Graph Convolutional Network |
| GDPR | General Data Protection Regulations |
| Grad-CAM | Gradient-weighted Class Activation Map |
| GT | Ground Truth |
| IOU | Intersection Over Union |
| LIME | Local Interpretable Model-agnostic Explanation |
| ML | Machine Learning |
| MSE | Mean Squared Error |
| NIH | National Institutes of Health |
| NLP | Natural Language Processing |
| PA | Posterior-Anterior |
| PCA | Principal Component Analysis |
| PR | Precision-Recall |
| ReLU | Rectified Linear Unit |
| ROC | Receiver-Operating Characteristic |
| ROI | Region Of Interest |
| SA-IOU | Size-Adaptive Intersection Over Union |
| SHAP | SHapley Additive exPlanation |
| STD | STandard Deviation |
| STN | Spatial Transformer Network |
| TN | True Negative |
| TP | True Positive |
| ViT | Vision Transformer |
| XAI | eXplainable Artificial Intelligence |

# Chapter 1

# Introduction

Thoracic pathologies pose substantial global health challenges due to their high prevalence and associated high mortality rates. These conditions place a significant burden on health systems, particularly in regions with limited access to healthcare resources (American Cancer Society, 2024). Continued efforts in prevention, early diagnosis, monitoring, and treatment are essential to reduce their worldwide impact. The use of chest radiography for such purposes, as well as the application of Artificial Intelligence (AI) for improving and optimizing the radiographs' analysis, are explored next. This introductory chapter motivates the PhD thesis and defines its objectives and contributions.

## 1.1 Thesis Motivation

Among the most popular medical imaging exams, the Chest X-Ray (CXR) is frequently requested by healthcare professionals to assess the presence of thoracic diseases, considering that cardiothoracic and pulmonary abnormalities are one of the leading causes of morbidity and mortality worldwide (Wang et al., 2016). CXR images, although less sensitive than Computed Tomography (CT), are widely requested due to their low-cost non-invasive nature (Garcia Santa Cruz et al., 2021). Nevertheless, the thorough analysis of CXR images is time-consuming, and their interpretation may be dubious even for expert radiologists (Shaw et al., 1990). In fact, Chassagnon et al. (2020) indicates that this analysis is prone to a higher detection error rate due to the low contrast between the represented lesions and surrounding structures, and the overall superposition of bone tissue. For this reason, the incorporation of Computer-Aided Diagnosis (CAD) systems for thoracic imaging during clinical practice has become an attractive solution to boost productivity and efficiency while interpreting these exams.

CAD seeks to assist clinical specialists in their daily tasks by providing a second opinion regarding a potential pathological finding or risk factor. These algorithms generally encompass multiple key steps for data processing, feature analysis, and classification (Shiraishi et al., 2000), and even though they can be applied to multi-modal data, they are often developed specifically

for medical image interpretation. Considering the recurring need to assess several types of thoracic pathologies, Deep Learning (DL) based systems have been preferred over traditional machine learning approaches, following the advances in computational capabilities and the increasing availability of medical datasets. This way, the data-driven nature of DL has proved to achieve great performance for multi-disease detection and classification tasks, being a preliminary diagnostic tool that reduces the physicians' workload. The detection and classification of cardiothoracic and pulmonary abnormalities often resort to Convolutional Neural Networks (CNNs), as a result of their ability to handle data with strong spatial relationships. These networks have been capable of matching or even exceeding human performance in multiple medical-related tasks, namely pneumonia detection (Rajpurkar et al., 2017), diabetic retinopathy detection (Ting et al., 2017), and skin cancer classification (Esteva et al., 2017).

DL can have a beneficial impact in radiology, not only by assisting the daily clinical practice, but also by contributing to the development of the radiologists' skills, and therefore help them achieve higher reliability throughout the years (Yasaka and Abe, 2018). Nevertheless, DL also poses certain limitations, enumerated in Lee et al. (2017), starting with the legal and ethical concerns that arise from employing clinical imaging data for the development and commercialization of CAD schemes. In a more practical manner, another relevant limitation of DL is its high dependency on the quantity and quality of training data, considering the risk of overfitting to a specific data distribution. Additionally, the opacity of DL algorithms has increasingly become a central topic of discussion (Dhar et al., 2023). The lack of understanding of the DL algorithms' inner working principles grants them a black-box nature, in a way that it is unclear which data features influence the model's performance. As such, there is no straightforward explanation of the reasoning behind the network's output, which deeply undermines its reliability or trust on behalf of clinical experts, and thus impairs its integration in their daily workflow.

To promote the adoption of such algorithms during clinical practice, strategies must be implemented to justify the predictions, by providing explanations to the physicians as a means to clarify how the model works (Rudin, 2019). This is the foundation of eXplainable Artificial Intelligence (XAI) in Medicine, where a DL-based diagnosis system must be transparent, understandable, and explainable to gain not only the physicians' trust, but also the regulators and patients'. Medical professionals can use XAI tools to rationalize specific diagnoses, allowing them to make informed decisions. For example, XAI can assist in interpreting medical images by highlighting key regions and explaining the factors that influence the diagnosis, as exemplified in Figure 1.1. Figure 1.2 illustrates the overall impact of XAI implementation from the perspective of clinicians as end-users. In the first scenario, a black-box system leaves clinicians uncertain about how and why a particular decision was made. In contrast, the second scenario demonstrates how XAI provides transparency by explaining the system's reasoning, enabling clinicians to better understand the decision-making process. This increased transparency helps doctors validate AI-driven diagnoses, ultimately leading to better patient outcomes.

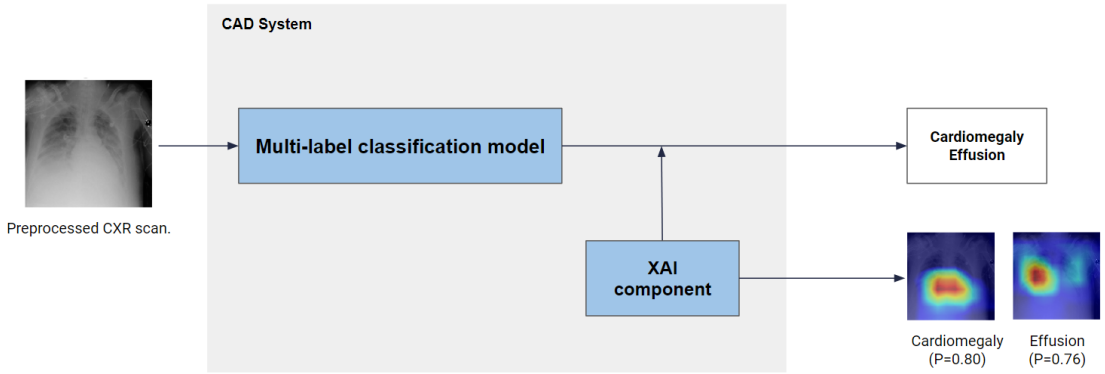



Figure 1.1: A typical XAI framework analyzes each CXR image to generate a final diagnosis, such as cardiomegaly and effusion in this example, and provides heat map-based explanations that highlight regions with higher activation values, indicating areas of the image most likely associated with the identified findings. Each prediction has an associated probability (*P*).

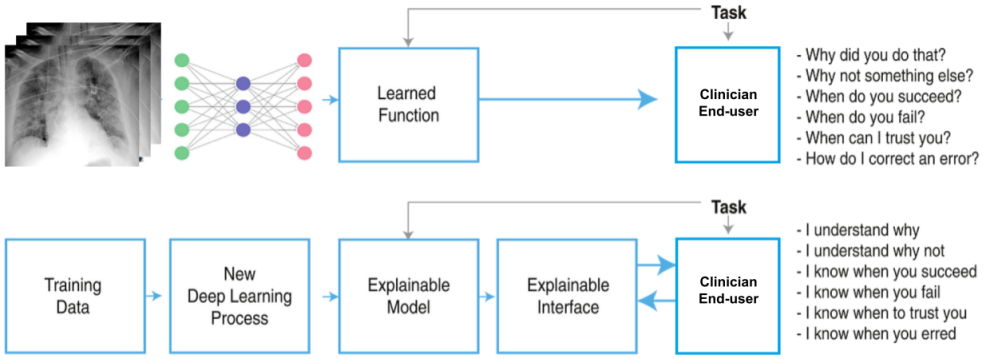



Figure 1.2: Impact of XAI implementation on clinician end-users. Adapted from Lunkad (2023).

## 1.2 Thesis Objectives and Contributions

Following the reported success of DL in medical-related computer vision tasks, the present PhD thesis focuses on the development and implementation of DL methods towards an innovative CAD system for the screening of thoracic pathologies, whose findings are essential to determine the prognosis and/or establish the most suitable treatment when analyzed in a clinical setting. More specifically, this research work is centered on the development of automated CNN-based frameworks to identify pathological findings represented in CXR images. The main goal of the proposed workflow is to optimize the diagnostic pipeline for thoracic pathology, with emphasis on abnormality detection and multi-label pathology classification, individually. This is accomplished through two distinct classification models, each designed to automatically extract relevant image

features. The first model identifies the presence of general abnormality, while the second categorizes any detected findings into specific pathologies. At the same time, it is intended to enhance the performance (and ultimately, usability) of these DL methods by enriching the algorithms with the capability of justifying their outcomes, alerting the specialists to anatomic regions that the models consider abnormal. This translates to the implementation of XAI strategies that are suitable for the medical image analysis domain and complement the performance of the developed methodologies.

This thesis addresses several key issues in this domain, specifically the bias introduced by image artifacts and the shortcomings of state-of-the-art visual XAI methods, which significantly impact the quality of the generated explanations (Dhar et al., 2023). These topics will be discussed thoroughly later in the document. In sum, this thesis research work aims to:

a) Improve model performance and consequent visual explanation maps by eliminating bias-inducing image artifacts on the outer region of the radiographs.

b) Enhance the interpretation of the current state-of-the-art visual explanation method during clinical practice to alert to possibly wrong predictions and/or explanations.

c) Address the limitations of the current state-of-the-art visual explanation method by developing an innovative alternative.

Consequently, the resulting contributions are outlined in Figure 1.3 and encompass the development of:

a) A spatial transformer network with an end-to-end attention mechanism for automatically focusing on the thoracic region of interest during abnormality detection.

b) A refined XAI map approach that identifies likely inaccurate explanations and introduces a confidence score metric to quantify the model's conviction in its class predictions.

c) A data-centric regression-based approach for multi-label classification of CXR images, incorporating an interpretable rule set for more transparent decision-making and a visual explainability method based on automatic retrieval of similar images.

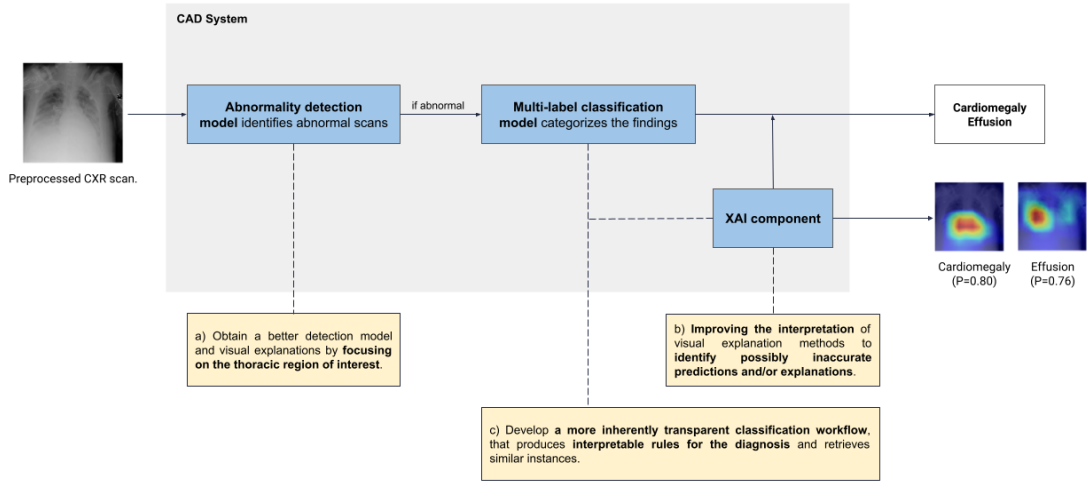



Figure 1.3: The main tasks and respective contributions of the present PhD work, contextualized in a typical CAD workflow.

## 1.3 Thesis Outline

The document comprises a total of seven chapters. This introductory chapter outlines the research focus, centered on thoracic abnormality detection and disease classification through CXR analysis. The motivation for automating these tasks via CAD systems is discussed, highlighting their importance in improving diagnostic accuracy and efficiency. The goal of this work is clearly defined: to develop methodologies that address these specific tasks and mitigate the inherent limitations of DL by integrating XAI into the workflow, thereby enhancing the interpretability of the models' predictions. Chapter 2 provides an overview of the clinical background and the use of DL for CXR image analysis, detailing the most relevant publicly available datasets.

Chapter 3 discusses the reasoning behind XAI for DL models, presenting the motivation behind this research topic and its intrinsic taxonomy. Furthermore, this chapter also emphasizes the significance of XAI in the clinical setting, namely its requirements and the potential risks associated with its use. Chapters 4, 5, and 6 detail the research conducted during the PhD to enhance both the performance of classification models and the quality of the explanations they generate. Finally, the main conclusions from this work are drawn in Chapter 7.

# Chapter 2

# Chest X-Ray Image Analysis

**The content of this chapter results from the following publication. The author of this thesis has the main contribution to the conception of the study.**
**Rocha, J.**, Mendonça, A. M., and Campilho, A. (2021). A Review on Deep Learning Methods for Chest X-Ray based Abnormality Detection and Thoracic Pathology Classification. U. Porto Journal of Engineering, 7(4), 16-32.

Contextualizing the research work is essential to address pressing knowledge gaps. This chapter provides an overview of chest anatomy and radiography, establishing the clinical background for the conducted experiments. The fundamentals of DL for CXR-based diagnosis are explored, with a particular focus on abnormality detection and multi-label classification of thoracic pathologies. This review is crucial for understanding how these tasks can be applied in an automated thoracic pathology screening process, resorting to some of the most relevant publicly available datasets.

## 2.1 Clinical Background

The thorax is the region of the body situated between the neck and the abdomen, enclosed by the 12 pairs of ribs and thoracic vertebrae (Kudzinskas and Callahan, 2023). It is supported by a complex network of ligaments and muscles, all of which play a crucial role in essential functions such as breathing and movement. Multiple vital organs are located in the thoracic region, notably the lungs/pleurae and the heart, but also the thymus gland and the breasts. Since the heart is located here, the major vessels associated with it are also present, including the pulmonary arteries and veins, the superior vena cava, and the aorta with some of its proximal branches. Figure 2.1 illustrates the positioning of these organs within the thoracic cavity.

The pleura is a serous membrane that folds back on itself to form a two-layered pleural sac (Mahabadi et al., 2018). The outer layer, known as the parietal pleura, attaches to the chest wall, while the inner layer, called the visceral pleura, covers the lungs, blood vessels, nerves, and bronchi. The right and left pleural cavities are anatomically separate. Serous fluid within

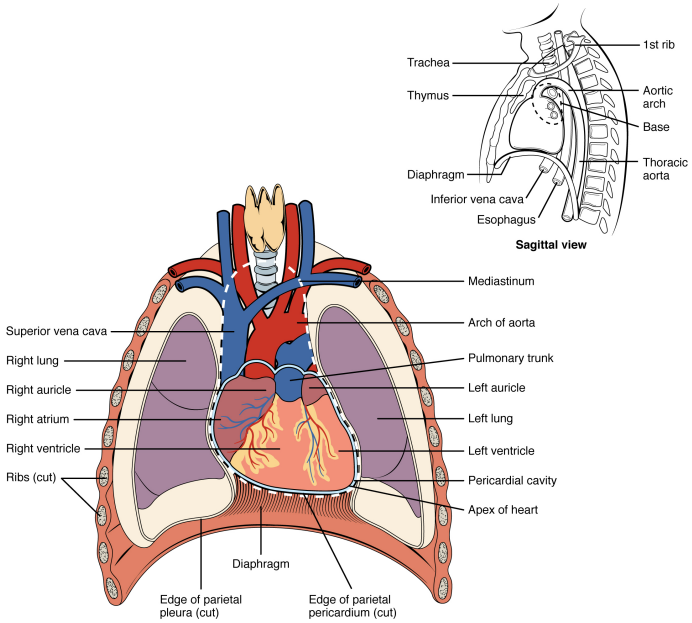



Figure 2.1: Illustration of the thoracic cavity and its constituents. OpenStax (2024).

the pleura continuously lubricates the surfaces, allowing them to slide smoothly over each other during lung inflation and deflation. The thoracic mediastinum is the compartment extending along the length of the thoracic cavity between the pleural sacs of the lungs. It reaches from the thoracic opening to the superior surface of the diaphragm. The mediastinum protects many vital structures, including the heart, great vessels, trachea, and key nerves.

The lungs are bilateral organs located on either side of the mediastinum within the thoracic cavity, and surrounded by the two pleural layers (Chaudhry et al., 2017). Each lung has an apex above the first rib, three borders (anterior, posterior, and inferior), and three surfaces (costal, medial, and diaphragmatic). The right lung has three lobes, while the left lung has two, accommodating the space taken by the heart. Air flows through the trachea, bronchi, bronchioles, and alveoli to reach and exit the lungs, given their uttermost importance in the respiratory system. On the other hand, the heart, located centrally in the middle mediastinum, is the core of the circulatory system (Rehman and Rehman, 2017). It comprises four chambers (two atria and two ventricles), and four valves (tricuspid, pulmonary, mitral, and aortic) that ensure blood flows in one direction only. The right side of the heart pumps blood into the pulmonary circulation, whilst the left side of the heart pumps blood into the systemic circulation. The heart is also surrounded by a protective fluid-filled sac, the pericardium.

Thoracic pathologies encompass a wide range of diseases and conditions that affect the structures within the thorax, such as the ones described above. The incidence of these conditions varies widely depending on the specific disease, geographic region, and population demographics. Their potential severity and often life-threatening impact dictate the importance of raising awareness regarding the associated risk factors and of promoting preventive measures. For instance, lung

cancer is the leading cause of cancer-related deaths worldwide, accounting for approximately 1.8 million deaths each year. This represents nearly 18.7% of all cancer fatalities. The incidence of lung cancer continues to rise, with an estimated 2.2 million new cases diagnosed globally each year (World Health Organization, 2024; American Cancer Society, 2024). Pneumonia, on the other hand, remains the largest cause of death among children under five, accounting for 15% of all deaths in this age group. It is also a significant cause of morbidity and mortality in the elderly, with millions of cases reported annually (World Health Organization, 2022). Heart conditions represent another major global health concern. Coronary heart disease and heart failure are common examples, affecting around 200 million people and 64 million people worldwide, respectively (British Heart Foundation, 2024). The latter is the leading cause of hospitalization and healthcare expenditure for people over 65 years old (Zehnder et al., 2022).

Chest radiography has proven to be essential in the prevention, diagnosis, and treatment of thoracic pathologies, as it offers a fast and accessible means of visualizing the internal structures of the thorax. The procedure's speed makes it particularly valuable both in emergency and routine situations, while its widespread availability in medical facilities ensures broad access. The simplicity of radiography, requiring minimal patient preparation and cooperation, is especially advantageous for those patients who may have difficulty remaining upright and still. Additionally, compared to CT radiographs, radiography typically involves lower radiation doses, which helps to minimize the risks associated with repeated imaging exams (Garcia Santa Cruz et al., 2021).

Unlike CT radiographs, which provide 3D representations, radiography produces 2D images. The procedure begins with an X-ray machine that emits a controlled beam of this high-energy electromagnetic radiation capable of penetrating the body. The patient is positioned between the X-ray source and a detector or photographic film, as illustrated in Figure 2.2. As the X-ray beam passes through the body, it is absorbed at different rates by various tissues. Dense tissues, such as bones, absorb more X-rays, while less dense tissues, like muscles and organs, absorb fewer X-rays. The X-rays that pass through the body reach the detector placed on the opposite side, which captures them and converts them into an image (Yaffe and Rowlands, 1997). The areas that absorbed more X-rays appear white or light gray on the image, while areas that absorbed fewer X-rays appear darker gray or black. The resulting image is reviewed by radiologists or medical professionals, who analyze the contrast between different tissues to identify any abnormalities, fractures, infections, or other conditions.

Note that different views and projections can be considered when acquiring a CXR image (Tafti and Byerly, 2022). Frontal and lateral views of the chest, as depicted in Figure 2.3(a), enable radiologists to more accurately assess the shape and volume of the organs, as well as the precise location of any abnormalities. In a posterior-anterior (PA) radiograph, the patient is positioned upright, facing the detector as illustrated in Figure 2.2. In contrast, an anterior-posterior (AP) projection is preferred when the patient is unable to stand and must remain laid down, with their back positioned against the detector. As a consequence, AP projections distort the representation of the anatomical structures, such as the clavicle and scapulae, which may overlap the lung fields and obscure details critical for evaluation. Additionally, AP projections tend to make the heart

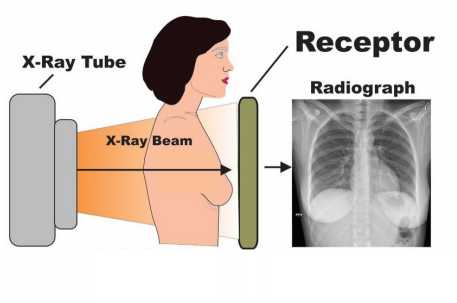
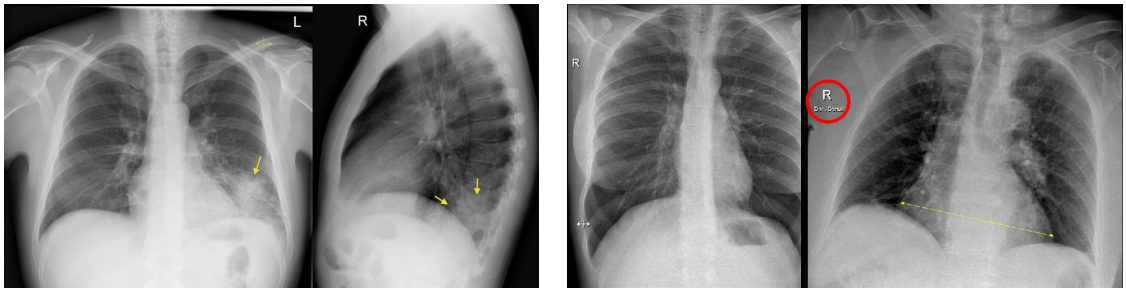



Figure 2.2: Schematic of a digital radiography system. Adapted from Sprawls Jr (1993).

appear enlarged and the mediastinum wider (yellow arrow in Figure 2.3(b)). These changes can complicate the interpretation of findings. Overall, the quality of AP images tends to be lower compared to PA images, considered the "gold standard" for radiographic imaging.




(a) (b)

Figure 2.3: CXR image with (a) frontal and lateral views, and (b) posterior-anterior and anterior-posterior projections, respectively. St Vincent's University Hospital (2024); Algarve Hospital and University Centre (2018).

As previously mentioned, CXR exams are very effective for visualizing structures with varying densities, thus portraying a wide range of findings. Irvin et al. (2019) proposed to group these findings using the following simplified categories. These categories are exemplified in Figure 2.4, and will be further addressed in Section 2.3 to discuss the available public CXR datasets.

a) **Enlarged Cardiomediastinum**: an abnormal increase of the chest area where the heart and the mediastinum are located. The specific enlargement of the heart is referred to as cardiomegaly, which can be indicative of various conditions, including cardiomyopathy. For this reason, an enlarged cardiomediastinum is a radiological finding that requires further investigation to determine the underlying cause.

b) **Lung Opacity**: refers to the hazy gray findings in the pulmonary region. The lower intensity of the lungs matches their expected lower density. However, the presence of higher-intensity gray regions indicates that certain findings are taking up lung space (e.g. fluid, nodule/mass). Further investigation may associate such findings with pneumonia infection, infiltration, atelectasis, etc.

c) **Pleural**: the pleura is only visible in a CXR when an abnormality is present. These are often subtle and may vary in appearance. Some pleural conditions cause pleural thickening, while others lead to fluid or air accumulation in the pleural spaces (effusion and pneumothorax, respectively).

d) **Others**: other conditions not considered in the previous categories (e.g. bone fractures).

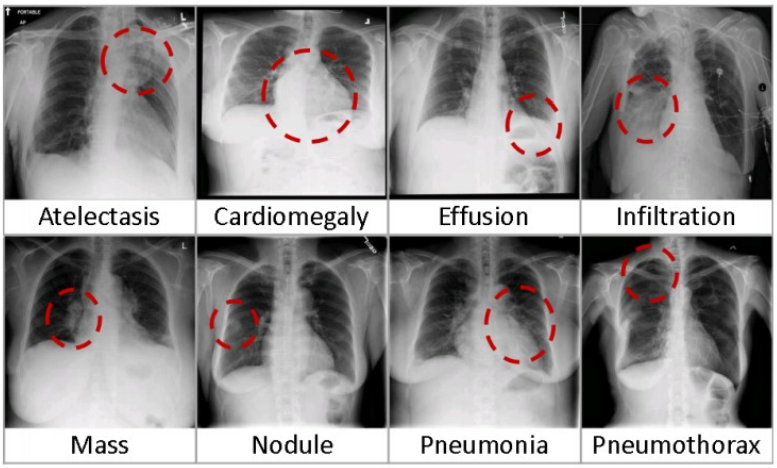



Figure 2.4: Eight common thoracic conditions represented in CXR images. Wang et al. (2017a).

## 2.2 Advances in Computer-Aided Diagnosis

While CXR images are easy to acquire, their interpretation is not as fast or as straightforward. The anatomical complexity of the thoracic region and its overlapping structures can obscure possible lesions, especially when dealing with low-contrasted images (Chassagnon et al., 2020; Zhang et al., 2023b). The appearance of normal and abnormal structures often varies significantly between patients due to multiple factors (e.g. age, lifestyle, anatomical features), and even the appearance of certain diseases can change rapidly within the same patient, requiring careful consideration of the clinical context. This variability can lead to different interpretations of the same image by different radiologists, resulting in diagnostic inconsistencies. Less experienced radiologists are more likely to overlook subtle signs of disease or misinterpreting normal variants as pathological (Shaw et al., 1990; Onder et al., 2021). Furthermore, their analysis is highly impacted by technical factors, as poor-quality X-rays resulting from improper patient positioning, movement during the acquisition, or suboptimal exposure promote misinterpretation. Not only that, the inherent resolution limitations of X-ray technology may also hinder the detection of very small lesions or fine details. These challenges evidence the need to make the diagnostic process more efficient for radiologists.

CAD systems have been extensively researched and are perceived as a key tool in the medical field, complementing human expertise and improving the overall quality of the healthcare workflow. A typical CAD system starts by receiving a CXR image and executes a series of tasks to reach a final decision, as illustrated in Figure 2.5. After acquiring an image, developers usually preprocess it to enhance its quality and better distinguish the represented structures, e.g. by increasing image contrast. Region Of Interest (ROI) selection can be applied to ensure the system focuses on the most relevant parts of the image, likely to contain useful information for the diagnosis, thus eliminating distracting background information. This step translates to an object detection or segmentation model, which crops the image with a bounding box or delimits specific regions with a more detailed mask of the structures of interest, respectively. For example, a CAD system designed specifically to detect pulmonary lesions may focus exclusively on the lungs to improve its accuracy. The most relevant features for the diagnosis are extracted from the resulting image, leading to the final classification stage of the process.

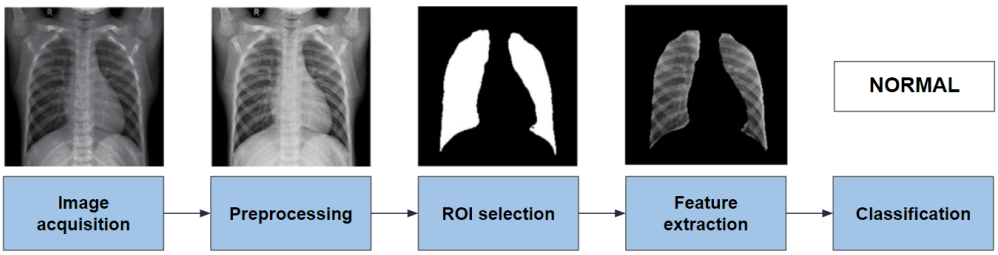



Figure 2.5: Typical CAD workflow for medical image analysis. Adapted from Nillmani et al. (2022).

Early work on image analysis manually designed each of these steps, guided by domain knowledge. Techniques such as edge detection, texture analysis, and thresholding were commonly used to identify patterns and structures within images. After feature extraction, decision-making was typically based on rule-driven logic or basic statistical models. However, these methods struggled with data variability and noise, limiting their ability to generalize to new unseen cases. Additionally, they demanded significant domain-specific knowledge and required fine-tuning for each unique task, making them less flexible and adaptable to novel problems without substantial manual intervention. With the emergence of AI, the progress of CAD tools has been exponential. In particular, the emergence of Machine Learning (ML) and DL has granted such systems the ability to analyze large datasets, make predictions, and provide insights that were previously more complex to obtain. Their impact continues to grow as data availability and computing power continue to increase, largely surpassing traditional image analysis in terms of accuracy and adaptability, especially in fields like medical imaging.

ML techniques automatically identify patterns in data to make predictions. DL is a specialized subset of ML that employs deep neural networks to automatically uncover and learn these hierarchical patterns in data. Although DL models demand larger datasets and computational resources, they can achieve exceptional accuracy in several applications of image recognition (Sharifani and

Amini, 2023). One of the most significant differences between both methods lies in the feature extraction process. In traditional ML, feature extraction still involves manually selecting and designing relevant features from raw data to maximize model performance. This process requires domain expertise and often involves techniques like statistical analysis or dimensionality reduction to identify the most important attributes, which are then provided to the model. On the other hand, feature extraction is largely automated in DL (Saba et al., 2019), whose models learn hierarchical features directly from raw data through multiple layers of neurons. The model fulfills both the feature extraction and classification steps of a CAD system by discovering and extracting complex patterns and features, reducing the need for manual feature engineering. This allows DL to handle high-dimensional data with less human intervention. Feng et al. (2019) extends this conclusion to radiography image analysis, highlighting that DL achieves breakthrough performance in complex chest imaging tasks because it solves most problems that traditional ML is unable to address effectively.

There is no doubt that DL offers exciting opportunities for radiologists to improve their diagnostic efficiency, with certain authors defending that automated CXR-based abnormality detection has already reached clinically acceptable performance (Feng et al., 2019). This mirrors the growing interest in the CAD research line, as shown in Figure 2.6. While multiple training strategies have been studied, supervised learning remains the most common for image classification in this field (Feng et al., 2019). Here, all available data is fully annotated, and the model is trained by minimizing the discrepancy between its predictions and the ground-truth labels. This discrepancy is quantified using a differentiable loss function, which guides the learning process. During training, the model's parameters, namely its weights, are iteratively adjusted to minimize the loss, thereby improving the accuracy of its predictions. To accelerate this convergence, reduce its computational cost, and make up for the lack of domain-specific training data, researchers often resort to transfer learning. In this approach, the model is initialized with weights learned from a different task, such as the ImageNet natural image analysis challenge (Deng et al., 2009), and fine-tuned on the domain-specific data to reduce the risk of overfitting.

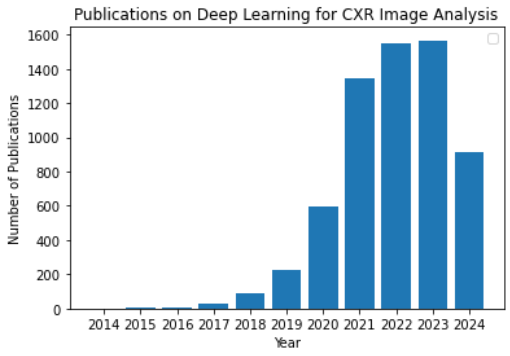



Figure 2.6: Number of publications per year, presenting chest radiography-focused DL applications. Data collected from *Scopus*, using the search criteria `(deep learning) AND (chest OR thoracic) AND (x-ray OR radiography)`. Results for 2024 are only recorded up to September, 2024.

According to McBee et al. (2018), CNNs are the most widely used DL architecture in radiology image classification. Its feature extraction layers use convolutional filters that slide over the input image, performing element-wise multiplications and summing the results to create feature maps. These feature maps capture spatial relationships, which are then processed through non-linear activation functions to highlight the distinctive characteristics of each finding/lesion type, and used to infer the corresponding predictions. The published studies on medical image analysis usually employ architectures that succeeded in the ImageNet challenge. Among these, there are the popular VGG (Simonyan and Zisserman, 2015), ResNet (He et al., 2016), and DenseNet (Huang et al., 2017) architectures, which are described next given their relevance for this PhD thesis.

In particular, the VGG16 is a CNN with 16 layers, known for its simplicity and effectiveness in image classification tasks. Developed by the Visual Geometry Group (VGG), it uses a uniform architecture of small $3 \times 3$ convolutional filters and $2 \times 2$ max-pooling layers, which allows it to capture detailed features in images. VGG16 is widely used for its straightforward design and solid performance on various computer vision benchmarks.

The ResNet50 is a CNN composed of 50 layers, designed for image recognition and classification tasks. It is part of the Residual Networks (ResNet) family, which introduced residual connections or skip connections, developed by Microsoft™ Research. In other deep neural networks, as the number of layers increases, the network often suffers from the vanishing gradient problem, where gradients become extremely small, making it difficult for the network to learn effectively during training. Residual connections address this issue by allowing the network to bypass one or more layers, effectively "skipping" them. These connections add the input of a layer directly to its output, allowing the network to preserve important information and making it easier to train deeper networks. ResNet50 can balance depth, capturing intricate details in images, and computational efficiency, making it practical for real-world applications.

DenseNet121, short for Densely connected convolutional Network, is an architecture with 121 layers that enhances feature propagation. Each layer receives information from all preceding ones, having direct access to the feature maps generated earlier in the network. This dense connectivity encourages feature reuse and also helps mitigate the vanishing gradient problem, allowing for more effective training. By reusing features from earlier layers, DenseNet achieves significant performance with fewer parameters compared to traditional CNNs of similar depth. Specifically, DenseNet121 strikes an optimal balance between depth and computational efficiency, making it a popular choice for image classification, object detection, and segmentation tasks. It outperforms many existing architectures, particularly in medical image analysis.

Introduced by researchers at Google™, Vision Transformers (ViTs) are a recent alternative to CNNs (Dosovitskiy et al., 2021). This novel architecture for image classification leverages the principles of transformer models, originally designed for Natural Language Processing (NLP), to model long-range dependencies within images. Unlike CNNs, which focus on local receptive fields, transformers capture global context more effectively, making them particularly suitable for large-scale images. Instead of processing the entire image as a whole, ViTs divide the image

into fixed-size patches, which are then linearly embedded into vectors. This allows the model to treat these patches similarly to tokens in NLP, facilitating a uniform processing mechanism across the input. The architecture employs a self-attention mechanism that evaluates the significance of different patches when making predictions, thereby enabling the model to understand relationships and dependencies between distant regions within the image. Overall, ViTs consist of a stack of transformer blocks, each comprising multi-head self-attention layers and feed-forward networks. This design grants the architecture flexibility and scalability, allowing it to perform exceptionally well when trained on large datasets. As a result, ViTs have already demonstrated competitive performance compared to CNNs on various benchmarks, including medical image classification tasks, particularly when ample labeled data is available.

In sum, DL models extract meaningful features from CXR images based on various properties, such as pixel intensity and spatial relationships, to identify patterns and make predictions. They analyze gradient changes to define the boundaries between different anatomical structures (e.g., the edge of the diaphragm or the silhouette of the heart), and consequently their shape and size. In turn, these boundaries offer a contextual understanding of the relative position of the structures. When a DL model learns these spatial relationships, it can infer when something looks abnormal (e.g. if one lung appears more opaque than the other, or if the heart is enlarged).

While multiple exercises arise within this domain, abnormality detection and thoracic pathology classification in CXR images are among the most common. Both tasks face challenges common to many DL applications in Medicine. Their performance is highly dependent on the available data, meaning it will learn different patterns depending on the quantity of annotated images, the noise in their annotations, their inter-patient and intra-patient variability, class distribution, etc. These issues hinder the model's ability to learn consistent patterns across similar lesion types. The following section describes the currently available public datasets, exploring their advantages and disadvantages for the mentioned exercises.

## 2.3 Public Chest Radiograph Datasets

The distinctive data-hungry nature of DL has motivated the release of multiple public CXR datasets over the last few years. These can be split into two main groups - those that tackle a specific thoracic pathology, and those that annotate multiple pathologies. Examples of the first group include the Japanese Society of Radiological Technology (JSRT) dataset of lung nodules (Shiraishi et al., 2000), the MC and Shenzhen and the Korean Institute of Tuberculosis (KIT) datasets of tuberculosis (Jaeger et al., 2014; Ryoo and Kim, 2014), and the Kermany dataset of child pneumonia (Kermany et al., 2018). Recently, other collections of data centered specifically on COVID-19 pneumonia have been released, in response to the pandemic that emerged in 2020 (Garcia Santa Cruz et al., 2021). Among these, there are the COVIDx (Wang et al., 2020b), the BIMCV COVID-19 (Vayá et al., 2020), and the COVID-19-AR (Desai et al., 2020).

The focus of this work lies in the datasets that encompass more than a single pathology, enumerated next and summarized in Table 2.1. One of the pioneering cancer screening trials was

conducted by Gohagan et al. (2000), creating the Prostate, Lung, Colorectal, and Ovarian (PLCO) dataset. This dataset includes 185,421 CXR frontal images labeled with 13 classes, collected from 56,071 patients and partially available to the public. Later on, the Open-I Indiana University dataset, published in Demner-Fushman et al. (2016), gathered 8,121 associated frontal images from two large hospitals in the Indiana Network for Patient Care, and addressed the 10 most prevalent conditions observed in 3,996 subjects.

The National Institutes of Health (NIH) released the ChestX-ray8 in Wang et al. (2017a) and compiled 108,948 frontal views belonging to 32,717 unique patients with a total of 8 associated pathologies. This dataset evolved to include 6 more categories, increasing the overall number of frontal CXR images and resulting in ChestX-ray14. It is argued that this version is more representative of the patients' distributions and diagnoses in comparison to the previously mentioned set (Wang et al., 2017b). ChestX-ray14 includes 112,120 images from 30,805 patients, covering 14 different pathologies. As will be further discussed, it is by far the most popular dataset being used in today's research.

Another staple among the most popular CXR datasets is the CheXpert, detailed in Irvin et al. (2019). It comprises 224,316 frontal and lateral chest X-ray images from 65,240 patients at Stanford Hospital. CheXpert is distinctive because it not only recognizes the presence of 12 pathology-related classes, but also the presence of medical support devices and fractures, all described in radiology reports that were released along with the images.

After that, Johnson et al. (2019) released the MIMIC-CXR, making available 377,110 frontal and lateral views of 65,379 patients from the Beth Israel Deaconess Medical Center Emergency Department, along with text radiology reports. The MIMIC-CXR was annotated with the same labels as CheXpert. PadChest became available in Bustos et al. (2020), containing a total of 193 labels applied to 160,868 frontal and lateral CXR images of 67,625 patients. It was collected from the San Juan Hospital, considering radiology reports written in Spanish.

The most recent dataset is the VinDr-CXR, provided by the Vingroup Big Data Institute (VinBigData) and made available through the VinBigData Chest X-ray Abnormalities Detection Kaggle Challenge (Nguyen et al., 2022). It comprises a total of 18,000 frontal CXR images, collected from Hospital 108 (H108) and the Hanoi Medical University Hospital, matching a total of 14 pathologies. Note that only 15,000 out of the 18,000 images have publicly available labels at the present moment. No information is detailed regarding the number of patients included in the dataset, and for that reason, it is assumed that there is a single CXR per patient.

Table 2.1: Description of multiple-pathology CXR datasets. In the CXR view column, *F* stands for Frontal and *L* stands for Lateral. In the text-mined labels column, *Y* stands for Yes and *N* stands for No, indicating if the majority of the dataset's annotations were obtained using NLP on radiology reports.

| Year | Dataset | Patients | Images | Image Format | CXR view | Non-normal Labels | Text-mined Labels |
|---|---|---|---|---|---|---|---|
| 2000 | **PLCO** | 56,071 | 185,421 | TIFF | F | 12 | N |
| 2016 | **Open-I Indiana** | 3,996 | 8,121 | DICOM | F | 10 | N |
| 2017 | **ChestX-ray8** | 32,717 | 108,948 | DICOM | F | 8 | Y |
| 2017 | **ChestX-ray14** | 30,805 | 112,120 | PNG | F | 14 | Y |
| 2019 | **CheXpert** | 65,240 | 224,316 | PNG | F/L | 13 | Y |
| 2019 | **MIMIC-CXR** | 65,379 | 377,110 | DICOM/JPG | F/L | 13 | Y |
| 2020 | **PadChest** | 67,625 | 160,868 | DICOM | F/L | 193 | Y |
| 2021 | **VinDr-CXR** | 18,000 | 18,000 | DICOM | F | 14 | N |

## NIH ChestX-ray14 *vs.* CheXpert

The ChestX-ray14 is undoubtedly the benchmark dataset for CXR-based computer-aided diagnosis (Liz et al., 2023). For this reason, it is important to carefully disclose its strengths and limitations. This section compares the ChestX-ray14 with the CheXpert dataset, as both were used in the PhD research herein reported in this document.

As mentioned previously, the NIH ChestX-ray14 is a large public dataset of 112,120 frontal CXR images, with the corresponding patients' clinical information (gender and age), and radiograph acquisition details (posterior-anterior or anterior-posterior projection). These images were collected between 1992 and 2015 from the National Institutes of Health Clinical Center in Maryland, and the respective labels were mined from the associated radiological reports using NLP regarding the presence of 14 different pathologies (Table 2.1). On the other hand, CheXpert was provided by the Stanford ML Group and consists of 224,316 CXR images collected at the Stanford Hospital in California between 2002 and 2017. This dataset details clinical information regarding each patient's gender and age, how the radiographs were acquired (view and projection), and how their labels were also automatically extracted from radiology reports. Of the total radiographs, 191,027 are frontal-view images. Besides labeling 12 thoracic findings and the presence of support devices, the original dataset also considers a *No Finding* label to capture the absence of any pathology. The pros and cons of both datasets are discussed next and condensed in Table 2.2.

After the ChestX-ray14 labels were text-mined from radiology reports, no expert validation was conducted to ensure that the final annotations accurately corresponded to the image content. Instead, the labeling model was validated on the Open-I Indiana University dataset. This approach raises concerns about the reliability of the annotations, namely regarding how well they reflect the actual pathologies in each image.

Table 2.2: Advantages of the ChestX-ray14 and CheXpert datasets.

| **ChestX-ray14** | **CheXpert** |
|---|---|
| Most frequent dataset in state-of-the-art publications, easing further method comparisons. | Larger than ChestX-ray14. |
| Has spatial annotations. | Has a *No Finding* label that truly represents the absence of any pathology. |
| Does not have many image artifacts in comparison to other datasets with available spatial annotations, such as VinDr-CXR. | Lower labeling error rates. Labeler was more thoroughly evaluated. |
| | Better label structure/hierarchy, also including uncertainty annotations. |

In addition to that, the established labels may not be ideal according to Oakden-Rayner (2019), because they do not provide information on the expected range of abnormalities (e.g. pacemakers and invasive lines). Unlike in CheXpert, the *No Finding* hypothesis simply ensures the absence of the 14 diseases and does not guarantee a healthy observation. It is argued that ChestX-ray14 lacks patient diversity and proper labeling structure and quality, with error rates that range from 30% to 90% in the various findings. The error rates in CheXpert are approximately between 5% and 15%. Even so, Oakden-Rayner (2019) highlights that CheXpert is also flawed, with 10% to 20% of instances labeled as *No Finding* actually showing abnormalities.

Indeed, both datasets face common issues, such as the class imbalance among pathologies, or the questionable relevance of certain proposed annotations. Even so, the same article concludes that CheXpert is larger and overall better for DL implementation, as its labeler was more thoroughly evaluated and proved to produce "labels that accurately reflect the reports".

On the contrary, ChestX-ray14 may be preferred when spatial annotations are necessary, as CheXpert does not provide them. These are essential when analyzing if the model is looking at the correct anatomical regions per instance, as will be further discussed in Chapter 3. To conclude, while the overall lack of diversity might impair the ChestX-ray14's generalization ability in the heterogeneous real-world settings, this dataset has fuelled innovation and is considered highly valuable by the research community. Table 2.3 shows the labels and corresponding distributions of ChestX-ray8, ChestX-ray14, and CheXpert. Their labels are presented according to the groups defined in Irvin et al. (2019).

Table 2.3: Comparison of the ChestX-ray8, ChestX-ray14, and CheXpert annotations, with the respective number of samples in each class. Adapted from Wang et al. (2017a) and Irvin et al. (2019).

| **Label group** | **ChestX-ray8** | | **ChestX-ray14** | |
|---|---|---|---|---|
| **Enlarged Cardiomediastinum** | Cardiomegaly | 1,010 | Cardiomegaly | 2,772 |
| **Lung Opacity** | Infiltration | 10,317 | Infiltration | 19,871 |
| | Atelactasis | 5,789 | Atelactasis | 11,535 |
| | Pneumonia | 1,062 | Pneumonia | 703 |
| | Mass | 6,046 | Mass | 5,746 |
| | Nodule | 1,971 | Nodule | 6,323 |
| | | | Consolidation | 4,667 |
| | | | Fibrosis | 1,686 |
| | | | Edema | 2,303 |
| **Pleural** | Effusion | 6,331 | Effusion | 13,307 |
| | Pneumothorax | 2,793 | Pneumothorax | 5,298 |
| | | | Pleural thickening | 3,385 |
| **Other** | | | Emphysema | 2,516 |
| | | | Hernia | 227 |

| **Label group** | **CheXpert** | |
|---|---|---|
| | No Finding | 16,627 |
| **Enlarged Cardiomediastinum** | Enlarged Cardiomediastinum | 9,020 |
| | Cardiomegaly | 23,002 |
| **Lung Opacity** | Lung Opacity | 92,669 |
| | Atelectasis | 29,333 |
| | Lung Lesion | 6,856 |
| | Pneumonia | 4,576 |
| | Consolidation | 12,730 |
| | Edema | 48,905 |
| **Pleural** | Pleural Other | 2,441 |
| | Effusion | 75,696 |
| | Pneumothorax | 17,313 |
| **Other** | Fracture | 7,270 |
| | Support devices | 105,831 |

## 2.4 Concluding Remarks

The incidence of thoracic diseases is rising globally, with severe conditions such as lung cancer becoming increasingly prevalent. CXRs are crucial for detecting cardiopulmonary diseases, providing a non-invasive way to identify abnormalities in the heart and lungs early on. However, their analysis requires significant time and effort, and consequently, radiologists struggle to interpret the growing volume of radiographs quickly, due to the complexity and high demand.

This opens the door for automated thoracic pathology screening, through the implementation of CAD systems during clinical practice. These are gaining traction in healthcare, complementing human expertise and improving workflow efficiency. The integration of DL allows these systems to analyze vast datasets, make predictions, and provide valuable insights with increasing accuracy, as data and computing power expand. CNNs, such as VGG and ResNet, are widely used in radiology for image analysis, extracting spatial features to make accurate predictions. Supervised learning, often enhanced by transfer learning, helps train these models by minimizing prediction errors and reducing overfitting risks.

Two publicly available datasets are of particular relevance to the scientific community working in this field, ChestX-ray14 and CheXpert, annotating multiple thoracic findings. Both are key datasets for CXR-based diagnosis, with CheXpert offering better validation and lower error rates, making it more suitable for DL applications. However, ChestX-ray14 remains the benchmark, as it provides valuable spatial annotations essential for certain studies, despite its higher labeling errors and limited diversity.

# Chapter 3

# Explainability in Deep Learning

The adoption of DL-based CAD systems is increasingly reliant on their ability to justify their decisions. This chapter builds on the previous discussion by exploring the role of XAI in DL and its applications in Medicine. Describing key concepts and taxonomy is necessary to clarify the significance of this trend and provide a comprehensive overview of state-of-the-art XAI methods. Additionally, this chapter examines the risks associated with these methods and explores possible evaluation strategies within the context of thoracic pathology screening.

## 3.1 Background

Since data is the foundation of any AI algorithm, the growing availability of large public datasets in the medical domain has led to significant improvements in the performance of DL models. Even so, DL presents certain limitations that hinder its broad adoption in clinical decision-making, considering that the model learns from a given set of data and may not fully capture the causal nature of the problem at hand (London, 2019). Yasaka and Abe (2018) enumerates the main disadvantages of this subset of ML methods, the first one being the difficulty of generalization across different datasets. This means that it is challenging to transport and apply the same model e.g. across different hospitals if their patients' data or acquisition equipment is highly variable. Overfitting issues deeply accentuate the barriers to broad DL adoption, given the lack of model robustness, as well as the complexity of clinical modeling tasks and high-stake decisions (Tonekaboni et al., 2019). Simultaneously, the implementation of less accurate or explicit models carries an inherent risk to the patients' health and welfare, who rely on the justifications given by the clinicians during treatment or diagnosis. Medical decision support systems should ideally mimic this process, exploiting causal relationships that will be validated by the experience and understanding of the clinicians to reach a conclusion (Samek et al., 2017). However, the highly non-linear nature of DL currently impairs the presentation of a meaningful explanation regarding its final decision, which leads to lack of transparency, and ultimately trust. DL generates the least interpretable outputs due to its large number of model parameters and is considered a black-box predictor whose internal decision process is unknown or uninterpretable to the observer.

While there is no standard definition of XAI, it generally refers to all research initiatives focused on model transparency and trust. It aims towards more than predictive diagnostic accuracy and imposes the presentation of justifications for the models' recommendations. The need for accountability is one of the several reasons why it is currently implemented, in compliance with recent legislation (Samek et al., 2017; Tjoa and Guan, 2021; Busch et al., 2024). Particularly in radiology, the need for regulation of DL-based technology is well recognized. Establishing protocols for the safe integration of these models into clinical practice has been receiving high priority. For example, in Europe, the General Data Protection Regulations (GDPR) set the "right to explanation," which mandates transparency in the development of automated decision support systems (Goodman and Flaxman, 2017). The ethical considerations surrounding these algorithms have also been explored in Geis et al. (2019). Besides accountability, XAI also promotes system verification and improvement, such as bias detection and debugging, for safety and quality control. Jia et al. (2020) demonstrates the importance of this aspect with a study on DL-based pneumonia detection in CXR images, in which the network was not learning from the anatomical features as expected, but instead simply identifying the hospitals with higher pneumonia incidence to make more accurate predictions. This is a clear example where the interpretation of the model contributes to detecting mistakes early on, that can otherwise lead to severe errors during clinical implementation.

## 3.2 Taxonomy

The majority of published reviews on XAI focuses on taxonomy, ethics, and the need for explanations. To explore the state-of-the-art and proposed methodologies in the following chapters, it is essential to introduce general key concepts of XAI, addressed in the most relevant studies. For example, concepts like causality and causability are often discussed in this field to better understand the implications of explainability. While the first aims to clarify the relationship between input and output, the latter is defined in Holzinger et al. (2019) as the "extent to which an explanation (...) achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use". For this reason, explainability is a complex term that goes beyond simple causality. Furthermore, Broniatowski (2021) also proposes to differentiate it from interpretability by defining the latter as the contextualization of a model's output that is easily related to the system's functional purpose, and the goals, values, and preferences of end users. This implies mapping the output class into an understandable domain example (Stano et al., 2020). On the other hand, an explanation seeks to accurately describe the mechanism that led to a certain output, by disclosing a set of domain features (e.g. image pixels) that contribute to the model's decision. Gilpin et al. (2018) reinforces the difference between both concepts by stating that explainable models are interpretable by default, but the reverse is not necessarily true. Even so, the terms explainability and interpretability are often used interchangeably in the literature, as seen in Allgaier et al. (2023). Note that all these concepts are not exclusive to DL and can be

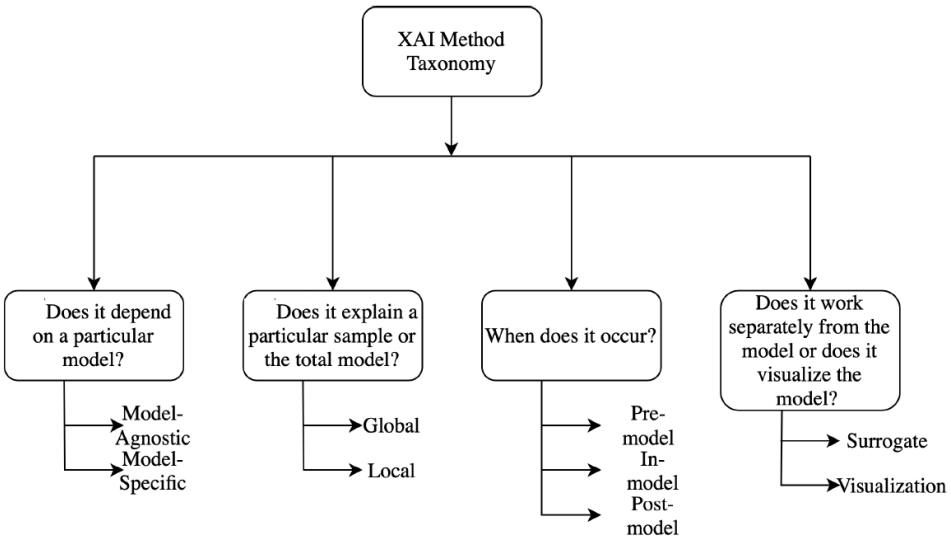



Figure 3.1: Taxonomy of XAI methods. Singh et al. (2020).

considered in multiple stages of the learning process to achieve different objectives, depending on the end user's goals.

The following paragraphs detail the main taxonomic categories for the current XAI methods, depicted in Figure 3.1. Several in-depth studies can be found in the literature providing a more extensive list of concepts and taxonomic definitions (Stiglic et al., 2020; Arrieta et al., 2020; Vilone and Longo, 2021).

### Model-Agnostic *vs.* Model-Specific

Explainability methods can be developed and implemented independently of the model's architecture and without any direct access to its parameters. In this model-agnostic scenario, the methods work independently of the internal structure of the model, using input-output behavior to infer explanations. As such, they are great for benchmarking explanations across models (Holzinger et al., 2019). Alternatively, in a model-specific context, explainability is tailored to specific model types and make use of the internal architecture or learned parameters of the model. For this reason, it may provide more accurate insights into how that specific model makes decisions, and is often more computationally efficient than model-agnostic methods for that target model (Singh et al., 2020).

### Pre-Model *vs.* In-Model *vs.* Post-Model

Explanations can be achieved in different stages of the DL workflow and consequently categorized in straightforward definitions according to the stage in which they are implemented (Figure 3.2). Pre-model methods are independent of the model's architecture and applied before the training

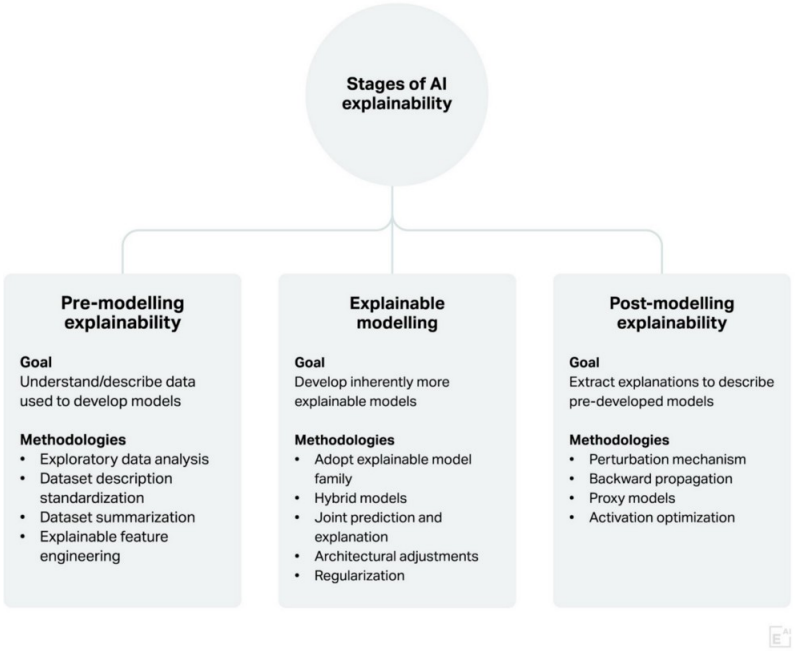



Figure 3.2: The three stages of explainability: pre-model, in-model, and post-model explainability. Khaleghi (2019).

stage, in-model methods are integrated into the model itself (in which case it is referred to as glass-box model), and post-model methods yield valuable insights after the model's training (Singh et al., 2020).

### Global *vs.* Local Explanations

One of the most important concepts in XAI translates to the range of the achieved explanations. Global explainability seeks a set of broad justifications for the model's predictions, by identifying the most significant patterns in the dataset driving those decisions. This way, as stated in Reyes et al. (2020), global explanations are obtained to assess if the learned patterns are coherent with existing domain knowledge, defining a common directive for all cases. These are particularly useful to detect biases in the training data, as they explain the general behavior of the model, e.g. with feature importance (Pereira et al., 2018). In contrast, local explanations evaluate a single outcome of the model and aim to explain that particular instance for a given input. This is typically implemented when the user is interested in the impact of specific features.

### Surrogate *vs.* Visualization

Lastly, a black-box model's decision can be interpreted through surrogate or visualization approaches. Singh et al. (2020) describes surrogate methods as model ensembles used to analyze other black-box models. In other words, a surrogate model refers to a simpler, interpretable model that is used to approximate the behavior of a more complex and opaque model, to provide insights

into how the latter makes decisions. A concrete example would be using a decision tree to approximate a neural network's predictions and reveal which features most significantly impact them. The goal is to benefit from the more transparent decisions of the surrogates to interpret the black-box outcomes, by comparing both. However, if the decision tree underperforms, it may be difficult to fulfill such purpose. Alternatively, visualization methods derive from the black-box model itself and attempt to explain its decision with visual hints, as will be further discussed in this work.

## 3.3 Methodologies

Explainability is gaining increasing attention within the scientific community, as it is recognized that a model's interpretability is shaped by both the model's complexity and the intricacy of the data features (Lipton, 2018). Allgaier et al. (2023) provides an extensive overview of model-specific and model-agnostic methods, with both global and local explanations, but for the sake of brevity, the current section will cover exclusively the most relevant approaches. Holzinger et al. (2019) categorizes interpretability methods for DL architectures into three distinct groups, described below:

a) **Uncertainty**: measures how training data variability is reflected in the model parameters, or how input parameter variation affects the outcome for one particular instance.

b) **Attribution**: focuses on connecting a particular output to the input variables, by analyzing the gradient when changing individual input variables. One possibility is to use activation maps as input to identify parts of images relevant to a prediction.Other widely-used attribution methods identify which features are associated with a specific classification outcome.

c) **Activation maximization**: determines input patterns that lead to maximal activations relating to specific classes in the output layer, which enables the visualization of class prototypes and the assessment of the class-specific properties captured by the model. Prototyping methods can construct samples that fit the representation learned by the classifier for each class (Erhan et al., 2009; Montavon et al., 2018).

Attribution and activation maximization are also referred to as perceptive interpretability, as these methods are more easily perceived by the end users (Tjoa and Guan, 2021). Attribution methods assign an attribution value (i.e. relevance) to each input feature or pixel (Singh et al., 2020), and are highlighted in the literature as the most common approach for DL model explainability in Medicine. More specifically, by determining the contribution of an input pixel to the output of a class, attribution methods generate heat maps, arranging all the attributions in the same shape as the input observation. This way, each pixel will provide negative or positive evidence of a certain intensity. These saliency maps can be easily integrated into a radiology workflow by merging the classification predictions with the input CXR data for easier interpretation.

Linardatos et al. (2021) subcategorizes attribution methods into model-agnostic and model-specific approaches. The first include techniques such as SHAP (SHapley Additive exPlanations,

Lundberg and Lee, 2017), Anchors (Ribeiro et al., 2018), and LIME (Local Interpretable Model-agnostic Explanations, Ribeiro et al., 2016). For example, LIME generates a new dataset consisting of perturbed samples and gets their corresponding black-box model predictions. The perturbed samples are a result of image segmentation into superpixels and are employed to train a new interpretable local model. This surrogate model is trained to approximate and individually explain the black-box model predictions, thus identifying the image superpixels that impacted that outcome the most.

Alternatively, model-specific attribution methods have been more thoroughly studied and are based on the backpropagation routine and gradient computation from the output of the DL model up to a certain layer. Such methods result in reconstructed saliency maps that show the activated image regions for a given class. Saliency (also known as Gradients, Simonyan et al., 2014), Integrated Gradients (Sundararajan et al., 2017), and Guided Backpropagation (Springenberg et al., 2015) are well-known early examples of attribution techniques that compute the gradient of the output with respect to the input layer. Figure 3.3 outlines how these gradient-based methods rely on forward and backward passes to reach the final activation of the class *Cardiomegaly*. In sum, given the input image, a class $c$ is maximally activated through forward passing along all the network layers. Then, the backpropagation step computes the gradients from the output to the input of the network, setting all activations to zero except for the class of interest $c$, thus enabling the visualization of the pixels that contribute to the activation of that specific class.

Zhou et al. (2016) argues that computing the gradient up to the input layer is not effective, and proposes Class Activation Maps (CAMs) to improve such aspect. The authors take advantage of a Global Average Pooling (GAP) layer at the end of the network to average the activations of each feature map. These averages are concatenated into a vector whose weights are considered to predict the final output class. They are then used in the backward step to calculate the gradient from the output to a specific hidden layer, usually the last convolutional one, performing a weighted combination of the corresponding feature maps. Consequently, the resulting CAM also highlights the relevant image regions for that particular class. The process is illustrated in Figure 3.4.

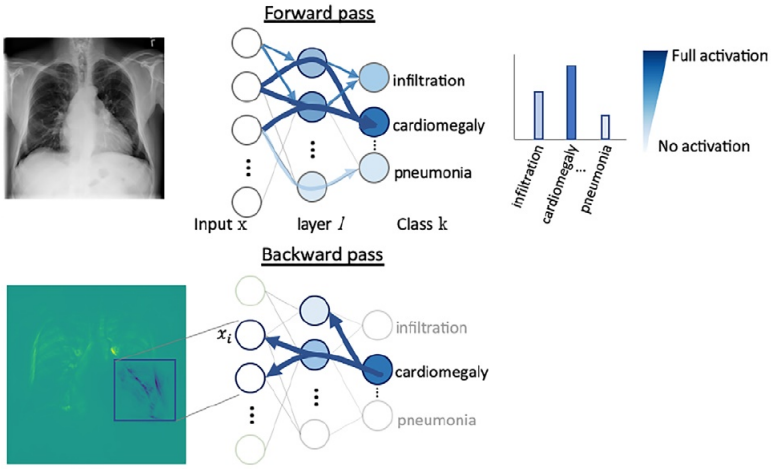



Figure 3.3: Gradient-based saliency maps for image classification. Reyes et al. (2020).

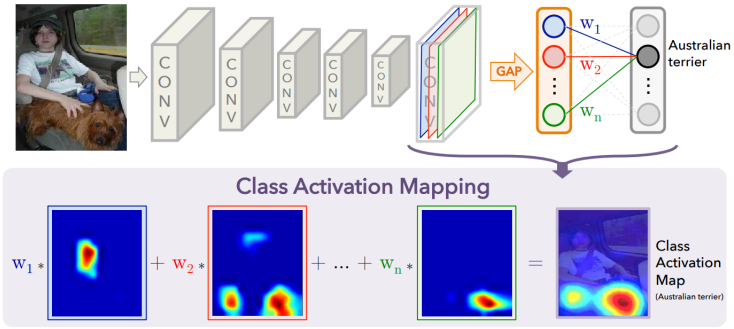



Figure 3.4: The predicted class score is mapped back to the previous convolutional layer to generate the CAMs, thus visualizing the class-specific discriminative regions. The $w_i$ values are the learned weights associated with each feature map after the GAP layer. Zhou et al. (2016).

Gradient-weighted Class Activation Maps (Grad-CAMs, Selvaraju et al., 2017) are a popular generalization of this method that extends the applicability of CAMs to any architecture (i.e. not requiring the GAP layer followed by the single fully-connected layer). They also consider the feature maps produced by the final convolutional layer, but differ from the previous methodology in how these are weighted to create the final class-specific heat map explanation. In CAM, the weights are taken out of the last fully-connected layer of the network (orange block in Figure 3.4), while in Grad-CAM, the feature maps are weighted using coefficients that are calculated based on gradients. A detailed discussion of this method will be provided in the following chapters. Grad-CAM++ builds on this method to better locate and explain occurrences of multiple object instances in a single image (Chattopadhay et al., 2018).

To conclude, note that the most popular XAI methodologies rely on *post-hoc* local visual explanations, which are not required to explain the complete decision process of the black-box model, but simply the reasoning behind a single prediction.

## 3.4 Risks and Evaluation

As seen earlier in this chapter, most DL exercises are affected by the same common challenges, namely the lack of training data, explainability, and reproducibility standards (Kallianos et al., 2019). In a high-stakes context, adding this XAI component is essential to promote the transparency of the DL technology and thus assist its users in rationalizing the predictions. On the other hand, there are certain risks and/or downsides that come with XAI implementation. For instance, adding this feature entails significant computational effort and expertise, since interpretability is domain-specific and is hardly generalized. Besides adding extra complexity to the CAD system, it may pose an accuracy *vs.* explainability trade-off. Multiple publications report how developing an inherently more interpretable model causes a drop in performance in comparison to less interpretable ones. However, the same publications also offer solutions to mitigate such trade-off, e.g. through the use of quality structured data, carrying a meaningful representation of significant features, or novel concept learning strategies (Rudin, 2019; Sheth and Kahou, 2023). Recent work also points out the inherent risk of the gap between law, regulation, and implementation (Lisboa et al., 2023). This gap may be accentuated by inadequate explanations due to incomplete constraints, low-quality training data, or even manipulation, and may lead to rushed assumptions (Tjoa and Guan, 2021).

To avoid these unreliable or misleading explanations, one must consider what defines good XAI. This is a challenging task because, once again, it is highly dependent on the domain and the targeted end user - e.g. the requisites of a satisfying explanation for a clinical specialist certainly differ from the requisites for a patient as an end user. Nevertheless, both groups are likely to reject explanations that are disconnected from the model's computations, lack coherence, or fail to provide sufficient information to enhance understanding of the results.

The scarcity of objective definitions to determine what exactly validates the explanations of a model in clinical practice is acknowledged in Tonekaboni et al. (2019), as very few publications address the quality and usability of the proposed explanation methods for these target users. The authors state that a well-designed explanation should complement the algorithm and its predictions using task-specific representations (e.g. confidence scores) and provide actionable steps that clinicians can undertake (e.g. potential interventions or data collection). While the best criteria for the selection of XAI techniques remain undetermined, two main publications stand out by suggesting more concrete evaluation frameworks. The first was proposed by Doshi-Velez and Kim (2017) and encompasses these three topics:

a) **Application-grounded evaluation**: evaluation of the method by human experts, e.g. by asking a clinician to use the proposed CAD system during clinical practice.

b) **Human-grounded evaluation**: non-experts analyze a less complex version of the experiments, while still maintaining the essence of the target application. Useful for preliminary decisions and/or when experts are not available. An example includes humans being presented with pairs of explanations and choosing the one that they find of higher quality.

c) **Functionality-grounded evaluation**: excludes humans from the evaluation process and uses more objective metrics instead. It is considered the most appropriate and easy to implement by the developers, as they establish a formal mathematical definition of explainability as a proxy measure.

This perspective is aligned with the "PDR" framework, proposed in Murdoch et al. (2019), with the following desiderata for selecting explainability methods:

a) **Predictive accuracy**: the assessment of the model's quality should precede the assessment of the explanations. An inaccurate model will most likely output inaccurate justifications.

b) **Descriptive accuracy**: the degree to which an interpretation method objectively captures the relationships learned by a model. Note that the predictive and descriptive accuracies may conflict in a trade-off, as previously described in this chapter. In complex settings such as image analysis, highly non-linear models can provide high predictive accuracy, but lower descriptive accuracy.

c) **Relevancy**: an interpretation is considered to be relevant if it provides insight for a particular audience of the domain at hand.

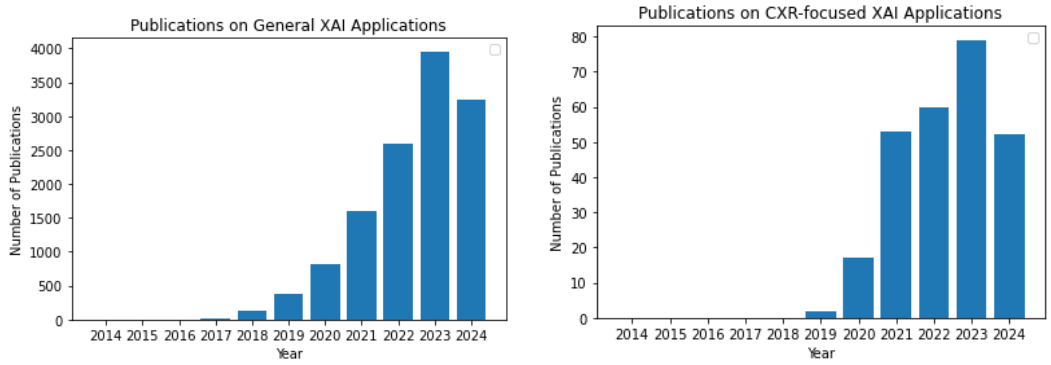
Other criteria often mentioned in the literature regard fairness, privacy, stability, and robustness, among other terms. The following section takes advantage of these ideas in a medical decision context, to explore how clinicians should be actively involved in the identification of certain guidelines and metrics early on, in order to build their trust in the XAI settings.

## 3.5 Explainable Thoracic Pathology Screening

This section focuses on the application of the previously discussed concepts in the clinical context of thoracic pathology screening. Figure 3.5 highlights the global increase in XAI publications over the years, with the medical field reflecting this growing attention as well. In response to this trend, it is crucial to examine what healthcare professionals seek in model explanations, categorizing and discussing the current methods for evaluating them.

### Explanation Categories

Tonekaboni et al. (2019) states that clinicians tend to prefer carefully designed visualizations to facilitate model understanding, so the most commonly used explanations in this domain are model-specific, local, post-hoc visualizations. A more concrete categorization of visual explanations is provided in Herm et al. (2023), which identifies the five main categories illustrated in Figure 3.6(a). These categories consider the location of the region of interest and associated findings, possible image adjustments (e.g. producing a healthy version of an abnormal radiograph), and similar cases for comparison. The study evaluated preferences for these categories using the scale shown in Figure 3.6(b) and found that radiologists particularly value access to radiographs with




(a) (b)

Figure 3.5: Number of publications per year, presenting general and chest radiography-focused XAI applications, respectively. Data collected from *Scopus*, using the search criteria `explainable AND (ai OR intelligence)` and `(explainable OR xai) AND (chest OR thoracic) AND (x-ray OR radiography)`, respectively. Results for 2024 are recorded up to September, 2024.

similar lesions for comparison, along with visualizations of the most relevant image regions linked to the outcome.

A particularly important study for CXR analysis is presented in Xie et al. (2020), where referring physicians and radiologists were surveyed to determine if, when, and what types of explanations are needed. The outcome of this study is a prototype featuring eight key features, as illustrated in Figure 3.7. These are summarized below:

a) **Augmenting CXR images with specific inquiries**: a directed explanation based on a user's target question (e.g. only paying attention to relevant regions of the CXR image).

b) **Presenting explanations hierarchically**: low-level examination (which body part to look at in a CXR image, e.g. lungs, trachea), mid-level observation (what characteristics or are observed, e.g. pleural effusion), and high-level impression (differential diagnosis based on observations, e.g. pleural effusion suggest pericardial disease or congestive heart failure).

c) **Contextualizing observations with contrastive examples**: contrasting a user-selected region with both normal and abnormal versions, as counterfactual explanations allow understanding a case in light of plausible alternatives.

d) **Interpreting observations probabilistically**: an array of CXR images with the same observation, sorted by their probability as unlikely (0%), likely (50%), and very likely (100%) of having that lesion.

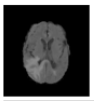
e) **Contextualizing impressions with prevalence and traces**: presenting prevalence and known factors that affect a user-selected impression (e.g. the patient's age, risk factors).

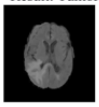
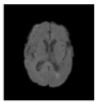
f) **Comparison across time**: accessing the patient's prior images.

g) **Comparison across patients**: comparing with CXRs from different patients having the same finding(s).

h) **Mediating system complexity by urgency level**: showing different amounts of information in urgent vs. non-urgent mode.

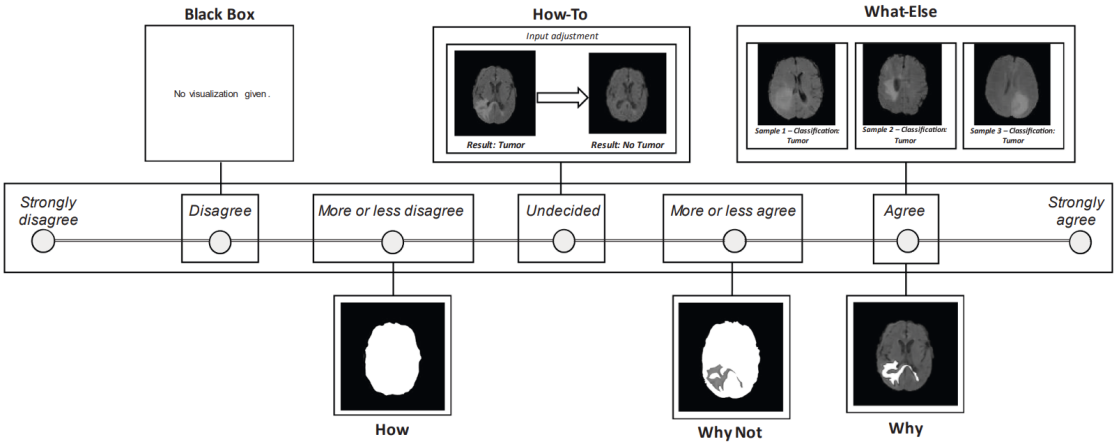
**Input Image**



| Explanation Type | Description of Explanation Types | Treatment Visualization |
|---|---|---|
| *How* | Explanation of which input areas are relevant to the trained model, i.e., the ML algorithm's inner logic.[a] |  Result: Tumor |
| *Why* | Explanation of which areas of the given input are relevant to the outcome of the prediction.[a] |  Result: Tumor |
| *Why-Not* | Explanation of which areas of the given input are not relevant to the outcome of the prediction.[a] |  Result: Tumor |
| *How-To* | Explanation of how hypothetical adjustments of the given input (e.g., the bright shades in the MRI) would result in a (different) model prediction. |  Result: Tumor *Input Adjustment:*  Result: No Tumor |
| *What-Else* | Explanation by example; showing similar inputs and their respective predictions. |  Sample 1 – Result: Tumor  Sample 2 – Result: Tumor  Sample 3 – Result: Tumor |

[a] The described XAI augmentations are highlighted as a white area in the treatment visualization.

(a)




(b)

Figure 3.6: Overview of (a) explanation types, and (b) their average end user rating. Herm et al. (2023).

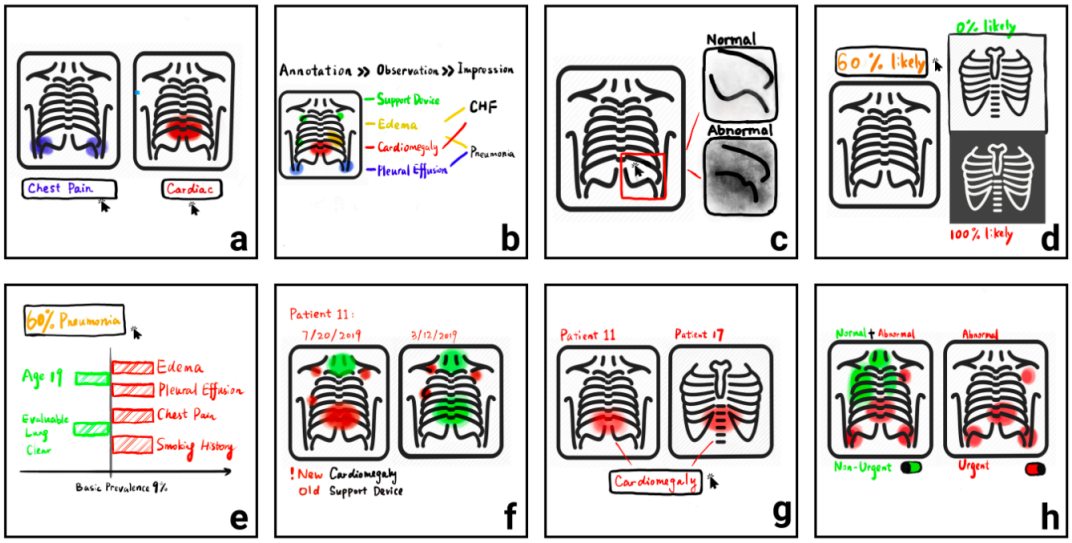



Figure 3.7: Eight key system designs for XAI in thoracic pathology screening. Xie et al. (2020).

In conclusion, the model's focus on clinically relevant features should match the reasoning used by radiologists, complementing each observation with valuable information. This way, models with lower accuracy might be acceptable if they are explicit regarding when and why the model underperforms.

### Evaluation

To the best of the author's knowledge, the current state of the art predominantly relies on functionality-grounded evaluation approaches across multiple domains, none of which are specifically tailored for medical explanations. To illustrate the concepts discussed in this section, a ResNet50 model was trained on the ChestX-Ray14 dataset. This model classifies the presence of 14 common thoracic pathologies and provides corresponding visual explanations for its predictions.

The most common evaluation setting consists in comparing the generated explainability heat maps with the available spatial annotations to assess the level of agreement between the highlighted areas and the bounding box (BB) pixels. The Size-Adaptive Intersection Over Union (SA-IOU) metric, as proposed in Sattarzadeh et al. (2021), works by first binarizing the heat map using its top-$n$ highest intensity pixels, where $n$ is the number of pixels contained within the annotated BB. This process is exemplified on a Grad-CAM in Figure 3.8. The SA-IOU is then calculated as the ratio between the $n$ highest-scoring pixels that overlap with the BB and the total $n$ pixels. This ratio, ranging from 0 to 1, indicates how well the heat map aligns with the BB annotation, with a value closer to 1 indicating a higher degree of overlap. Less popular options include binarizing the map heuristically or with conventional methods such as Otsu's (Saporta et al., 2022).

Alternatively, Energy-Based Pointing Game (EBPG, Wang et al., 2020a) can be used to get a similar measure without binarizing the heat map. As its name suggests, EBPG quantifies the percentage of the map's energy that is contained within the corresponding ground-truth mask.

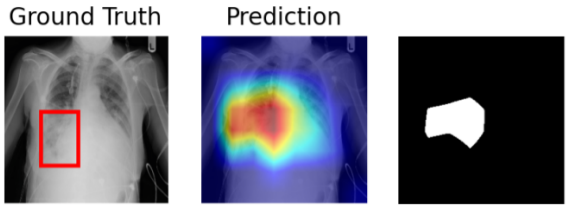



Figure 3.8: Infiltration case from the ChestX-ray14 dataset, with the spatial annotation marked in a red box. The ResNet50 model's Grad-CAM is displayed to its right. The explanation's resulting mask on the far right selects the same number of high-intensity pixels as the ones present in the BB.

This is done through point-wise multiplication, as illustrated on a Grad-CAM in Figure 3.9 and generalized in Equation 3.1 for a heat map $M$ obtained for class $c$.

$$\text{EBPG} = \frac{\sum M^c_{(i,j)\in bbox}}{\sum M^c_{(i,j)\in bbox} + \sum M^c_{(i,j)\notin bbox}} \tag{3.1}$$

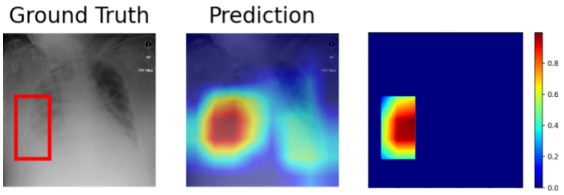



Figure 3.9: Effusion case from the ChestX-ray14 dataset, with the spatial annotation marked in a red box. The ResNet50 model's Grad-CAM explanation is displayed to its right. The EBPG calculations are derived from the point-wise multiplication of the annotation and the heat map, shown on the far right.

However, most public datasets do not provide spatial annotations. In these cases, other metrics must be considered to measure the relevance of the highlighted areas. Chattopadhay et al. (2018) proposes to compute the average probability drop and increase after restricting the input image to the selected activations. Specifically, it is expected that removing portions of an image will decrease the model's confidence in its decision compared to when the full image is provided as input. The average drop quantifies this reduction in confidence for a particular class by measuring how much the model's confidence decreases when only the areas highlighted by the explanation map are used as input. This metric is calculated per image and then averaged across the entire dataset of $N$ images (Equation 3.2). A lower average drop is desirable, indicating that the explanation map effectively identifies the most relevant areas, allowing the model to maintain a certain level of confidence even with reduced input. $Y_i^c$ is the original probability score for an image $i$ of class $c$, and $O_i^c$ is the new output score for the constrained image.

$$\text{Average Drop} = \left( \sum_{i=1}^{N} \frac{\max(0, Y_i^c - O_i^c)}{Y_i^c} \right) / N \tag{3.2}$$

Complementing the previous metric, there are situations where using only the regions highlighted by the explanation map as input can increase the model's confidence in its prediction, e.g. in cases where the surrounding context is distracting. The average increase metric captures this by counting the instances in the dataset where the model's confidence improves when given only the selected regions (Equation 3.3, where $sign = 1$ if the condition is true). The objective is to maximize this metric, indicating that the model is consistently confident in its predictions when focused only on the most relevant areas.

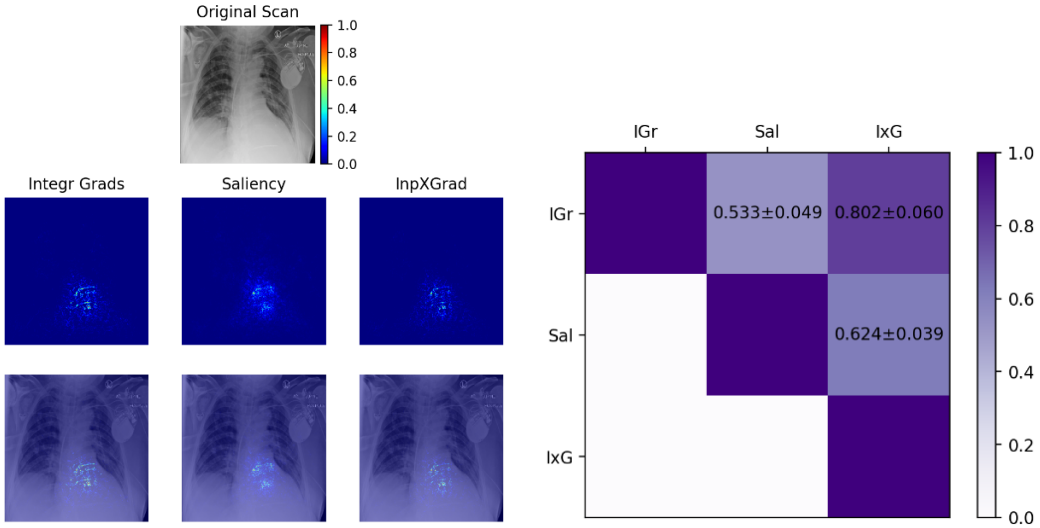
$$\text{Average Increase} = \sum_{i=1}^{N} \frac{\text{sign}(Y_i^c < O_i^c)}{N} \tag{3.3}$$

Purwita and Qomariyah (2021) reports that using different XAI approaches may yield different explanation maps for the same model and instance, given the specificities of each method. While the previously discussed metrics can help identify the most suitable one, it is also of interest that the explanation maps are not vastly different from one another to maintain the overall trustworthiness of the model. In this context, a model that produces more consistent maps may be preferred (Yang, 2022). The consistency of the maps can be evaluated using a similarity metric, such as cosine similarity, defined in Equation 3.4 for two heat maps M with $P$ pixels, obtained with an XAI method A and an XAI method B.

$$\text{Cosine Similarity} = \frac{\sum_{p=1}^{P} M_p^A M_p^B}{\|\mathbf{M}^A\| \|\mathbf{M}^B\|} \tag{3.4}$$

Figure 3.10 exemplifies the results obtained by the ResNet50 model, minding three distinct XAI methods: integrated gradients, saliency, and input $\times$ gradient. Saliency computes the gradient of the output with respect to the input features. Integrated gradients address some of saliency's shortcomings by considering how the model's output changes as the input moves from a baseline (usually a neutral or zero-value input) to the actual input. Input $\times$ gradient multiplies the input values by their corresponding gradients to provide an explanation, building on the idea that the gradient gives the sensitivity of the output to a small change in the input. In detail, Figure 3.10(a) presents these explanation maps for a cardiomegaly instance, while Figure 3.10(b) displays the average cosine similarity matrix calculated for the entire dataset. From this matrix, it can be observed that the saliency method produces the most discordant map compared to the others, in this specific model. In other words, saliency's maps differ more significantly from the integrated gradients' maps, and input $\times$ gradient yields more similar maps to integrated gradients. Note that this analysis employs three high-resolution XAI methods, as comparing maps with differing resolutions would distort the interpretation of the results. For example, Grad-CAM produces lower-resolution maps, with less detail, which contrast with the more granular and detailed explanations provided by the methods under consideration, preventing a fair comparison. Alternatively, Grad-CAMs

could be compared with other lower-resolution methods, or with higher-resolution maps that were adjusted to match the level of blurriness of Grad-CAM using e.g. a Gaussian filter.




(a) (b)

Figure 3.10: Explainability maps obtained by the ResNet50 model for a cardiomegaly instance from the ChestX-ray14 dataset, using high-resolution XAI methods: integrated gradients (*IGr*), saliency (*Sal*), and input $\times$ gradient ($I \times G$). The matrix on the right shows the average cosine similarity across the entire dataset for the same methods.

In sum, the discussed metrics have been presented mostly as a way of choosing the best visual explanation method for a given model. One should opt for the method that maximizes the SA-IOU and EBPG between the activations and the BB annotations, or, if these are not available, the one that maximizes the average increase and minimizes the average drop. Cosine similarity, on the other hand, can be used as an estimate of the model's trustworthiness. Consistent results across different XAI techniques would suggest that the model is confident regarding the locations of the lesion(s). However, these metrics can be as valuable for comparing the explainability of two distinct models, enabling an evaluation of their maps alongside their predictive performance. This helps to validate if their predictions are genuinely accurate and reliable, rather than the result of potential biases. To exemplify such a comparison, the author trained a second model, a DenseNet121, on the same dataset and conditions as the ResNet50 introduced in this section. Their predictive performance was measured with five-fold cross-validation using the Area Under the ROC Curve (AUC), defined in Appendix A and disclosed in Table 3.1. An additional independent test set of 984 images, containing BB annotations of the first eight pathologies, was kept to further evaluate their explainability.

The results show that ResNet and DenseNet deliver comparable predictive performance, with the latter outperforming ResNet by a margin of 1%. This difference is evident in all pathology classes except for atelectasis and fibrosis, in which ResNet outperforms DenseNet. As discussed in Chapter 2, the ResNet and DenseNet architectures were both designed to improve gradient flow in deep networks, using different strategies. ResNet uses residual connections, where the input

Table 3.1: Classification results of the ResNet50 and DenseNet121 models per pathology class (mean and standard deviation in the last row).

| **Class** | **AUC** | |
|---|---|---|
| | **ResNet50** | **DenseNet121** |
| **Atelectasis** | 0.813 | 0.807 |
| **Cardiomegaly** | 0.903 | 0.912 |
| **Effusion** | 0.875 | 0.877 |
| **Infiltration** | 0.710 | 0.721 |
| **Mass** | 0.822 | 0.833 |
| **Nodule** | 0.755 | 0.766 |
| **Pneumonia** | 0.765 | 0.789 |
| **Pneumothorax** | 0.854 | 0.881 |
| **Consolidation** | 0.793 | 0.801 |
| **Edema** | 0.887 | 0.890 |
| **Emphysema** | 0.891 | 0.907 |
| **Fibrosis** | 0.797 | 0.794 |
| **Pleural_Thickening** | 0.774 | 0.791 |
| **Hernia** | 0.872 | 0.880 |
| **Mean** | **0.822** $\pm$ 0.057 | **0.832** $\pm$ 0.062 |

to a layer is added to the output of a later layer, enabling better gradient propagation in deep networks. DenseNet, on the other hand, uses dense connections, where each layer is connected to all subsequent layers, encouraging feature reuse and improving information flow between layers. DenseNet is typically more parameter-efficient, while ResNet is known for being simpler and easier to train in very deep configurations. This aspect could explain the slightly different mean AUC values.

Following this initial analysis, and given the availability of BB annotations, it is only natural to examine the Grad-CAM results of both models to assess their explainability capability. Figure 3.11 and Figure 3.12 display the Grad-CAM maps for identical instances, generated using the ResNet and DenseNet models, respectively. At first glance, DenseNet exhibits more dispersed activations, highlighting larger areas compared to the other model. Figures 3.12(a,b) are clear examples of that. Furthermore, in Figure 3.12(c), DenseNet identifies two possible lesions, whereas the annotation specifies only one. This discrepancy is not observed in the ResNet map in Figure 3.11(c). DenseNet produces unsatisfactory results in Figures 3.12(d,e), showing larger and misplaced activations, while ResNet aligns well with the BBs. In Figure 3.11(f), although ResNet's explanation is also imprecise, it remains closer to the correct anatomical region in comparison to the DenseNet's output in Figure 3.12(f). Furthermore, both networks struggle to delineate the nodule in Figure 3.12(g), due to the low resolution of the training set images. The dataset was rescaled to 256 $\times$ 256 pixels, so very small lesions become more challenging to identify. Finally, Figure 3.12(h) illustrates a case where both models fail to highlight the correct anatomical region.

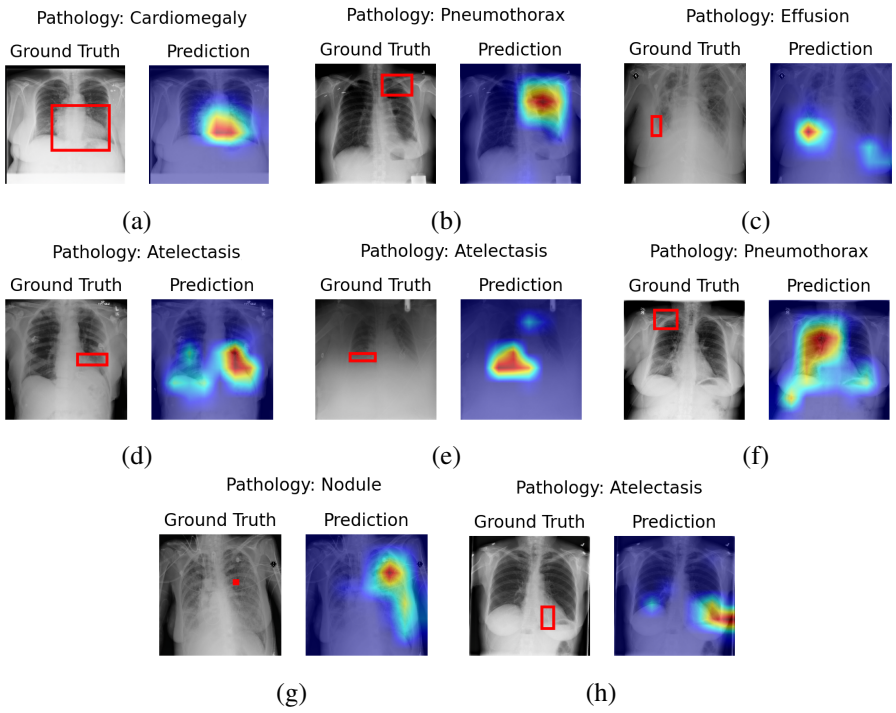



Figure 3.11: Explainability maps obtained for eight instances of the independent test set of ChestX-ray14, using the ResNet50 model.

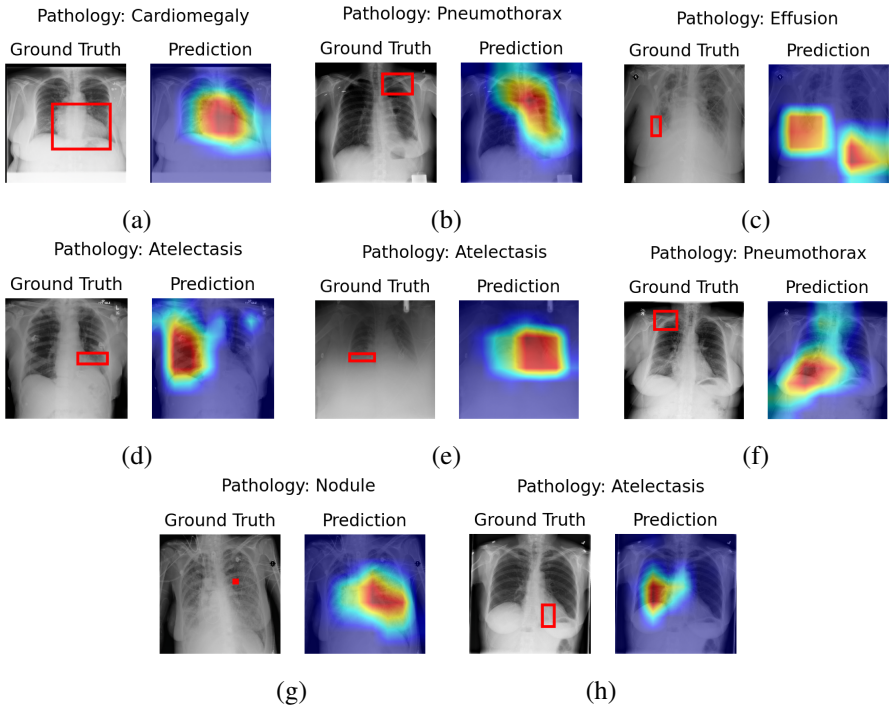



Figure 3.12: Explainability maps obtained for the same eight instances of the independent test set of ChestX-ray14 as Figure 3.11, using the DenseNet121 model.

This analysis is objectively supported by the SA-IOU and EBPG metrics presented in Table

3.2, which demonstrate that the explanations generated by the ResNet50 model exhibit larger overlap with the annotated regions in the independent test set. The ResNet's superiority, in this case, could be attributed to several key factors related to its architectural design and intrinsic feature processing, while DenseNet's dense connections may be leading to more complex interactions that produce broader activation areas and dilute attention on specific areas, ultimately affecting the clarity of its Grad-CAM outputs. This analysis could be further extended to assess the average drop and average increase metrics, if the BBs were not available and/or the results were not as clear.

Table 3.2: Mean overlap between Grad-CAM activations and lesion locations, for all instances of the independent test set predicted as positive.

| | **ResNet50** | **DenseNet121** |
|---|---|---|
| **SA-IOU** | 0.354 | 0.340 |
| **EBPG** | 0.282 | 0.243 |

To exemplify the implementation of the cosine similarity metric, the previous three high-resolution XAI methods were considered. Figure 3.10(b) presents the average cosine similarity matrix for the entire dataset using the ResNet model, while Figure 3.13 displays the equivalent matrix for the DenseNet model. The same key observations hold: the maps produced by integrated gradients and input $\times$ gradients exhibit the highest similarity. Additionally, the total average cosine similarity across all XAI maps for ResNet and DenseNet models is 0.653 and 0.679, respectively. This suggests that the latter generates more consistent outputs. Figure 3.14(a,b) provides two examples of instances where that pattern is confirmed, and Figure 3.14(c) presents a case where the opposite occurs.

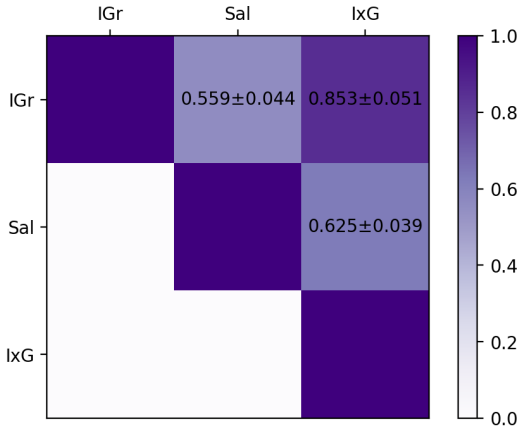
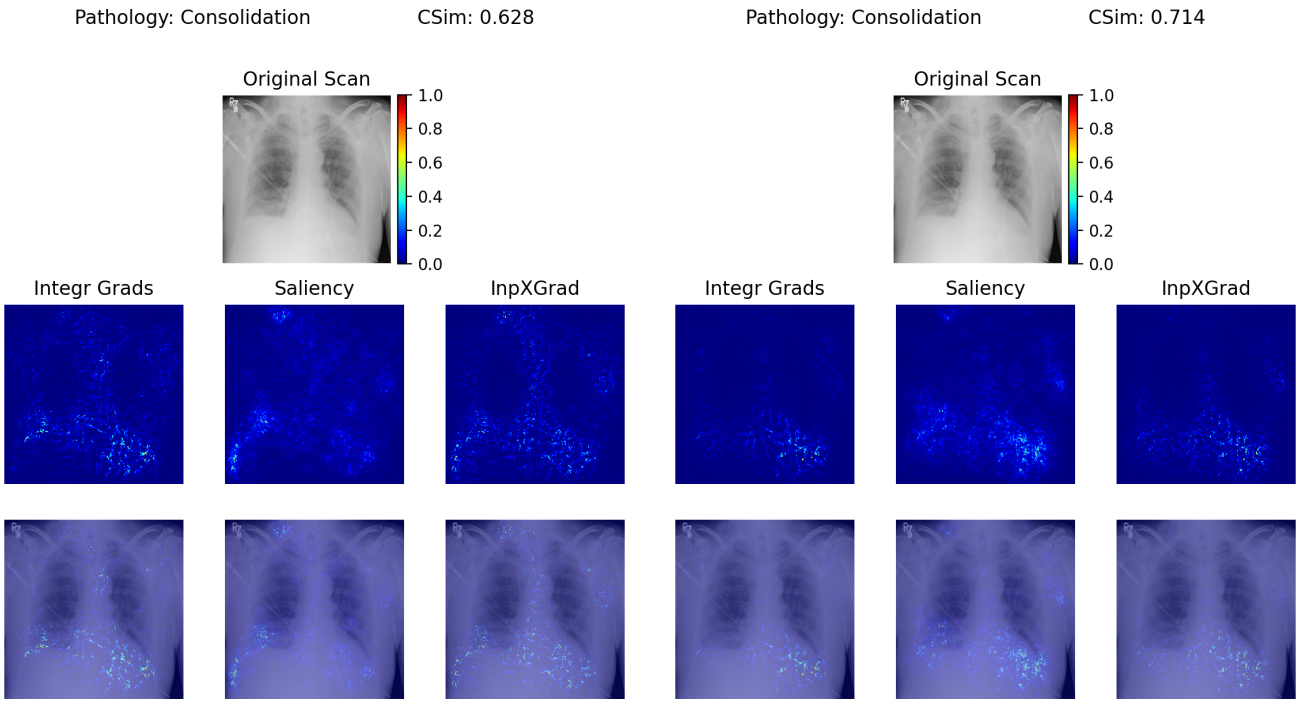



Figure 3.13: Average cosine similarity across the entire dataset for the DenseNet121 model, using high-resolution XAI methods: integrated gradients (*IGr*), saliency (*Sal*), and input $\times$ gradient ($I \times G$).

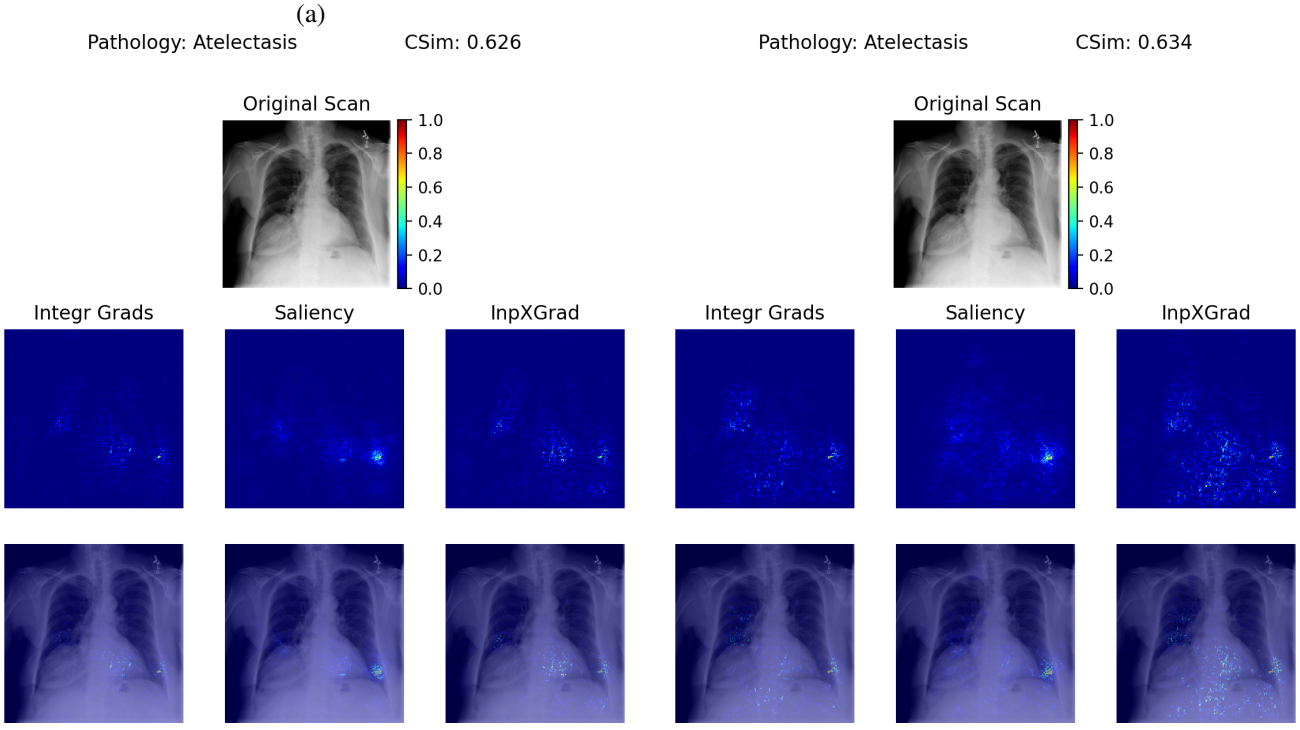
To conclude, it is essential to consider how model selection should be guided by the prior analysis. In terms of predictive performance, both models demonstrate a similar ability to classify

the radiographs, with DenseNet outperforming ResNet by a marginal 1.0%. This tendency is in agreement with the fact that DenseNet produces explainability maps that are 2.6% more consistent, indicating a higher likelihood of agreement on the probable locations of the pathologies. On the other hand, when inferring the overlap between the selected and annotated regions of those pathologies, the ResNet surpasses the DenseNet by 1.4% in SA-IOU and 3.9% in EBPG values.

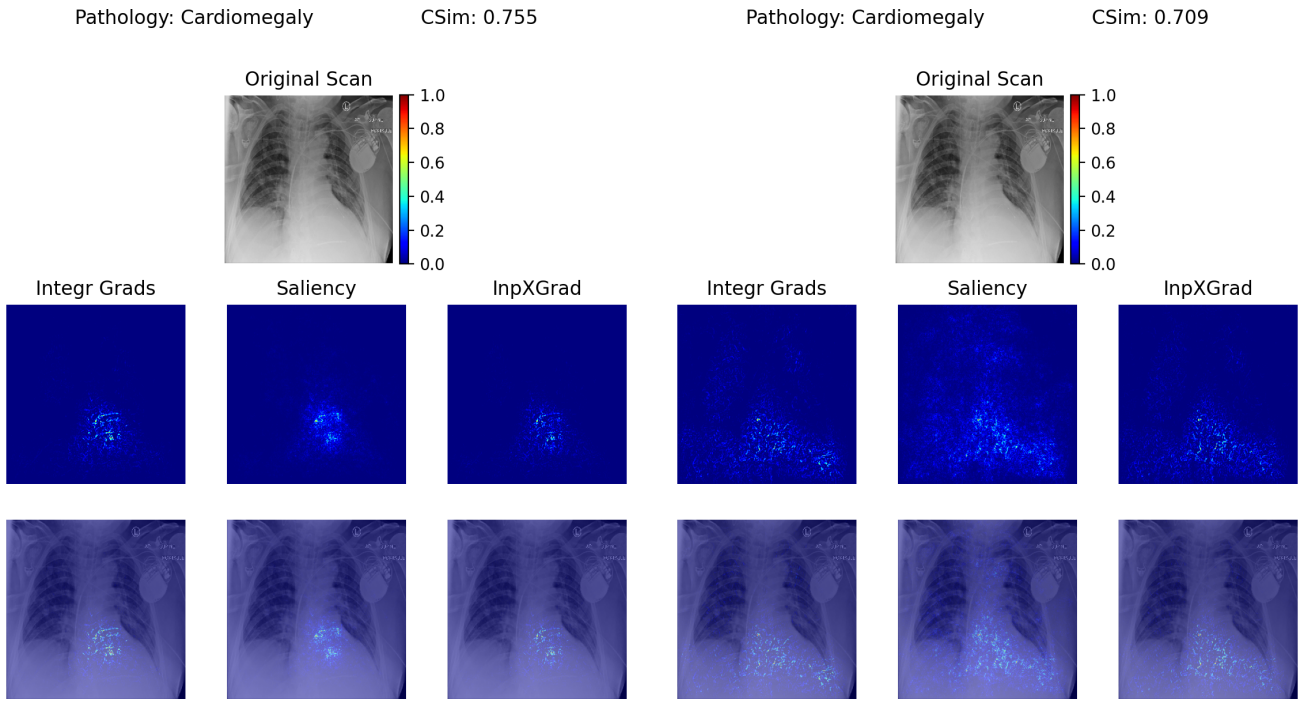
This discussion emphasizes the challenge posed by the absence of standardized guidelines for evaluating medical explanations, which leads to the use of varied and sometimes conflicting metrics to determine the most suitable model. As a result, developers should prioritize the metrics that best align with their specific goal. Given the minimal difference in predictive performance between the models, the author chose to focus on the quality of the explanations, particularly by maximizing the overlap between the heat maps and the BBs, which are the most objective reference. In this context, ResNet was ultimately considered the most reliable option, in a task where both prediction accuracy and interpretability are crucial. Further details on this ResNet model and its training settings are provided in Chapter 5.


Pathology: Consolidation
CSim: 0.628
Original Scan
1.0
0.8
0.6
0.4
0.2
0.0
Integr Grads
Saliency
InpXGrad
Pathology: Consolidation
CSim: 0.714
Original Scan
Integr Grads
Saliency
InpXGrad

(a)


Pathology: Atelectasis
CSim: 0.626
Original Scan
1.0
0.8
0.6
0.4
0.2
0.0
Integr Grads
Saliency
InpXGrad
Pathology: Atelectasis
CSim: 0.634
Original Scan
Integr Grads
Saliency
InpXGrad

(b)




(c)

Figure 3.14: Explainability maps obtained for three instances of the ChestX-ray14 dataset, using high-resolution XAI methods (integrated gradients, saliency, and input × gradient, respectively) on the ResNet50 model (left) and DenseNet model (right). The cosine similarity (*CSim*) between the three maps of that particular instance and model is displayed on top of the original radiograph.

## 3.6 Concluding Remarks

The availability of large medical datasets has improved model performance, but DL's limitations hinder its clinical adoption. Key challenges include difficulty in generalizing across datasets, overfitting, and lack of transparency in decision-making, which is crucial in high-stakes medical contexts. DL's black-box nature raises concerns about trust and accountability, particularly in radiology. XAI seeks to address these issues by promoting model transparency, interpretability, and compliance with regulations like the GDPR. XAI is essential for ensuring model safety, detecting biases, and improving system reliability. This chapter covers the most important taxonomic concepts and state-of-the-art methods, and concludes that the medical field typically favors model-agnostic, local, post-hoc visualizations.

Effective XAI should address domain-specific needs and target users, with clinicians requiring actionable, reliable explanations. The current evaluation frameworks emphasize assessing XAI based on predictive accuracy, descriptive accuracy, and relevancy, along with expert involvement. In clinical settings, early engagement of clinicians in developing XAI guidelines is crucial for building trust and ensuring the usability of these models. This discussion is extended into thoracic pathology screening, linking broader XAI concepts to this specific clinical application, and thus exploring what constitutes a good explanation and how it can be objectively evaluated to benefit

clinical decision-making. Current prototypes address these issues by contextualizing observations, incorporating probabilistic interpretation, and enabling comparisons both over time and across different patients. Finally, it is concluded that several metrics can be employed to select the best model and/or XAI technique. Developers must select the most adequate ones e.g. based on the specific requirements of the task, the availability of spatial annotations for estimating explanation quality and accuracy, and the model's predictive performance.

# Chapter 4

# STERN: Attention-driven Spatial Transformer Network for Abnormality Detection in Chest X-ray Images

**The content of this chapter results from the following publication. The author of this thesis has the main contribution to the conception of the study, method development, and results assessment.**

**Rocha, J.**, Pereira, S. C., Pedrosa, J., Campilho, A., and Mendonça, A. M. (2024). STERN: Attention-driven Spatial Transformer Network for abnormality detection in chest X-ray images. Artificial Intelligence in Medicine, 147, 102737.

Further implementation details of STERN can be found in the **GitHub repository** provided by the author. https://github.com/JoanaNRocha/STERN-attention-driven-spatial-transformer-network

While DL models can be employed to automate the interpretation of CXR images to prioritize urgent cases, the presence of image artifacts (e.g. lettering) often generates a harmful bias in the classifiers and an increase in false positive results. To address this, healthcare could benefit from a system capable of accurately selecting the thoracic ROI before determining whether an image shows potential signs of pathology. Although attention mechanisms are not explicitly categorized as an XAI technique, they are valuable in this context as they help the model focus on the correct anatomical regions, thereby enhancing both the model's classification performance and explainability outcomes.

The work herein described tackles a binary classification exercise, in which an image is either normal or abnormal, using an attention-driven and spatially unsupervised **S**patial **T**ransform**er** **N**etwork (STERN), that takes advantage of a novel domain-specific loss to better frame the ROI. Unlike the state of the art, in which this type of network is usually employed for image alignment, this work proposes a spatial transformer module that is used specifically for attention, as an alternative to the standard object detection models that typically precede the classifier to crop

out the ROI. In sum, the proposed end-to-end architecture dynamically scales and aligns the input images to maximize the classifier's performance, by selecting the thorax with translation and non-isotropic scaling transformations, and thus eliminating artifacts. Additionally, this paper provides an extensive and objective analysis of the selected ROIs, by proposing a set of evaluation metrics.

The results indicate that the STERN achieves similar results to using YOLO-cropped images, with reduced computational cost and without the need for localization labels. More specifically, the system is able to distinguish abnormal frontal images from the CheXpert dataset, with a mean AUC of 0.857 - a 2.6% improvement vs. the 1.0% improvement achieved by the YOLO-based counterpart in comparison to a standard baseline classifier. At the same time, the STERN approach requires less than 2/3 of the training parameters.

## 4.1 Introduction

A CXR-based abnormality detection system may help clinicians prioritize more urgent abnormal exams, outputting the image-level binary labels normal or abnormal. DL can be employed for such a purpose, given its inherent ability to select and combine image features in an efficient learning process. However, the resulting models may also be subject to data-related biases (Gianfrancesco et al., 2018). For instance, the letters in a CXR image may be wrongfully interpreted as pathologic features (Figure 4.1(a) in red), according to DeGrave et al. (2021). Such characteristic is a major setback for their broad adoption in a clinical setting, whose domain is grounded in the ability to exploit causal relationships. This way, researchers often resort to a preprocessing stage in which they crop the images' ROI before feeding them to a classifier, selecting only the portion of the image that contains the thorax. Object detection models are often used for this purpose, whose outcome is exemplified in Figure 4.1(b). This approach contributes to the reduction of false positive predictions by removing image artifacts outside the selected ROI, but at the same time, it also implies training a completely separate model with BB Ground Truths (GTs) before training the main classifier, creating significant extra human and computational costs.

Spatial Transformer Networks (STNs) were originally introduced in Jaderberg et al. (2015) and refer to any CNN that includes a spatial transformer module. This module can be trained end-to-end with a classifier to transform the original input images in a way that benefits the classification task. In sum, the present chapter focuses on replacing the typical object detection model (e.g. a YOLO network, Çallı et al., 2021; Karacı, 2022) for a more efficient workflow, using an attention-driven STN to simultaneously select the ROI and classify the images, thus suppressing the need for two independent models. More specifically, the proposed STERN approach implements a model that is fine-tuned on a domain-specific loss function, enabling the module to more precisely locate and select the thoracic region. The main contributions of this study are summarized below:

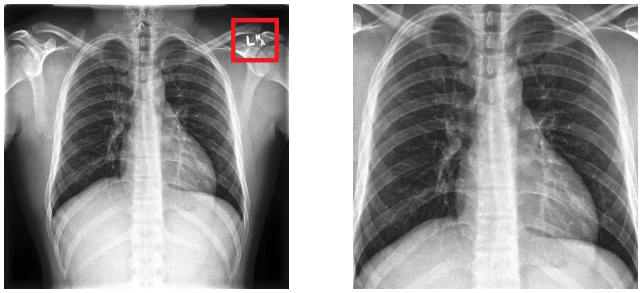
1. A single STN whose attention-driven module and classifier are trained end-to-end to crop and classify the CXR images.

2. A novel domain-specific loss function to further refine the selection of the thoracic ROI in a fine-tuning stage.

3. A reduced computational cost in comparison to the typical object detection models, which require localization GT annotations and larger architectures.

4. A set of proposed metrics to further evaluate the selected ROI, given the lack of spatial annotations in the dataset.

Section 4.2 presents a review of the state-of-the-art publications on the topic. A detailed overview of the dataset and framework is presented in Section 4.3. The results are further discussed in Section 4.4, and the main conclusions presented in Section 4.5.



(a) Original image with artifacts marked in red.

(b) ROI selected by a thorax detection model.

Figure 4.1: Example of an image before and after cropping, in which the outer frame of the image is discarded, consequentially removing the artifacts outside that region.

## 4.2 Related Work

Binary classification models that focus on general abnormality detection have received far less attention than multi-label pathology classification models, but are still considered a relevant line of research in the medical imaging field (Alloqmani et al., 2021; Tschuchnig and Gadermayr, 2022). In fact, including such detection task in a computer-aided diagnosis system can have a high impact when it comes to building a triage system for the analysis of CXR images. In general, standard off-the-shelf CNN-based methods are frequently applied to the data in question, establishing a comparison between well-known architectures. For instance, the work in Yates et al. (2018) studies the performance of an Inception model on frontal CXR images, while the work in Tang et al. (2020) does the same using AlexNet, VGG, ResNet, Inception and DenseNet. It is also possible to combine the outcome of multiple models (AlexNet, ResNet) in an ensemble as seen in Karthik and Sowmya Kamath (2022), to further improve the prediction accuracy.

Alternatively, other state-of-the-art publications opt for autoencoder-based one-class learning approaches to infer normality patterns and distinguish the abnormal radiographs. In such scenario, and given only normal instances in the training stage, an autoencoder can be employed to reconstruct the input image as best as possible. This is the main idea behind Nakao et al. (2021): the

reconstruction error will be higher if the test instance has an abnormality, as the model has not seen any lesions in the training stage to be able to replicate them. On the other hand, according to Mao et al. (2020), an autoencoder can be used to not only output the reconstructed normal version of that image using a mean squared error loss, but also a pixel-wise uncertainty prediction, which correlates with the likelihood of the radiograph being abnormal. To soften the exclusivity of normal images during optimization and thus obtain a more realistic scenario, Shvetsova et al. (2021) allows a reduced subset of abnormal examples in the training set. In this case, the authors resort to a perceptual loss function.

Several studies have considered the importance of detecting anatomical regions within the thorax, such as the lungs (Furutani et al., 2019; Zhao et al., 2022) and heart (Junior et al., 2021). This task has been incorporated into more elaborated algorithms for ROI detection, in order to look solely at the areas in which lesions may appear. That can be achieved e.g. through a segmentation exercise that delimits and crops the thorax accordingly before classification (Baltruschat et al., 2019; Liu et al., 2019b; Moradi et al., 2020; Teixeira et al., 2021), or through an object detection network with the same goal (Karacı, 2022). These are considered common state-of-the-art approaches for "ROI attention". The work in Teixeira et al. (2021) is particularly relevant, as it discusses the impact of ROI segmentation preceding CXR classification and highlights the biases that arise in such workflow, introduced by the use of different database sources for training and/or inference. The authors concluded that ROI selection mitigates potential bias by eliminating confounding factors outside the lung area, thus providing a more reliable and realistic model for the task at hand.

STNs have been used in different domains mostly for object localization (Meethal et al., 2021; Li et al., 2022; Baruah et al., 2023), and scene recognition (Guo et al., 2016; Liu et al., 2019a), benefiting from the alignment or attention provided by their module, but resorting to overall complex architectures with multiple branches to detect more than a single object. Image registration and super-resolution are also domains in which STNs are beneficial to align the input and get a uniform view across the dataset (Rempakos et al., 2023; Deng et al., 2023). Similarly to other state-of-the-art end-to-end architectures (Xiong et al., 2023), these networks provide higher computational efficiency with reduced inference times. In the medical field, the implementation of STNs ranges from image registration in brain magnetic resonance scans (Lee et al., 2019) and aortic valve segmentation in computed tomography scans (Pak et al., 2020) in 3D data, to skeleton-based gait recognition on 2D images (Zhang et al., 2023a).

Three relevant papers have recently implemented STNs in CXR images to achieve an invariant canonical pose through affine transformations (rotation, shear, scaling, and translation). The first one, Liu et al. (2019c), focuses on multi-label pathology classification using spatial annotations, and takes advantage of aligned image pairs (one normal and one abnormal image) to further improve this task. The two-branch network extracts convolutional features from the input pairs and then computes the contrast-induced attention on the extracted feature map to estimate where the lesion is located. Binary cross entropy loss and multi-instance learning loss take charge of box-level annotated data and class-level data, respectively. A third loss function, combining the

perceptual and consistency losses, is considered in the ST module, to guide the alignment by minimizing the difference between the transformed radiographs and a canonical CXR reference. The same loss function is considered in Kim et al. (2023) for image alignment given a fixed reference. In this case, the authors seek to synthesize the aligned image pairs, with a normal and abnormal version of the same radiograph obtained by a generative adversarial network, before feeding them to an object detection model for lesion localization. Both the work in Liu et al. (2019c) and in Kim et al. (2023) use a ResNet18 as a backbone for their STNs.

Alternatively, a third paper addresses binary abnormality classification using multi-modal data and image-level labels (Bharati et al., 2020), combining an STN with a standard VGG16 network to align all instances. In this example, the patients' clinical data is considered along with the CXR images to train the model, but no details are provided regarding the implemented loss function. All three publications employ the ChestX-ray14 dataset, but the work in Kim et al. (2023) also considers the VinDr-CXR dataset.

Multiple research questions arise from these three articles. For instance, one may ask if it is possible to address STNs specifically for region attention rather than just image alignment, and if it is simultaneously possible to simplify the problem by restricting the number of transformation parameters to predict. Furthermore, would it be possible to do so without any fixed image reference or spatial annotations, and using exclusively image data? If so, it would also be needed to evaluate the relevance of the selected ROI without resorting to GT BBs. A recent conference publication (Rocha et al., 2022) addresses these issues by proposing an STN specifically for attention purposes, in addition to image alignment, to crop the thorax/exclude the outer regions before inferring if an image is normal or abnormal (binary setting). This was implemented using image-level labels only, by restricting the transformations that are allowed to take place to only scaling and translation, thus simplifying the number of parameters to predict. In comparison to that publication, this chapter's work offers improvements in the following areas:

1. Training the STN-based classifier with more data - the original conference publication considers only posterior-anterior frontal radiographs, while this publication considers both posterior-anterior and anterior-posterior frontal radiographs.

2. Adding a fine-tuning stage with a domain-specific loss to further refine the selection of the thorax.

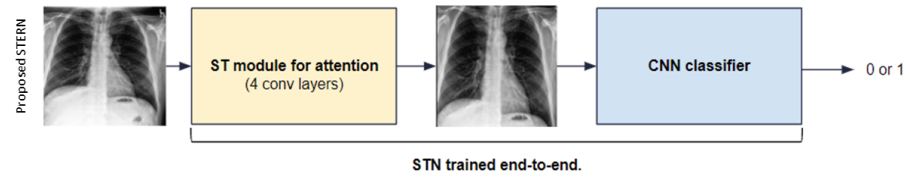
3. Result comparison with a new thorax detection baseline that considers non-isotropic scaling for object detection.

4. Proposal of a set of evaluation metrics that objectively assess the selected ROI without the need for spatial GT annotations.

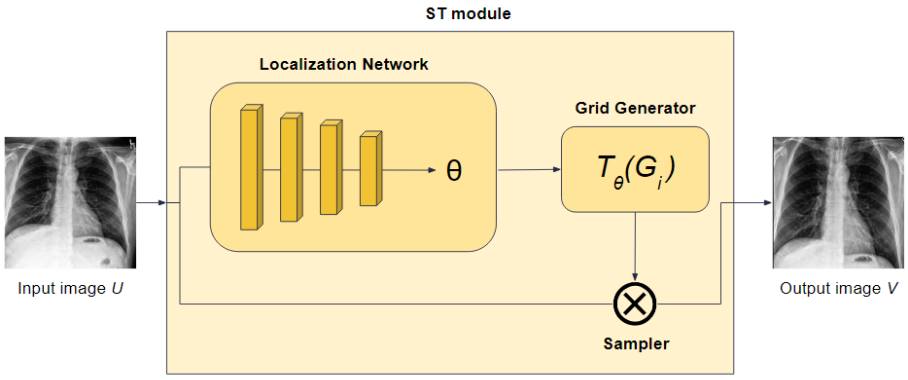
## 4.3 Method

### 4.3.1 STERN

The STERN architecture consists of the ST module, followed by a CNN classifier, outlined in Figure 4.2(a). The present section focuses primarily on the description of the module and the proposed loss function, due to their relevance for the ROI selection task. The ST module consists of three main components: a localization network, a grid generator, and a sampler, represented in Figure 4.2(b).




(a)




(b)

Figure 4.2: Schematic of the STERN components. (a) ST module and classifier within the STN. Class 0 is considered a normal instance, while class 1 is an abnormal instance. (b) Components of the ST module. The localization network predicts the transformation parameters $\theta$, the grid generator uses an inverse transform $T_\theta$ to calculate the corresponding position in the input image, and the sampler samples the pixel value from that input image using bilinear interpolation to produce the transformed output. Rocha et al. (2022).

Firstly, an input image is fed to the localization network that predicts the *theta* parameters ($\theta$). These parameters encode the affine transformations to be applied to the image, and originally encompassed operations of scaling, translation, rotation, and shearing. However, the last two operations are not needed for the purpose of attention, and so it is proposed to reduce the complexity of the transformations by predicting only four $\theta$ parameters instead of six, corresponding to vertical and horizontal scaling ($s_x$,$s_y$) and translation ($t_x$,$t_y$) in Equation 4.1. Here, a shallow CNN was considered for the architecture of the localization network, composed of four convolutional blocks, a dense layer with Rectified Linear Unit (ReLU) activation, and a final dense layer to predict the

four $\theta$ variables. Each convolutional block contains a convolutional layer with ReLU activation followed by max pooling.

In this particular type of images, a larger zoom in the vertical axis is needed rather than in the horizontal axis to accurately delimit the thoracic region. For this reason, non-isotropic scaling was considered to better frame the ROI, particularly because the original radiographs already have shorter margins on the left and right sides of the thorax, in comparison to the top and bottom margins. Consequently, the original image proportion may or may not be distorted. This way, the parameters were initialized as follows: $s_x = 1$, $s_y = 0.1$, $t_x = 0$, and $t_y = 0$. Note that the original publication proposes to initialize all parameters with an identity matrix (Jaderberg et al., 2015), meaning that the model will implement no transformations in the first epoch, and then progress towards optimal $\theta$ values during training. However, in this case, the $s_y$ component can be initialized with a low positive value (0.1), meaning that early transformations start with a larger vertical zoom and progress towards the optimal $s_y$ prediction in the following epochs, enabling a faster convergence since a low $s_y$ (i.e. large vertical zoom) is expected.

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix} \tag{4.1}$$

The last portion of the ST module uses a grid generator and a sampler to output the new transformed image, focused on the ROI. This is achieved through reverse mapping to determine and sample the corresponding position in the input image and use that value to fill in each output image pixel $(x_i^t, y_i^t)$. More specifically, this process is initiated with the grid generator iterating over the output image grid $(G_i)$ and applying the inverse transformation $T_\theta$ to calculate the corresponding sample positions $(x_i^s, y_i^s)$ in the input image, based on the predicted $\theta$ transformation matrix. The result is a sampling grid $T_\theta(G_i)$ containing this set of coordinates where the input map will be sampled to produce the transformed image, as formulated in Equation 4.2. The sampler will then iterate over the entries of the generated sampling grid $T_\theta(G_i)$ and extract the corresponding pixel values from the input map using bilinear interpolation. In short, the module yields a transformed image with the same size as the original input, which is passed along to the CNN for classification, in a way that both the module and the classifier are optimized together and trained end-to-end. Details on the classifier are provided in Section 4.3.3.

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G_i) = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix} \times \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{4.2}$$

The training routine of the STERN approach comprises two stages. First, the described STN model is optimized based on the previously appointed constraints and the Binary Cross Entropy (BCE) loss in Equation 4.3, where $y_i$ is the GT label (0 or 1) and $\hat{y}_i$ is the predicted probability (between 0 and 1) for each of the $N$ images. That initial model is then fine-tuned using the STERN loss described in Equation 4.4, which encompasses a scaling factor component ($S_f$). This function seeks to prevent the model from cropping out pixels belonging to the ROI due to an exaggerated

vertical zoom, by penalizing the model when a prediction has a $s_x/s_y$ ratio greater than the scaling factor - more precisely when the vertical zoom is at least $S_f$ times larger than the horizontal one.

$$\text{BCE Loss} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i)) \tag{4.3}$$

$$\text{STERN Loss} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i) + ReLU(s_{xi}/s_{yi} - S_f)) \tag{4.4}$$

The scaling factor is a fixed scalar parameter, defined by the mean of the $s_x/s_y$ ratio of the training set data in the initial model. As such, each image whose ratio is higher than the mean ratio gets penalized during the second training stage (thus decreasing their vertical zoom), while the remaining images keep their initial ratio.

### 4.3.2 Dataset Preprocessing

This work considered the CheXpert dataset. The employed subset comprises the 191,027 frontal CXR images from the original dataset, including both PA and AP views (the corresponding distribution is shown in Table 4.1). Note that a very small subset of images is incorrectly annotated as having a "left lateral" or "right lateral" view, but those instances were still kept in this study's subset, since they were simultaneously labeled as frontal.

Besides labeling 12 thoracic findings and the presence of support devices, the original dataset also considers a *No Finding* label to capture the absence of pathologies. The *No Finding* observation is assigned a positive label if there is no pathology classified as positive or uncertain (here, the pathologies analyzed extend beyond the pathologies used for labels, according to Garbin et al., 2021). The frequency of each pathologic finding considered in this work is in Table 4.2, yielding 16,974 normal images (class "0") and 174,053 abnormal images (class "1"). All radiographs were resized to 256×256 pixels to standardize their dimensions and accelerate the training routine.

The conversion of the original annotations into binary ones was done following CheXpert's *U-Ones* approach in Irvin et al. (2019), in which uncertain labels are considered positive. This way, an image is considered normal only when none of the discriminated pathologies are present, and the *No Finding* label is affirmative. Note that the presence of support devices, originally labeled in the dataset, was disregarded as it is not indicative of any abnormality. This selected subset of data was split into five folds, preserving the original pathology proportion and without patient overlap (i.e. two distinct folds cannot encompass data belonging to the same patient). More specifically, a five-fold cross-validation scheme was implemented in all experiments, in which three folds were used for training, one for validation, and one for testing. Further details on the matter are given in Section 4.3.3.

Table 4.1: Radiographic projection and view distribution of the CheXpert frontal subset. The 17 images falsely annotated as left or right lateral were also annotated as frontal radiographs, and for that reason were kept in this study's set.

| **View** | **Number of Instances** |
|---|---|
| Anterior-posterior (AP) | 161,590 |
| Posterior-anterior (PA) | 29,420 |
| Left lateral (LL) | 16 |
| Right lateral (RL) | 1 |

Table 4.2: Label distribution of the CheXpert frontal subset.

| **Findings** | **Number of Instances** | **Binary Setting** |
|---|---|---|
| Atelectasis | 59,583 | 174,053 abnormal images |
| Cardiomegaly | 30,092 | |
| Consolidation | 37,364 | |
| Edema | 61,493 | |
| Enlarged Cardiomediastinum | 19,473 | |
| Fracture | 7,935 | |
| Lung Lesion | 8,148 | |
| Lung Opacity | 98,642 | |
| Pleural Effusion | 86,477 | |
| Pleural Other | 4,317 | |
| Pneumonia | 20,656 | |
| Pneumothorax | 20,401 | |
| No Finding | 16,974 | 16,974 normal images |

### 4.3.3 Experimental Setup

Three experimentation setups were established. In addition to the proposed STERN approach, two baseline experiments were conducted as control setups for comparison. These baseline models provide a reference to assess the effectiveness of the STERN method against more conventional approaches, whose general workflow is shown in Figure 4.3.

In all cases, the training data was normalized initially to a 0-1 range and then using ImageNet's mean and standard deviation (STD) values. Datasets are standardized with ImageNet's mean and standard deviation when using pre-trained models, because these models were initially trained on ImageNet images with that specific normalization. Aligning the input data distribution helps the model generalize better to new data by ensuring that features are on a similar scale. Then, the same data was augmented to double the samples through soft random affine transformations (i.e. 5 degrees of rotation range and 3 degrees of shear range). A weighted random sampler was implemented when loading the training set to tackle the considerable class imbalance by ensuring that the batches contain a similar number of instances belonging to each class.

The CNN classifier is common to all three experiments and consists of the widely-used VGG16

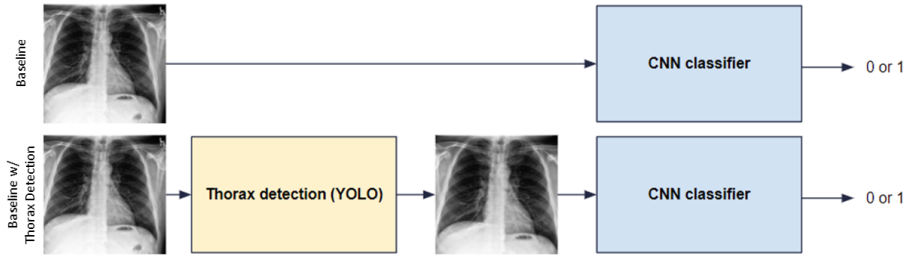



Figure 4.3: Experimental setup baselines. Class 0 is considered a normal instance, while class 1 is an abnormal instance.

architecture, whose last fully-connected layer was modified to consider a single node for normal/abnormal instance classification. This layer is followed by a sigmoid activation function to yield the corresponding predictions. It was trained using the BCE loss and the stochastic gradient descent optimizer, with an initial learning rate of $10^{-3}$ and a total of 15 epochs. Multiple architectures and hyperparameters were initially tested, but ultimately the described conditions provided the strongest classification baseline for this specific learning task. The VGG16 was chosen due to its simplicity and effectiveness in diverse image classification tasks (Simonyan and Zisserman, 2015). The hyperparameter search process was carried out using a grid search, i.e. looping through different values and evaluating the model's performance for each combination. This optimization routine was replicated in the other experiments.

When used by itself, the VGG16 constitutes the first baseline, i.e. a straightforward classification network with no preceding task. The second baseline, further addressed as the baseline with thorax detection, focuses on refining the initial baseline workflow by restricting the image regions being analyzed by the model, i.e. by keeping only the thorax portion of the radiograph and eliminating any structures outside this ROI. Note that the aim here is to prevent the influence of information represented outside the thorax region, such as left/right lettering markers or irrelevant anatomical features (e.g. the patient's neck and shoulders). Artifacts within the ROI will not be removed.

The thorax detection model was based on a YOLOv5 (Redmon et al., 2016) and trained with 956 CXRs from publicly available lung segmentation databases. More specifically, the GT was obtained by drawing a BB around the manual lung field segmentation masks provided in the JSRT, MC, and Shenzhen datasets. The YOLOv5 network was initialized using weights pre-trained on the COCO dataset (Lin et al., 2014), and training was performed using a stochastic gradient descent optimizer with an initial learning rate of $10^{-2}$ for a duration of 150 epochs. This combination of hyperparameters was also proven to be ideal following multiple tests. Once trained, the detection model was applied to the CheXpert data to yield BBs of the thorax region. For post-processing, low confidence predictions (i.e. probability inferior to 90%) were discarded to keep only the most relevant cropped images. The BB was used to crop the images without enforcing isotropy, which were then resized back to $256\times256$ pixels and fed to the binary classifier.

As previously discussed, the STERN approach trained the same classifier architecture, but

in an end-to-end manner alongside the ST module. STERN includes an additional fine-tuning phase, performed over 5 epochs and using the final learning rate and optimizer state resulting from the initial training phase. Several experiments were conducted to determine the optimal hyperparameters for this stage, and only the best-performing values are reported.

### 4.3.4 ROI Selection Evaluation Metrics

Existing evaluation metrics in the field, such as Intersection Over Union (IOU) and F1-score, typically focus on comparing the selected regions to a spatial GT. By measuring the overlap between the region selected by each method and a GT BB, it is straightforward to conclude on the most efficient approach. However, in this case, there are no spatial annotations available, which led to the necessity of defining new metrics to evaluate the performance of the proposed method and distinguish it from the YOLO baseline. As such, the author has delineated specific topics to further comprehend and evaluate the applied transformations:

1. Do the STERN and YOLO-based approaches yield mostly concordant results?
2. Do the STERN and YOLO-based approaches privilege vertical and/or horizontal scaling?
3. How are the intensity profiles of the images before and after the STERN transformation?
4. Are the images more aligned after transformation, i.e. presenting a canonical view of the thoracic ROI?

One must assess the similarity between the region selected by the YOLO and the STERN approaches, as well as whether they focus more on vertical or horizontal scaling (if not both). Additionally, one may extend the evaluation of the proposed method by analyzing how the STERN impacts the intensity profile of the images after transformation, and if the said transformations end up aligning the thorax so that it is represented in similar coordinates across the multiple radiographs. These aspects are covered in the present section, in which several metrics were established as a means of comparing the YOLO-cropped and the transformed images, and further evaluating the ability of the STERN to select the ROI.

Foremost, one may use the original dataset radiographs as a reference and compare the YOLO BB with the region selected by the STERN algorithm before the cropping/transformation takes place. By inferring the selected regions in both situations and plotting them on the original radiograph, it is possible to measure their overlap with the IOU, as well as the mean of both vertical (V1, V2) and horizontal (H1, H2) margins, represented in Figure 4.4.

Alternatively, one may take the original radiographs and the transformed radiographs to compare their corresponding vertical projections, as a result of averaging the pixel intensity of all individual columns. These can be used to trace a profile and assess how the STERN transformation affects image intensity when it comes to removing artifacts and/or noisy areas outside the ROI, as will be further explored in Section 4.4.1. An example of a single radiograph is provided in Figure 4.5.

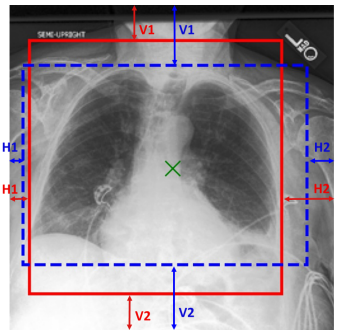
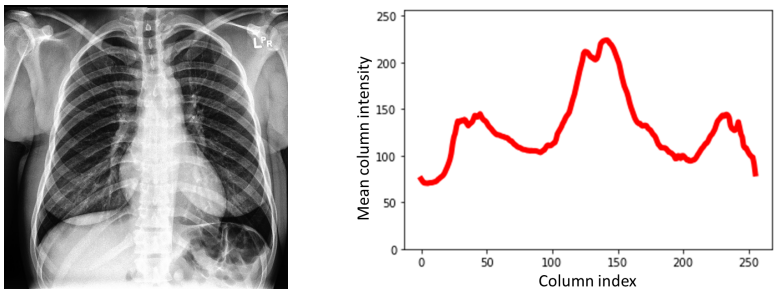
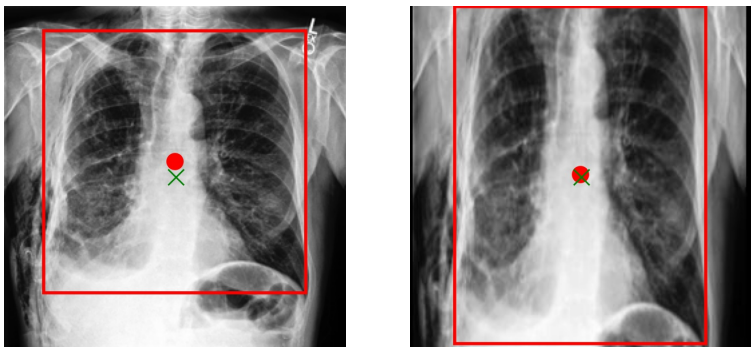



Figure 4.4: Example of ROI selected by the YOLO (red line) and the STERN (blue dashed line) models in an original radiograph before the cropping and transformation, respectively. The four horizontal (H) and vertical (V) margins of each approach are marked using the same color scheme, and the green cross marks the center of the image.




(a) Original radiograph. (b) Corresponding vertical projections.

Figure 4.5: Example of vertical projections for an original CXR, where the horizontal axis corresponds to the column index and the vertical axis corresponds to the average intensity per image column.



(a) YOLO BB on original radiograph. (b) YOLO BB on transformed radiograph.

Figure 4.6: Example of ROI BBs before and after transforming an image with the STERN approach. The BBs and corresponding centers are marked in red, and the center of the image is marked with a green cross.

Finally, it is also desired that the positioning of the ROI is more uniform across all transformed

images. As STNs have often been employed for image alignment, it is expected that the thorax location in the final images will vary less in comparison to the original ones, meaning that the thorax shares a more consistent region in the final images. In order to corroborate this, the YOLO model is fed with the same original and transformed images to get the corresponding thorax location. Note that the YOLO model is used as a reference here to get the ROI location before and after the transformation, as no spatial GT is available for the CheXpert dataset. Based on these results, it is possible to analyze the location of the ROI before and after the STERN approach, by calculating the mean center coordinates of the resulting BBs, as shown in Figure 4.6. A lower variation in the thorax center coordinates, measured e.g. through STD, would indicate that the STERN generated images in which the thorax is overall more aligned, as its location is more homogeneous across all transformed radiographs. A summary of all the criteria proposed in this section and their respective purpose is presented in Table 4.3.

Table 4.3: Summary of the proposed criteria to evaluate the relevance of the STERN-selected ROI.

| | **Description** | **Measured on** | **Objective** |
|---|---|---|---|
| **#1** | IOU between the regions selected by YOLO and STERN. | Original images. | How significant is the overlap between both selected regions? Are the regions selected by both methods mostly coincident? |
| **#2** | Four margin distances (V1, V2, H1, H2) to the radiograph's edges, for the regions selected by YOLO and STERN. | Original images. | Does the method prioritize scaling along the vertical and/or horizontal axis? |
| **#3** | Vertical intensity projection (column average) of the radiographs. | Original and transformed images. | How does the intensity profile of the transformed radiographs differ from the original ones? What is the impact of removing the outer regions? |
| **#4** | Center coordinates of the thoracic BB using the YOLO model as an independent reference. | Original and transformed images. | Are the thorax center coordinates more uniform across the transformed radiographs, meaning these images are more aligned than the original? |

## 4.4 Results and Discussion

This section analyses the results of the three scenarios previously described, in order to determine if it is possible for the STERN to achieve similar results to using YOLO-cropped images, with less computational expense and without the need for localization labels. To do so, the analysis focuses on two main aspects: the results of the ROI selection using the YOLO model and the ST module, and the overall performance of the classifier in both of these settings. The algorithms presented in this work were implemented using the *PyTorch* framework and an Nvidia GeForce GTX 1080 GPU (8 GB).

### 4.4.1 Evaluation of ROI Selection

Both the YOLO detection model and the ST module are employed for attention, forcing the classifier to focus on the image regions most relevant for the diagnosis. The main difference between these two approaches is the fact that the first is trained independently of the classifier using spatial annotations (BBs locating the thorax), and the second is optimized simultaneously with the classifier, using exclusively the image-level binary labels. Figure 4.7 presents several examples of the resulting radiographs for both approaches, which are further discussed below. Starting with the YOLO model, an average precision of 0.998 at an IOU$>$0.5 was obtained in the JSRT/Montgomery/Shenzhen validation set. This detection network, free from any isotropic constraints, was able to perfectly delimit most left and right boundaries of the thorax, but on the other hand, the vertical cropping is not as precise, for instance in Figures 4.7(a-c,f). As a consequence, this model is often not able to eliminate certain artifacts, which are visible in Figures 4.7(a-d).

The STERN also considers non-isotropic scaling, meaning the original proportions of the image may or may not be preserved. In this study, the classifier benefited from a larger vertical scaling than a horizontal one. As such, the horizontal delimitation of the module is not as precise as in YOLO, but rather focusing its attention on the top and lower limits of the thorax. Note that the STERN is, in general, able to successfully delimit the top and bottom of the lungs, but may occasionally crop out a slight portion of those organs, as exemplified in Figures 4.7(b,d,e). This may indicate there are no statistically relevant features in those portions of the lungs, considering that the STN-based network is trained in a spatially unsupervised manner to select the ROI. Even so, the YOLO model also suffers from the same issue, as it tends to crop out a few pixels in the lateral edges of the thorax (Figures 4.7(c,e,h)). Figure 4.7(g) illustrates a case where YOLO successfully removed the artifact, whereas STERN did not. Additionally, Figure 4.7(h) presents an example where both approaches were unable to remove the artifact.

To get a more objective comparison of both methods and their selected ROI, the metrics defined in Table 4.3 are further analyzed. For instance, when plotting the selected ROIs in the original radiographs before cropping or transformation (exemplified in Figure 4.8), one concludes that the BBs of both approaches have a mean IOU of $0.781 \pm 0.120$. Although there is a significant overlap and agreement in the vast majority of selected pixels (Figure 4.8(c)), most plots reinforce that the STERN considers more relevant to define the vertical limits of the thorax rather than the horizontal ones. This is also supported by the average of both vertical and horizontal margins, whose values are displayed in Table 4.4. Once again, it is noticeable that the YOLO privileges larger horizontal margins, while the STERN privileges larger vertical margins. The latter consequently crops out more confounding factors since these are mostly located on the top of the image (Figures 4.8(b)), thus achieving a better performance when it comes to reducing FP predictions. As no artifacts are usually found in the lateral margins of the radiographs, a less precise cropping in that region is less likely to have any significant negative impact on the classifier's performance.

The vertical projections in Figure 4.9 reveal the averaged profiles of all original and transformed radiographs. Two main differences can be pointed out, starting with the intensity of the

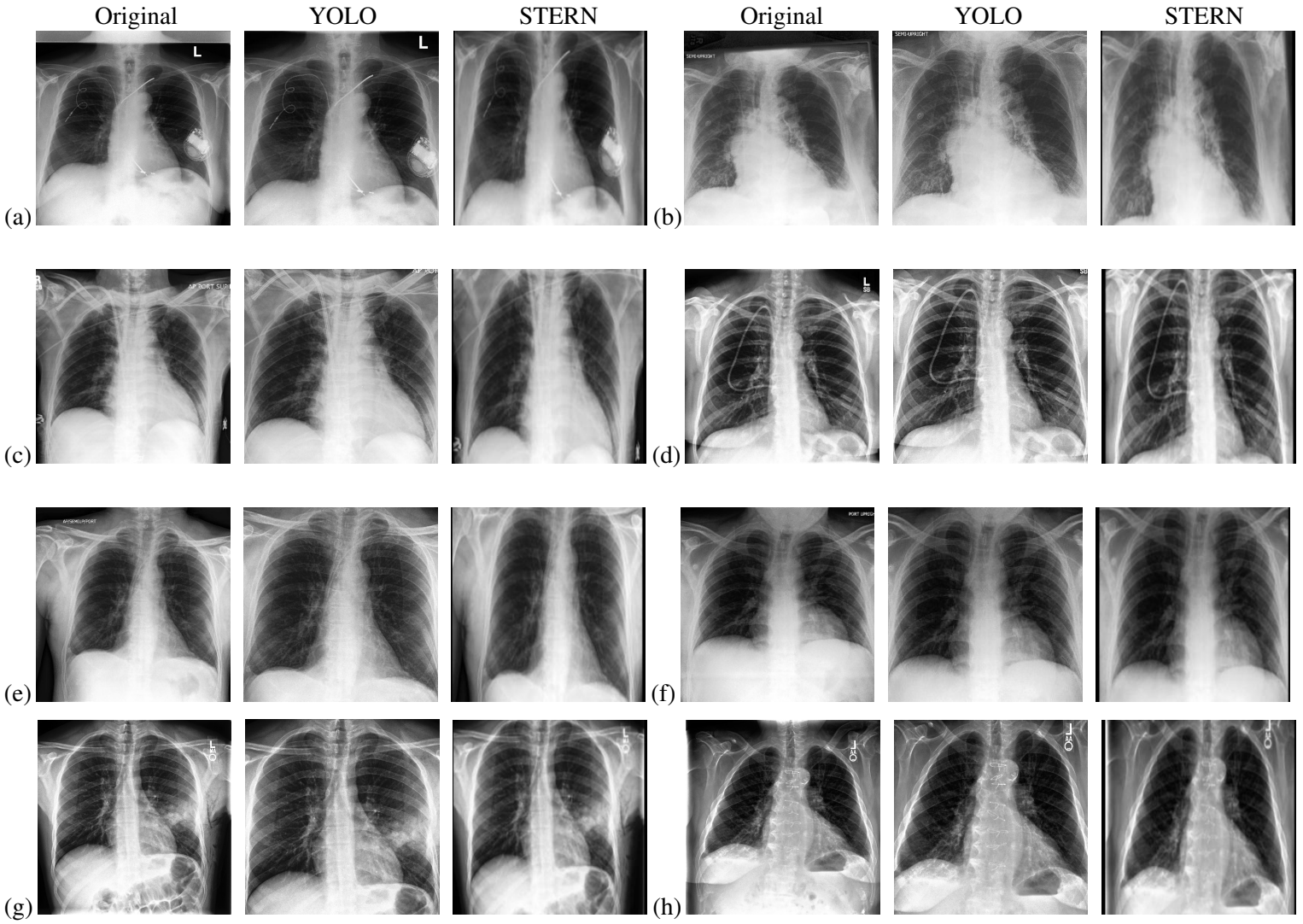
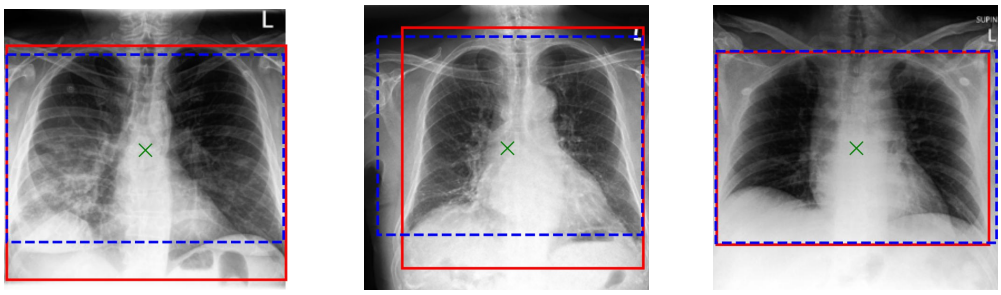



Figure 4.7: Visual comparison between the ROI selected by the proposed method and the two baselines. Each row displays an example of an original, cropped (YOLO), and transformed (STERN) images, respectively. The proposed method focuses on the vertical delimitation of the ROI.

radiographs' edges. Figure 4.9(a) shows that there is noise in the initial and final columns of the original images, causing a very discrete increase in intensity in the edges of the traced profile (black arrows). This tendency is not observed in Figure 4.9(b), since the STERN slightly crops out that lateral noise in the transformed images, thus exhibiting a much cleaner profile at the edges.

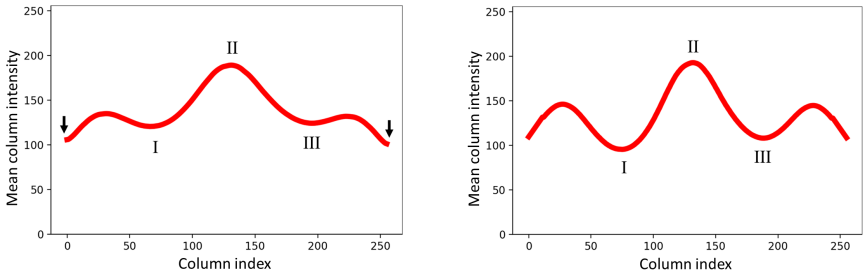
Additionally, three main sectors can be identified in both profiles. The highest peak (sector II) is centered in the images and matches the positioning of the spine and sternum, which corresponds to a high-intensity area. The other two distinctive sectors (I and III) match the positioning of the lungs, that typically present lower intensity values in the concave peaks located in columns number 50-100 and 150-200. Here, there is a clear contrast in terms of intensity between both profiles, which can be explained by the non-relevant neck/shoulder area present in the original images. This area is lighter than the lungs, thus increasing the mean intensity of the corresponding columns (I



(a) (b) (c)

Figure 4.8: Example of original radiographs with regions selected by the YOLO (red line) and STERN (blue dashed line) models before cropping and transformation. The green cross marks the center of the image.

and II in Figure 4.9(a)). In the transformed images, the noisy area above the thorax is removed, thus decreasing the mean intensity. Consequently, the second profile in Figure 4.9(b) exhibits more defined concave peaks, closer to the mean intensity of the lung pixels.




(a) Original images. (b) Transformed images.

Figure 4.9: Vertical projections of both approaches, representing the mean image intensity (vertical axis) per image column (horizontal axis). Sectors I and III correspond to the positioning of the lungs, Sector II corresponds to the positioning of the spine and sternum, while the black arrows indicate the extremities of the original image.

At last, Table 4.5 displays the analysis performed using the BB predictions on both the original and transformed radiographs, exemplified in Figure 4.10. As previously mentioned in Section 4.3.4, another goal of the STERN approach would be to produce transformed images in which the ROI presents a more homogeneous location in the final images. The results support this conclusion because the ROI center location is more uniform in the transformed radiographs, as noticeable by the lower STD values, even though the thorax BBs in the original and transformed radiographs share similar mean center coordinates. The most relevant difference in terms of STD is found in the vertical axis (YY), reinforcing the precise scaling that the STERN applies along that direction. Figures 4.10(a-f) support these conclusions, showing that the thorax occupies a central and more prominent position in the transformed images and confirming the STERN's ability to select the ROI and align the images accordingly. Figures 4.10(g,h) showcase instances where STERN struggled to accurately zoom in on the thorax, likely due to darker intensity spots below the lungs that

may have confused the model regarding their relevance for the diagnosis.

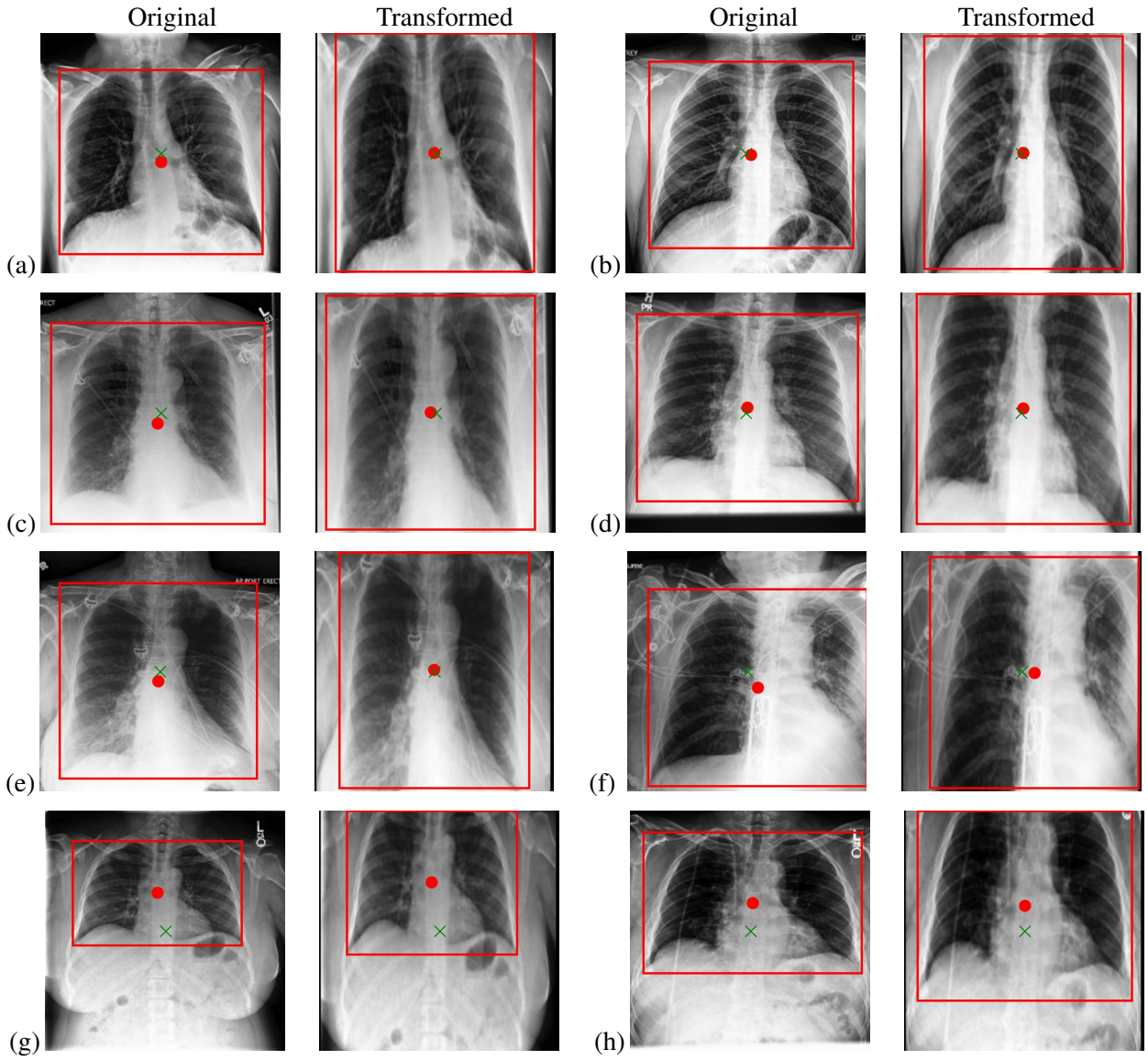



Figure 4.10: Example of original and transformed (STERN) images, respectively, with the corresponding ROI BBs as reference. These BBs and their centers are marked in red, and the center of the image is marked with a green cross.

Table 4.4: Mean length of the margins of the selected regions in the original radiographs, before cropping and transforming. Measured in pixels. Each margin is defined in Figure 4.4.

| | **Vertical Margin V1** | **Vertical Margin V2** | **Horizontal Margin H1** | **Horizontal Margin H2** |
|---|---|---|---|---|
| **Region of YOLO** | $30.2 \pm 14.5$ | $35.0 \pm 18.0$ | $19.4 \pm 13.6$ | $21.4 \pm 11.2$ |
| **Region of STERN** | $39.8 \pm 8.2$ | $40.0 \pm 11.6$ | $11.9 \pm 9.0$ | $5.5 \pm 4.0$ |

Table 4.5: Mean thoracic BB center coordinates before and after the STERN transformation. The image center coordinates are (128,128) in both cases. Measured in pixels.

| | **Center Coordinate XX** | **Center Coordinate YY** |
|---|---|---|
| **Original Images** | $126.0 \pm 10.6$ | $121.8 \pm 15.5$ |
| **Transformed Images** | $127.2 \pm 10.3$ | $121.9 \pm 10.5$ |

### 4.4.2 Classifier Performance

The VGG16 classifiers were evaluated using the following standard metrics: AUC, Precision-Recall AUC (PR-AUC), and precision, defined in Appendix A. A high recall operating point of 95% was defined to suit the medical domain premise, in which a false negative prediction is more severe than a false positive one (Jangam and Annavarapu, 2021; Hicks et al., 2022). Table 4.6 presents the final results averaged across five different test sets, including the rate of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) predictions.

It was observed that the STERN classifier outperformed both baseline settings in most established metrics: the baseline with thorax detection offers a 1.0% increase in AUC in comparison to the single classifier, while the STERN offers an increase of 2.6%. The STERN loss proved to be beneficial for fine-tuning, enabling an AUC performance increase of 0.2% versus the initial model, by recovering most of the lung portions initially cropped out of the transformed images and broadening the ROI area that is encompassed in the final radiograph. Original radiographs after the first and second training stages are displayed in Figure 4.11 to illustrate how this loss allowed to better frame the vertical limits of the ROI and consequently improve the classification results. The fine-tuning largely succeeded in preventing excessive scaling of the lung area in most radiographs, as demonstrated in Figures 4.11(a,b). However, it occasionally led to the reintroduction of artifacts, as seen in Figure 4.11(c).

Additionally and as expected, both the YOLO model and the ST module contributed to reducing the number of FP predictions by eliminating misleading image artifacts, while at the same time increasing the number of TP predictions. The latter was able to reduce the FP rate by 6.3% and increase the TP rate by 7.1%. On the other hand, both attention-focused approaches exhibited a lower rate of TN predictions and a higher rate of FN predictions in comparison to the baseline (nearly 1.0% and 0.4% difference, respectively). While the classifier is not able to discriminate specific types of abnormal findings, the top-three pathologies that most contribute to the STERN's TP rate are pleural effusion, lung opacity, and edema, and the top-three pathologies that most contribute to the FN rate are also lung opacity, atelectasis, and enlarged cardiomediastinum. The overall TP and FN rates per method and per finding are detailed in Figure 4.12. The previous conclusions are also supported by the Grad-CAM examples in Figure 4.13(a) (cardiomegaly) and Figure 4.13(b) (pleural effusion), demonstrating that the presence of artifacts has a harmful effect over the classifier's learning process, diverging its attention from the thorax.

Finally, it is also of importance to establish an objective comparison in terms of computational cost between the STERN and the baselines. Table 4.7 seeks to provide that, displaying the number

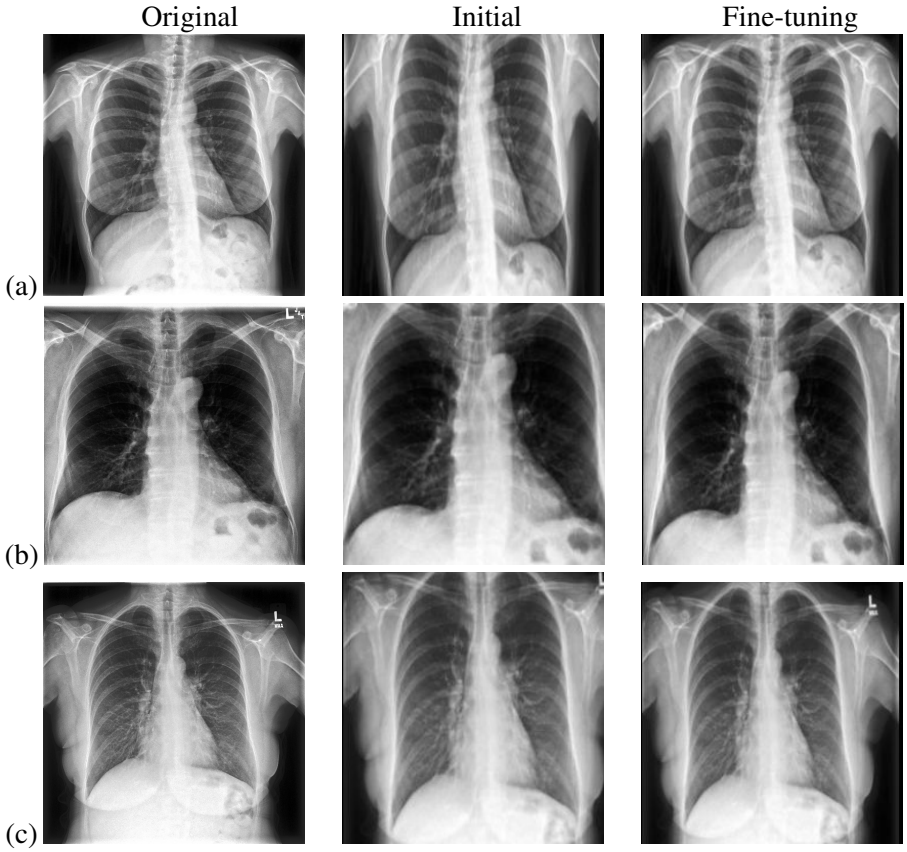


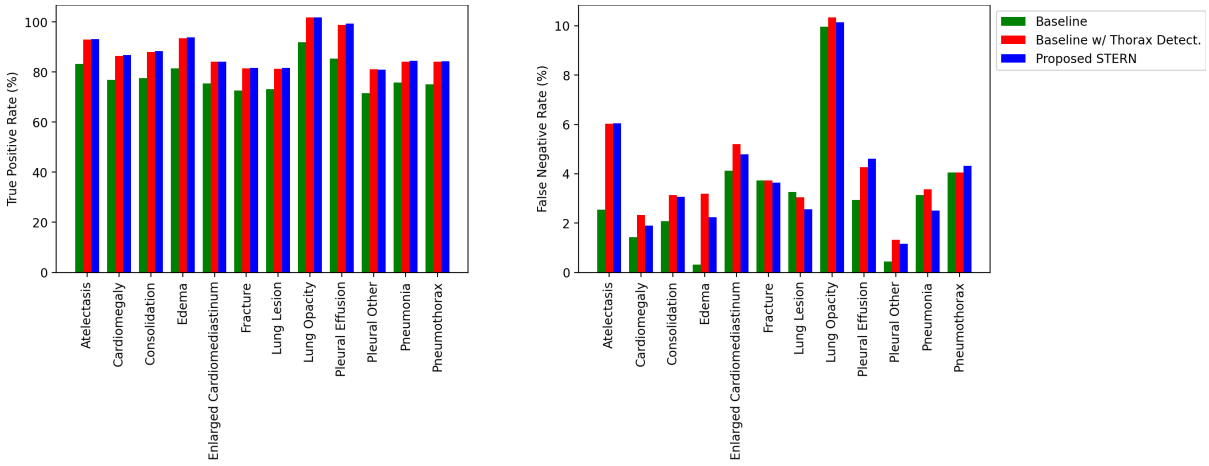

Figure 4.11: STERN transformations before and after the fine-tuning stage.

of trainable parameters of each method, as well as their inference time per batch. One may conclude that using the STERN requires nearly the same number of parameters of the classifier-only baseline and half the number of parameters of the YOLO-based architecture, while having a competitive inference time with the last one. Again, both the STERN and the YOLO-based architecture select the ROI, crop the image accordingly, and predict the instance's class, but the STERN does it in an end-to-end fashion. For these reasons, it is associated with a lower computational cost. This conclusion was also expected since the ST module is composed of four convolutional layers, whereas the YOLOv5 has a much more complex architecture.

Table 4.7: Computational cost analysis of the three methodologies. Inference time is calculated per batch of 8 images, in an Nvidia GeForce GTX 1080 GPU (8 GB).

| | **# Trainable Parameters (M)** | **Inference Time (ms)** |
|---|---|---|
| **Baseline** | 134.2 (VGG16) | 15.614 |
| **Baseline with Thorax Detection** | 220.8 (YOLO and VGG16) | 373.615$^{\star}$ |
| **STERN** | 135.1 (ST module and VGG16) | 375.016 |

$^{\star}$ Consecutive prediction of BB, cropping of the image accordingly, and prediction using the baseline classifier.




(a) (b)

Figure 4.12: True positive and false negative rates per method and finding.

Direct comparisons with the state of the art are challenging, due to the lack of publications addressing the same binary classification setting with an STN-based approach and/or the same data. Nevertheless, there are two publications that employ an STN to distinguish between normal and abnormal radiographs (Bharati et al., 2020; Rocha et al., 2022). The first reports a precision increase of 1.0% between their baseline and the STN using the ChestX-ray14 dataset, but does not disclose AUC scores. The second minds a subset of CheXpert and achieves a 1.0% increase in AUC and an increase of 0.9% in precision, in comparison to the baseline classifier. As mentioned earlier, this work increases the AUC by 2.6% and the precision by 7.0% when using the STERN approach.

Regarding abnormality detection in general, most papers focused on CXR imaging which are cited in Section 4.2 (Yates et al., 2018; Tang et al., 2020; Nakao et al., 2021; Mao et al., 2020; Shvetsova et al., 2021) employ the ChestX-ray14 dataset (or a subset of it), in opposition to the more recent CheXpert one. The Indiana University and/or the pediatric Kermany datasets are also considered by a few publications (Tang et al., 2020; Karthik and Sowmya Kamath, 2022). The justification for choosing CheXpert over ChestX-ray14 in this study is grounded in Chapter 2: both cases generate most of the annotations through natural language processing labelers, but their overall consistency and reliability are presumably different, with CheXpert exhibiting a lower labeling error rate.

To the best of the author's knowledge, there is no state-of-the-art publication that specifically addresses the evaluation of the ROIs selected by an STN in a spatially unsupervised way, particularly using objective metrics. The proposed work also differs from the state of the art in that sense, as it provides not only an analysis of the classifiers' performance but also an analysis of the attention-driven components and their outcomes.

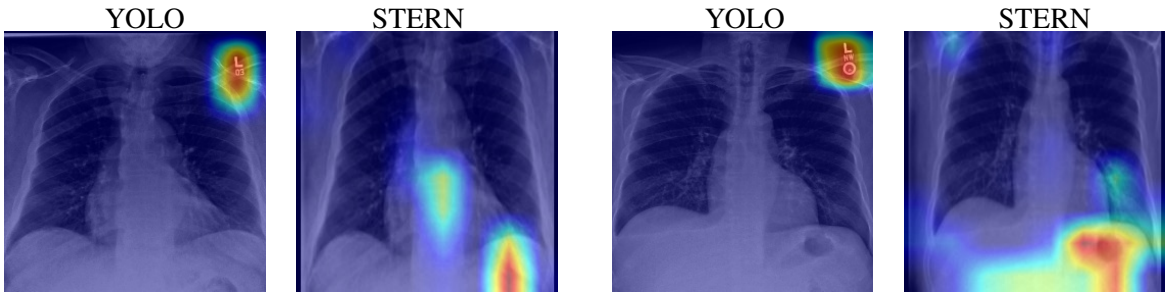
To conclude, the main distinction from the STN-based methods proposed in Liu et al. (2019c), Bharati et al. (2020), and Kim et al. (2023) is the restriction in allowed transformations to encompass only non-isotropic scaling and translation operations, which enabled the attention-driven

Table 4.6: Classification results of the baselines and STERN method. The best result for each metric is shown in bold.

| | **AUC** | **PR-AUC** | **Precision** |
|---|---|---|---|
| **Baseline** | 0.831±0.005 | 0.977±0.001 | 0.870±0.004 |
| **Baseline with Thorax Detection** | 0.841±0.005 | 0.983±0.006 | 0.939±0.004 |
| **Proposed STERN** | **0.857**±0.003 | **0.989**±0.001 | **0.940**±0.000 |

| | **TP Rate** | **FP Rate** | **TN Rate** | **FN Rate** |
|---|---|---|---|---|
| **Baseline** | 0.795±0.001 | 0.119±0.004 | **0.043**±0.004 | **0.042**±0.000 |
| **Baseline with Thorax Detection** | **0.866**±0.000 | 0.057±0.002 | 0.032±0.002 | 0.046±0.000 |
| **STERN** | **0.866**±0.000 | **0.056**±0.000 | 0.033±0.000 | 0.046±0.000 |

component of the work while using a much simpler architecture for the module and no spatial GT. Additionally, the work in Rocha et al. (2022) has been improved with the proposed STERN approach by adding more radiographs from the CheXpert dataset, an additional training stage with a domain-specific loss, and a relevant analysis of the selected ROIs to better comprehend the capability of the models to be used for attention purposes. The obtained results have been compared with two baselines as well, but unlike in Rocha et al. (2022) where the YOLO had an isotropy constraint on the BBs, this condition was removed on this study's baseline for a fairer comparison.




(a) Patient with cardiomegaly.

(b) Patient with pleural effusion.

Figure 4.13: Grad-CAM examples of cropped (YOLO) and transformed (STERN) images, respectively.

## 4.5 Concluding Remarks

The growing ability of deep learning models to analyze medical images has been decisive for the advances of computer-aided diagnosis systems. These can be employed in a clinical setting to prioritize more urgent cases, for instance by implementing an automated binary classification model that detects thoracic abnormalities, defining a triage step that is useful for patient management and follow-up. However, among other challenges, the presence of image artifacts may mislead the model into mostly false positive predictions. For this reason, an attention-driven approach is considered in this work as a means of restricting the learning process exclusively to the area of interest for the diagnosis. In CXR data analysis, this becomes particularly relevant to ensure that the classifier is in fact making predictions based on the correct anatomical region.

The proposed STERN approach can be described as a convolutional neural network, composed of an attention-driven spatial transformer module and a classifier, which is able to select the thoracic ROI, transform the input image accordingly, and then classify it as normal or abnormal. A novel domain-specific loss function is proposed to better frame the ROI in a fine-tuning stage. This method was compared to a more traditional approach, in which a YOLO object detection model is trained independently of the classifier and used as an *a priori* processing stage to select the thorax. The main findings of this study indicate that the STERN is able to delimit the ROI in a non-isotropic way that benefits the most the performance of the classification task, thus outperforming the YOLO-based method. The latter privileges a more accurate delimitation of the lateral margins of the radiograph, while the proposed architecture focuses mostly on identifying the top and bottom limits of the thorax. The STERN-transformed images also exhibit a cleaner intensity profile and are more homogeneous regarding the thorax position across multiple radiographs, in comparison to the original images.

In sum, the merits of the proposed work lie in the STERN architecture, which does not need any spatial GT labels and is trained end-to-end exclusively on image-level annotations. The four convolutional-layer ST module is also computationally more efficient in comparison to the YOLO and able to crop out more confounding factors, such as lettering, which are mostly located on the top of the images. On the other hand, note that this work targets only image artifacts found outside the ROI, and does not consider the artifacts that may be present within that region, such as arrows or other annotations. It is also worth mentioning that it is not intended to display the distorted STN images to the radiologists during their diagnostic process, in order to prevent possible misdiagnoses. These images are exclusively intended to be used in an internal step of the computer-aided diagnosis system, to further improve this binary triage stage.

# Chapter 5

# Confident-CAM: Improving Heat Map Interpretation in CXR Image Classification

**The content of this chapter results from the following publication. The author of this thesis has the main contribution to the conception of the study, method development, and results assessment.**

**Rocha, J.**, Pereira, S. C., Campilho, A., and Mendonça, A. M. (2023). Confident-CAM: Improving Heat Map Interpretation in Chest X-Ray Image Classification. In 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 4116-4123). IEEE.

Further implementation details of Confident-CAM can be found in the **GitHub repository** provided by the author. https://github.com/JoanaNRocha/Confident-CAM

The integration of explanation techniques promotes the comprehension of a model's output and contributes to its interpretation e.g. by generating heat maps highlighting the most decisive regions for that prediction. However, there are several drawbacks to the current heat map-generating methods. Probability by itself is not indicative of the model's degree of confidence in a prediction, as it is influenced by multiple factors, such as class imbalance. Thus, even if a model correctly predicts the outcome, it does not necessarily generate an accurate explanation map. In other words, the model can arrive at the correct prediction for the wrong reasons, leading to an explanation that does not align with the intended factors for the application at hand. Current state-of-the-art explanations are not able to distinguish both scenarios and alert the user to dubious explanations. The goal of this chapter is to improve on the Grad-CAM technique, by representing its maps more intuitively based on how confident the model is regarding the diagnosis. This works by adding an extra validation step over its initial results to indicate whether the user should trust the initial explanation or not. The proposed method, Confident-CAM, facilitates the interpretation of the results by measuring the distance between the output probability and the corresponding class

threshold, using a confidence score to generate nearly null maps when the initial explanations are most likely incorrect. This study implements and validates the proposed algorithm on a multi-label CXR classification exercise, targeting the 14 radiological findings in the ChestX-ray14 dataset with significant class imbalance. Results indicate that confidence scores can distinguish likely accurate and inaccurate explanations.

## 5.1 Introduction

The interpretation of Grad-CAMs may be quick, but it is hindered by multiple factors, including their low resolution and normalization process. In a clinical setting, their performance is often affected by severe class imbalance, as certain findings occur much more frequently than others, further complicating a robust evaluation, specially when spatial GT annotations are scarce (Qu et al., 2020).

The optimization of the classification thresholds plays a vital role in this scenario to handle the imbalanced distribution, defining the probability value above which an instance is classified with a certain pathology. This class estimation can be done using the ROC or PR curves. In practice, these models are often optimized for recall by lowering the prediction threshold, which increases the number of instances classified as positive (Esposito et al., 2021). This approach favors FPs over FNs, which are more severe in a medical environment (Hicks et al., 2021). However, false positives are generally uncertain "borderline" cases, where the output probability hovers close to the optimal class threshold.

This work considers a multi-label classification scenario, in which a CXR image can be annotated with 0, 1, or multiple classes of pathologies. The goal here is to use the information of the distance between the output probability and the corresponding optimal class threshold to improve the XAI maps' interpretation, i.e. to represent these maps more intuitively based on how "sure" the model is regarding a specific diagnosis. This will allow the user to more easily interpret and compare maps that would otherwise be normalized in the same color range, whether the model is more or less certain about that diagnosis. The proposed methodology was termed Confident-CAM and adds an extra validation step to alert the user to potential inaccurate heat maps. The main contributions of this chapter are listed below:

1. Definition of a confidence score metric that represents the degree of confidence of the model on a given class prediction.

2. Creation of a novel XAI map approach that incorporates that information to distinguish potential inaccurate explanations.

3. Validation study for the proposed method's results.

The following section provides a review of state-of-the-art publications on the subject. Section 5.3 offers a detailed overview of the dataset and proposed methodology, with the corresponding results being discussed in Section 5.4. Finally, the key conclusions are summarized in Section 5.5.

## 5.2 Related Work

The vast majority of the articles on multi-label CXR classification employs standard off-the-shelf pre-trained networks, either individually (Kumar et al., 2018; Urinbayev et al., 2020) or in ensemble structures (Chen et al., 2019a; Jeong et al., 2023). Even without incorporating spatial information, these classification architectures are reported to effectively capture discriminative image features. Other authors suggest a weakly supervised strategy to leverage the limited available data and enhance the model's sensitivity to subtle differences in disease-specific features (Yan et al., 2018). Among these, Zhang et al. (2020) proposes to compare an unannotated observation with images of known pathology, measuring their similarity to distinguish between classes. Another common strategy is to use attention learning to selectively focus on relevant image regions for diagnosis. This improves classification performance by concentrating on the lesion areas and ignoring the surrounding irrelevant information (Guan and Huang, 2020; Wang et al., 2021). Recently, more complex structures have been studied to tackle this exercise. For instance, Chen et al. (2020) employs a Graph Convolutional Network (GCN), and Sun et al. (2023b); Wu et al. (2024); Jiang et al. (2024) employ ViTs. The latter have become very popular for better capturing global and local information of the images and/or enhancing the richness and diversity of feature representations, e.g. through spatial attention.

When searching for papers addressing classification in CXR images, introducing explainability as a prerequisite significantly reduces the number of relevant studies found. This reinforces both the relatively recent emergence of XAI in the field and the growing need for further research in this area. Yet, it is clear that most published works in this area rely on attention-based heat map explanations, as previously discussed in Chapter 3 and evident in several cited studies (Urinbayev et al., 2020; Yan et al., 2018; Guan and Huang, 2020; Wang et al., 2021). Other examples include Brunese et al. (2020); Karim et al. (2020); Padmaja et al. (2023), which opted for Grad-CAM and/or Grad-CAM++, and Ahsan et al. (2020); Koyyada and Singh (2023), which also considered LIME to identify the superpixels that contributed the most to the predicted label. Other works have taken this further by incorporating XAI maps into the training routine. These approaches leverage explainability maps to guide the model's learning, using information on the likely location of findings to enhance performance without relying on BB annotations (Purwita and Qomariyah, 2021; Kabiraj et al., 2024).

The present study focuses on Grad-CAM as the current state-of-the-art technique, due to its prominence in the literature. While it is a fast and straightforward method, several drawbacks have been identified over the years. For instance, this technique produces low-resolution maps as a result of down-scaling the input image, with accentuated blur/dispersion (Adebayo et al., 2018; Schöttl, 2022). Additionally, the 0-1 normalization of the maps also poses a great disadvantage. It can obscure the relative importance of different regions within an image, possibly leading to a loss of fine-grained information about which parts of the image contribute most to the model's decision. Furthermore, normalization scales all heat maps to the same range and thus distorts the relative importance of regions among multiple images, making it difficult to compare their high-

lighted features. For instance, an area of moderate significance in one image might appear more prominent after normalization than a highly significant area in another, misleading the comparison of the maps and the overall interpretation of model behavior across different inputs. In sum, these disadvantages may lead to incorrect conclusions about the model's decision-making process.

Several publications aimed to improve this method by increasing its spatial sensitivity, e.g. using gradient-free algorithms (Wang et al., 2020a) or considering different ways to weight the feature maps (Chattopadhay et al., 2018; Sattarzadeh et al., 2021; Yamauchi and Ishikawa, 2022). Alternatively, Barkan et al. (2021) addresses negative gradient impression and suggests extracting information from multiple layers, and not exclusively from the last convolutional one. Others tackle specifically the low-resolution issue and seek to output higher-resolution maps with a more precise location of the objects of interest (Draelos and Carin, 2021). The work in Schöttl (2022) reduced the dispersion and blur of the highlighted areas by incorporating the maps' information in an additional training loss term.

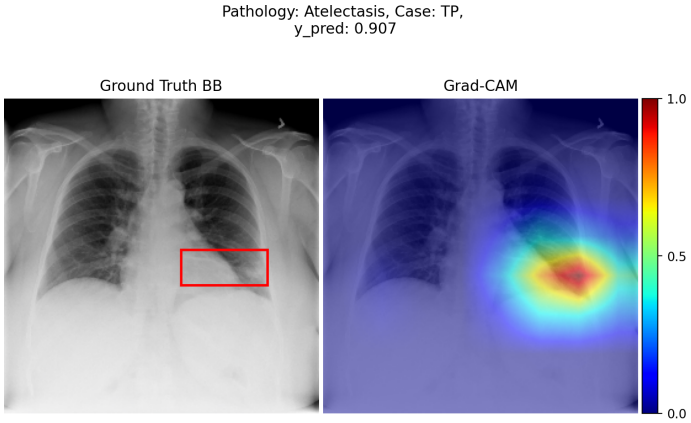
Grad-CAM often outperforms other approaches in the medical domain (Zhang et al., 2021). For this reason, it is still used in many applications besides CXR imaging, such as brain magnetic resonance (Zhang et al., 2021) and histopathological imaging (Afify et al., 2023). Alternative Grad-CAM-based methods in this domain's state of the art include the combination of multiple XAI maps (Abeyagunasekera et al., 2022), or enhancing the original Grad-CAM results by adding numerical explanations over the heat maps to further quantify the importance of each region for the diagnosis (Marmolejo-Saucedo and Kose, 2022). To the best of the author's knowledge, no published paper specifically seeks to improve the interpretation of normalized maps, which is the focus of the current publication.

## 5.3 Method

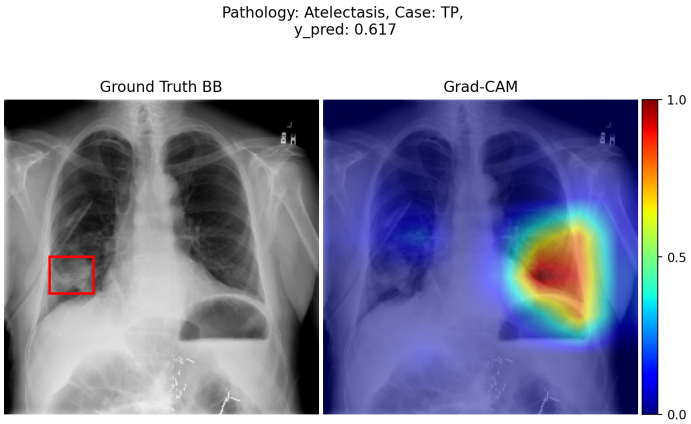
### 5.3.1 Goals and Premises

Grad-CAMs are not directly comparable between themselves, due to their 0-1 normalization. As exemplified in Figure 5.1 for atelectasis, a classifier can make two TP predictions with very distinct output probabilities. Still, in the second case, it is more "unsure" about its decision (i.e. lower probability value) and the explanation fails, pointing out a wrong location. Even so, both maps are still represented in the same 0-1 color scale, independently of the confidence of the model, which may not be ideal in terms of user interpretation.

In fact, prediction probability alone is not indicative of the model's sureness regarding a specific instance, because different pathologies have different optimal class thresholds. For instance, a probability of 0.7 given a 0.6 threshold is a much less confident positive prediction, than the same probability given a 0.3 threshold, because the distance to the threshold is lower in the first scenario. An imbalanced setting significantly impacts the optimal threshold values, as less represented pathologies will be assigned lower thresholds to favor FPs over FNs.




(a)




(b)

Figure 5.1: Example of two TP instances for atelectasis, in which the model activates on the correct and incorrect regions, respectively. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability.

The work described in this chapter focuses on modifying the original Grad-CAM heat maps to add information regarding the distance of the probability output to the optimal class threshold, in order to visually represent the confidence of the model in that specific outcome. The author proposes to do so by resorting to a confidence score that rescales the probability per instance per class ($y^c_{pred}$), based on its distance from the class-specific threshold ($Th^c$), as shown in Equation 5.1 and graphically represented in Figure 5.2. This way, considering only the positive predictions (i.e. $y^c_{pred} > Th^c$), each instance will be attributed a confidence score per class between 0 and 1, expressing the degree of confidence of the model about that pathology.

$$\text{Confidence Score} = \frac{1}{1 - Th^c} * (y^c_{pred} - Th^c) \tag{5.1}$$

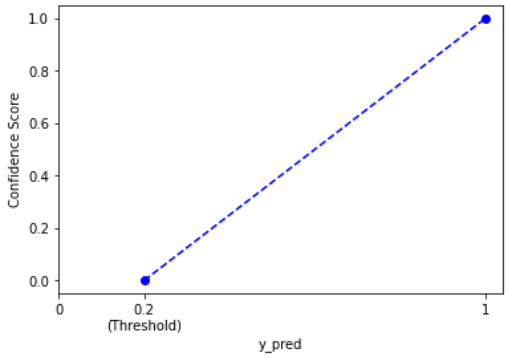



Figure 5.2: Confidence score calculation based on the rescaling of the predicted probability (*y_pred*), minding the hypothetical class-specific threshold.

### 5.3.2 Confident-CAM Algorithm

The Confident-CAM algorithm can be implemented over any XAI heat map method that is normalized between 0 and 1, on instances predicted as positive for the class(es) of interest. In this study, the author has considered Grad-CAM, which obtains the class-discriminative localization map $L^c_{Grad-CAM}$ by computing the gradient of $y^c$ (score for class $c$ before the sigmoid activation) with respect to the feature map activations $A^k$ of the last convolutional layer. These gradients are global-average-pooled over the width and height dimensions (indexed by $i$ and $j$, respectively) to obtain the importance weights $\alpha^c_k$ in Equation 5.2. Note that $Z$ is the number of pixels in the activation map.

$$\alpha^c_k = \frac{1}{Z}\sum_i\sum_j \frac{\delta y^c}{\delta A^k_{ij}} \tag{5.2}$$

The final Grad-CAM heat map is a weighted combination of the feature maps, followed by a ReLU activation function (Equation 5.3), which is then upsampled to the original input image size.

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k) \tag{5.3}$$

Let us assume that for each image, a classification model will predict a certain class $c$ and the corresponding Grad-CAM. The proposed algorithm starts by calculating the optimal class threshold based on the ROC Curve, i.e. maximizing TP cases and minimizing FP ones. Then, the confidence score (per instance, per class) is determined using Equation 5.1. The Confident-CAM is then obtained according to Equation 5.4, by multiplying the initial Grad-CAM by that confidence score. The resulting heat map will be in the range [0, 1] since both the confidence score and the Grad-CAM lie within that range. The overall workflow is represented in Figure 5.3, for a FP cardiomegaly example.

$$L^c_{Confident-CAM} = L^c_{Grad-CAM} * \text{Confidence Score} \tag{5.4}$$

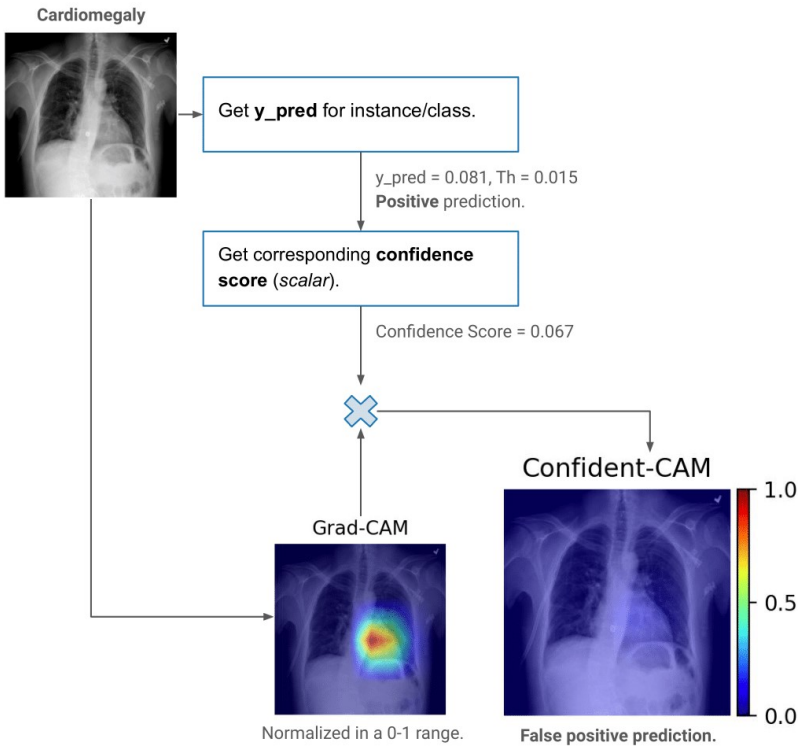


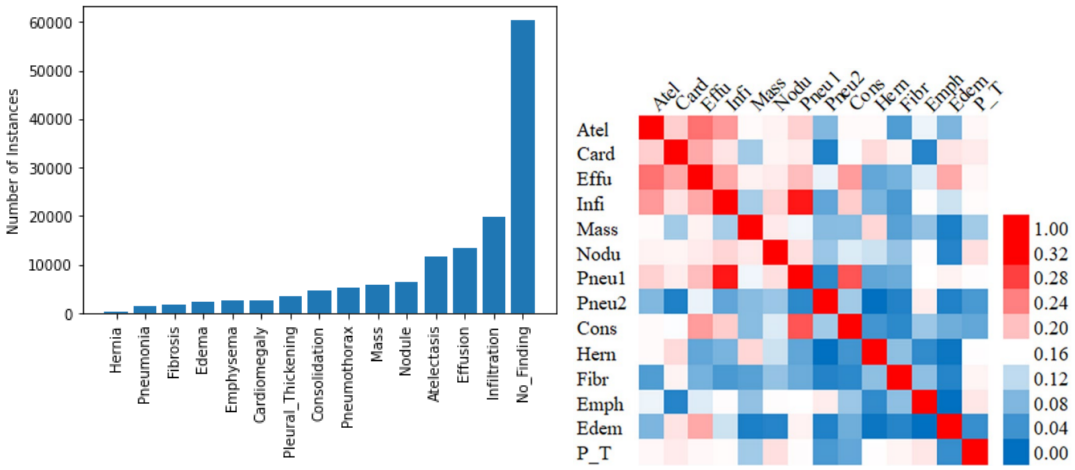

Figure 5.3: Confident-CAM algorithm workflow. Example for a FP cardiomegaly prediction (*y_pred*).

### 5.3.3 Dataset Preprocessing

In the ChestX-ray14 dataset, an image can be annotated with one or more of these 14 classes: atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia. The total number of instances per pathology reported in ChestX-ray14 is represented in Figure 5.4(a), in which the *No Finding* observation is assigned if none of the pathologies are present. The high imbalance between the multiple classes is evidenced in that figure. Besides that, certain classes are highly correlated, as can be observed in Figure 5.4(b) from Chen et al. (2019a).

As previously discussed in Chapter 2, the dataset includes a subset of 984 BBs, covering the first eight pathologies. This availability of BBs is the key reason for choosing the ChestX-ray14 dataset over CheXpert in this study, since these GT BBs are essential for evaluating the areas highlighted in the explanations.

All available radiographs were resized to 256×256 pixels to accelerate the training routine, and the data was split into five folds, preserving the original pathology proportion and without patient overlap. More specifically, a five-fold cross-validation scheme was implemented in all experiments, in which three folds were used for training, one for validation, and one for testing. The images with spatial annotations were kept as an independent test set and not considered in the cross-validation. Further details on the optimization routine are given in the following section.




(a) (b)

Figure 5.4: Number of instances per pathology in the ChestX-ray14 dataset (left), and their respective pathology correlation matrix (right, Chen et al., 2019a).

### 5.3.4 Multi-label Classifier

Previous work in Xu et al. (2020) compared multiple DL models that could be used for transfer learning on ChestX-ray14, and concluded that ResNet demonstrated a good potential for fine-tuning in this domain. Not only that, the preliminary work in Section 3.5 indicated that a ResNet50 model is capable of providing reliable explanations for this specific dataset, with competitive predictive performance. As such, the baseline model corresponds to the implementation of a standard multi-label classifier, a widely-used ResNet50 architecture, whose last fully-connected layer was modified to consider 14 output nodes. This layer is followed by a sigmoid activation function to yield the corresponding predictions. The model was pre-trained on ImageNet, and for this reason, all images were normalized initially to a 0-1 range and then standardized using ImageNet's mean and STD values. Note that the training data was also augmented to double the samples through soft random affine transformations (i.e. 5 degrees of rotation and 3 degrees of shear).

A BCE loss was employed in the optimization routine and validated on 6,400 image steps over a total of 100 epochs, using a batch size of 8 images. This is a standard practice when dealing with large datasets that converge rapidly in the first epochs, as is the case here. The Adam optimizer was initialized with a learning rate of $10^{-4}$, which was then reduced on plateau with a patience of 3 epochs. Multiple architectures and hyperparameters were tested, namely the ones in Section 3.5, but ultimately the conditions described above provided the most suitable classification baseline for this specific learning task.

## 5.4 Results and Discussion

In order to understand the results of the proposed Confident-CAM method, it is first necessary to analyze the model's performance for this particular ChestX-ray14 distribution. With that in mind,

Table 5.1 provides the AUC values and the frequency of each pathology in the dataset. Overall, the model is able to successfully distinguish the 14 classes, with a mean AUC of 0.822. Table 5.2 compares the predictive performance of the proposed ResNet50 model with state-of-the-art approaches discussed in Section 5.2 that use the same dataset. The consistent use of the AUC metric in these studies prompted its adoption in this analysis for a fair comparison. Despite its simplicity, the ResNet achieves comparable AUC scores, even against more complex architectures, such as the ViTs recently introduced in Wu et al. (2024) and Jiang et al. (2024). Remarkably, the ResNet even outperforms one of these transformer-based models, underscoring its efficiency and competitive edge over other approaches. This highlights that while transformer architectures offer increased complexity, they do not necessarily translate into significantly better performance. The ResNet, despite its straightforward design, proved to be a strong contender. Note that the evaluation methodology varies across the cited publications. For instance, some studies, like this one, employ cross-validation by rotating the test set, while others rely on a single independent test set provided by the NIH. This difference in evaluation schemes introduces variability in the performance metric, hindering direct comparisons. Nevertheless, the overall consistency in performance across the range of reported values indicates that the proposed approach performs as expected and reinforces the reliability of the model's results.

The mean confidence scores per class are provided in the last column of Table 5.1. The first conclusion one may draw is that less-represented classes tend to have a lower confidence score. This seems intuitive, as the model has seen fewer examples of such pathologies and consequently will not be as discerning regarding their distinctive features when establishing a diagnosis. It is also clear that less represented classes are associated with lower optimal threshold values (as incorrect positive diagnoses are preferred over incorrect negative ones).

Table 5.1: Classification results of the ResNet50 model.

| **Pathology Class** | **AUC** | **Class Frequency (%)** | **Optimal Threshold** | **Mean Confidence Score** |
|---|---|---|---|---|
| **Atelectasis** | 0.813 | 14.2 | 0.154 | 0.297 |
| **Cardiomegaly** | 0.903 | 3.42 | 0.015 | 0.154 |
| **Effusion** | 0.875 | 16.4 | 0.187 | 0.402 |
| **Infiltration** | 0.710 | 24.5 | 0.202 | 0.203 |
| **Mass** | 0.822 | 7.1 | 0.031 | 0.204 |
| **Nodule** | 0.755 | 7.8 | 0.068 | 0.138 |
| **Pneumonia** | 0.765 | 1.8 | 0.012 | 0.017 |
| **Pneumothorax** | 0.854 | 6.5 | 0.038 | 0.286 |
| **Consolidation** | 0.793 | 5.7 | 0.060 | 0.072 |
| **Edema** | 0.887 | 2.8 | 0.010 | 0.079 |
| **Emphysema** | 0.891 | 3.1 | 0.026 | 0.149 |
| **Fibrosis** | 0.797 | 2.1 | 0.019 | 0.042 |
| **Pleural_Thickening** | 0.774 | 4.2 | 0.034 | 0.094 |
| **Hernia** | 0.872 | 0.3 | 0.001 | 0.009 |
| **Mean** | **0.822** $\pm$0.057 | | | |

The trained ResNet50 was used to generate the class-wise predictions and the corresponding XAI maps, i.e. the Grad-CAMs and the Confident-CAMs, when a certain pathology is considered

Table 5.2: Comparison of AUC scores for state-of-the-art methods evaluated on the ChestX-ray14 dataset. The reported scores were obtained minding different evaluation strategies/conditions across studies.

| **Method** | **Atelectasis** | **Cardiomegaly** | **Effusion** | **Infiltration** | **Mass** | **Nodule** | **Pneumonia** | **Pneumothorax** |
|---|---|---|---|---|---|---|---|---|
| Kumar et al. (2018) | 0.762 | 0.913 | 0.864 | 0.692 | 0.750 | 0.666 | 0.715 | 0.860 |
| Chen et al. (2019a) | 0.784 | 0.888 | 0.831 | 0.705 | 0.838 | 0.796 | 0.727 | 0.876 |
| Urinbayev et al. (2020) | 0.810 | 0.910 | 0.870 | 0.720 | 0.850 | 0.780 | 0.740 | 0.900 |
| Zhang et al. (2020) | 0.845 | 0.905 | 0.877 | 0.817 | 0.859 | 0.824 | 0.804 | 0.871 |
| Guan and Huang (2020) | 0.781 | 0.883 | 0.831 | 0.697 | 0.830 | 0.764 | 0.725 | 0.866 |
| Chen et al. (2020) | 0.786 | 0.893 | 0.832 | 0.699 | 0.840 | 0.800 | 0.739 | 0.876 |
| Wang et al. (2021) | 0.779 | 0.895 | 0.836 | 0.710 | 0.834 | 0.777 | 0.737 | 0.878 |
| Sun et al. (2023b) | 0.789 | 0.889 | 0.842 | 0.694 | 0.843 | 0.803 | 0.742 | 0.896 |
| Wu et al. (2024) | 0.742 | 0.918 | 0.717 | 0.838 | 0.820 | 0.811 | 0.783 | 0.890 |
| Jiang et al. (2024) | 0.791 | 0.885 | 0.842 | 0.715 | 0.837 | 0.803 | 0.745 | 0.885 |
| **Proposed model** | **0.813** | **0.903** | **0.875** | **0.710** | **0.822** | **0.755** | **0.765** | **0.854** |

| **Method** | **Consolidation** | **Edema** | **Emphysema** | **Fibrosis** | **Pleural Thickening** | **Hernia** | **Mean** |
|---|---|---|---|---|---|---|---|
| Kumar et al. (2018) | 0.784 | 0.888 | 0.898 | 0.756 | 0.774 | 0.802 | 0.795 |
| Chen et al. (2019a) | 0.746 | 0.852 | 0.942 | 0.837 | 0.796 | 0.912 | 0.823 |
| Urinbayev et al. (2020) | 0.790 | 0.910 | 0.920 | 0.810 | 0.790 | 0.990 | 0.842 |
| Zhang et al. (2020) | 0.810 | 0.862 | 0.896 | 0.849 | 0.829 | 0.927 | 0.855 |
| Guan and Huang (2020) | 0.758 | 0.853 | 0.911 | 0.826 | 0.780 | 0.918 | 0.816 |
| Chen et al. (2020) | 0.751 | 0.850 | 0.944 | 0.834 | 0.795 | 0.929 | 0.826 |
| Wang et al. (2021) | 0.759 | 0.855 | 0.933 | 0.838 | 0.791 | 0.938 | 0.826 |
| Sun et al. (2023b) | 0.757 | 0.858 | 0.944 | 0.863 | 0.799 | 0.915 | 0.831 |
| Wu et al. (2024) | 0.684 | 0.763 | 0.697 | 0.674 | 0.596 | 0.820 | 0.768 |
| Jiang et al. (2024) | 0.753 | 0.859 | 0.944 | 0.849 | 0.803 | 0.924 | 0.831 |
| **Proposed model** | **0.793** | **0.887** | **0.891** | **0.797** | **0.774** | **0.872** | **0.822** |

| **Method** | **Validation Strategy** |
|---|---|
| Kumar et al. (2018) | Holdout (80% training set, 20% test set) |
| Chen et al. (2019a) | Holdout (70% training set, 10% validation set, 20% test set) |
| Urinbayev et al. (2020) | Holdout (70% training set, 20% validation set, 10% test set) |
| Zhang et al. (2020) | 10-fold cross-validation |
| Guan and Huang (2020) | Holdout (77% training set, 23% test set) |
| Chen et al. (2020) | Holdout (70% training set, 10% validation set, 20% test set) |
| Wang et al. (2021) | Holdout (77% training set, 23% test set) |
| Sun et al. (2023b) | Holdout (77% training set, 23% test set) |
| Wu et al. (2024) | Unclear. Do not use the full dataset. |
| Jiang et al. (2024) | Holdout (80% training set, 10% validation set, 10% test set) |
| **Proposed model** | 5-fold cross-validation |

to be present in that instance. Several examples are displayed in Figures 5.6-5.9, given their individual probability output and confidence score values. Figure 5.6 shows FP predictions, which typically exhibit a low confidence score. In these cases, the Grad-CAM maps are inherently mistaken, as there are no annotated lesions to highlight, even though the model indicates so. Consequently, the weighted component of the Confident-CAM method clears most maps, providing a more accurate explanation - an empty map. On the other hand, Figure 5.7 represents TP predictions, associated with an expected high confidence score (the GT is consistent with the model output). The weighted component does not significantly impact these maps, preserving the orig-

inal correct explanation. However, there is also the possibility that TP predictions are associated with lower confidence scores, as seen in Figure 5.8. Here, the model accurately predicted the instance class but provided incorrect explanations maps (i.e. not pointing to the lesion's GT location). In this scenario, the confidence score becomes useful to filter these maps by "deleting" highlighted regions that are most likely imprecise.

In practice, both the Grad-CAMs and the Confident-CAMs should be provided to the physicians, so they can better assess the predicted outcome by comparing the original highlighted pathologic areas with the newly obtained maps. For instance, when provided with a low-activated Confident-CAM (Figures 5.6 and 5.8), the radiologist is informed not to trust the initial Grad-CAM explanation, and thus take a double look on this instance's diagnosis to validate it further. On the other hand, when provided with a high-activated Confident-CAM (Figure 5.7), the radiologist knows that the predicted outcome is most likely correct and can be trusted for interpretation as it is. Note that Confident-CAMs do not establish whether a prediction is "right" or "wrong", but rather ask the physicians for an extra validation step on the areas selected as potentially pathologic. A low-confidence CAM asks the doctor to be more careful and double-check on this class output.

The proposed method is effective in scenarios characterized by significant class imbalance, where each class has a distinct optimal threshold. It enhances the interpretability of the original probabilities based on the specific threshold value, and thus enables the comparison of the final maps in terms of confidence in the diagnosis, e.g. pointing out that one pathology is more likely to be present than other, which was not possible using the Grad-CAMs exclusively. In contrast, in a balanced class setting, rescaling the maps using confidence scores is unnecessary, as all classes display similar optimal thresholds. In this case, the maps would be equally weighted, offering no additional information. Overall, there are three possible scenarios for the XAI map outcomes, summarized in Figure 5.5:

- **TP cases with high confidence scores** are associated with trustworthy explanations.
- **TP cases with lower confidence scores** are associated with unreliable explanations (most likely highlighting the wrong anatomical regions).
- **FP cases with lower confidence scores** are associated with unreliable predictions and consequently unreliable explanations (there are no lesions related to the predicted class to highlight in the image).

In a practical unsupervised context, these conclusions are generalized as follows:

- **Instances with high confidence scores**, and thus coherent Grad-CAM and Confident-CAM maps, are associated with trustworthy class predictions and explanations.
- **Instances with lower confidence scores**, and thus nearly null Confident-CAM maps, are associated with unreliable class predictions and/or explanations. The radiologists should pay special attention to these cases.

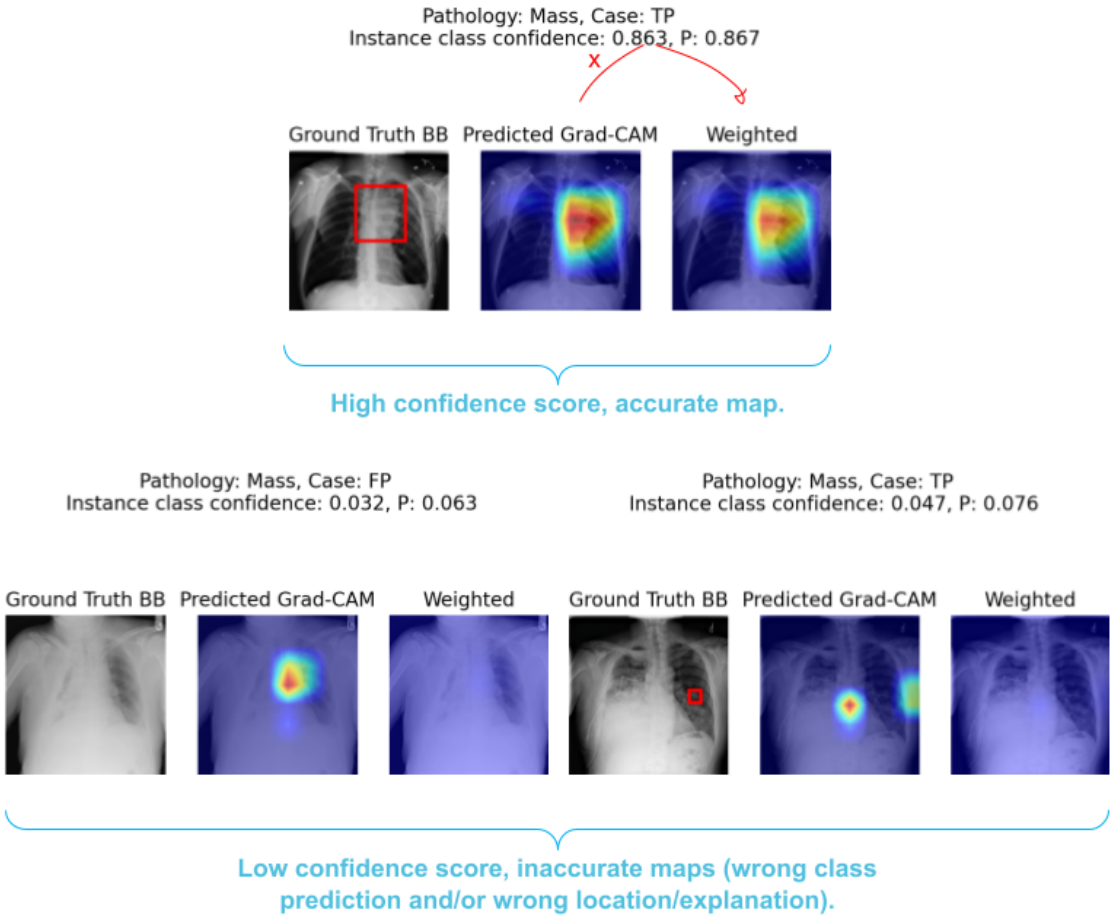



Figure 5.5: Schematic of the three possible Grad-CAM scenarios and corresponding "rectified" maps, named Confident-CAMs. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability.

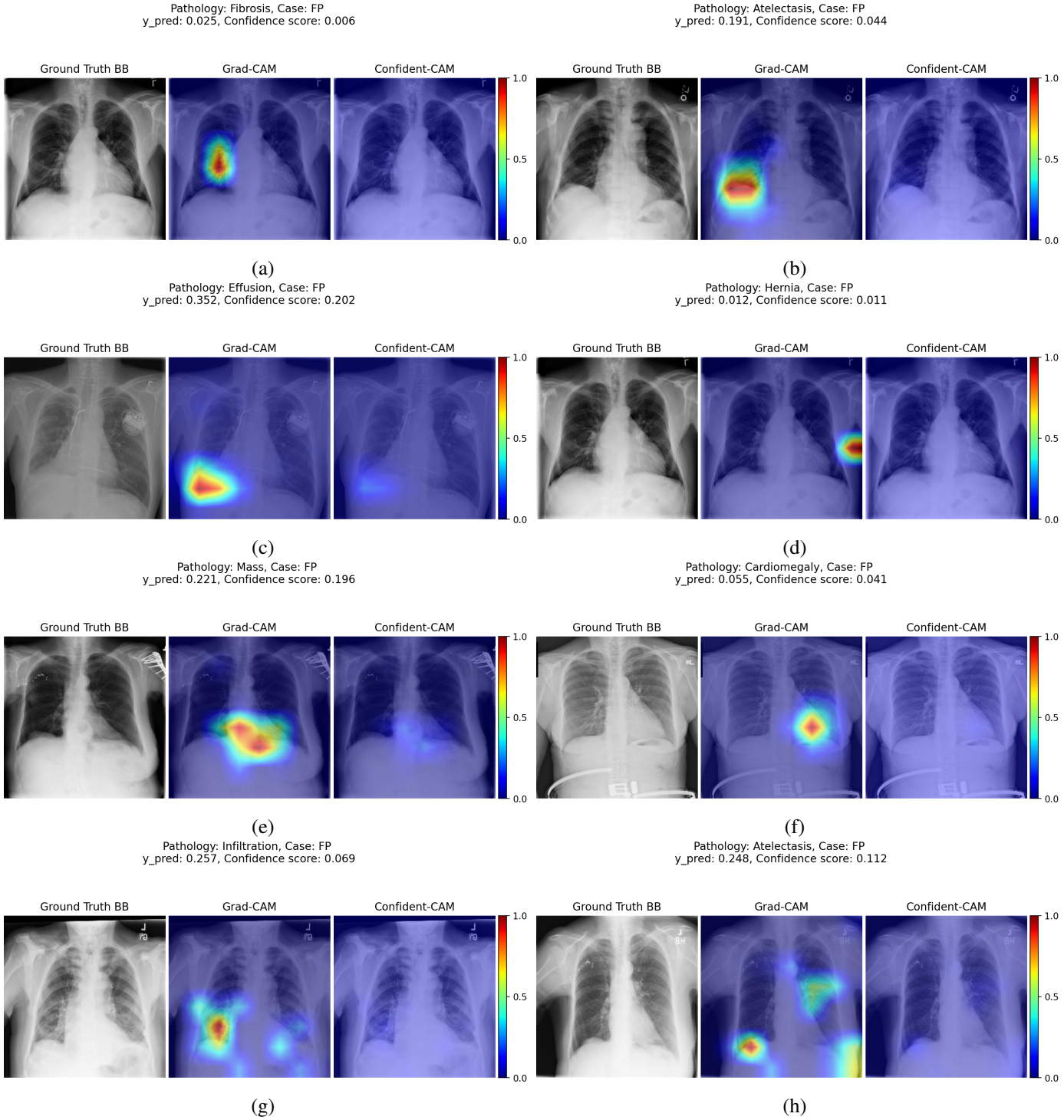



Figure 5.6: Confident-CAM results for several FP predictions with unsatisfactory Grad-CAMs. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability.

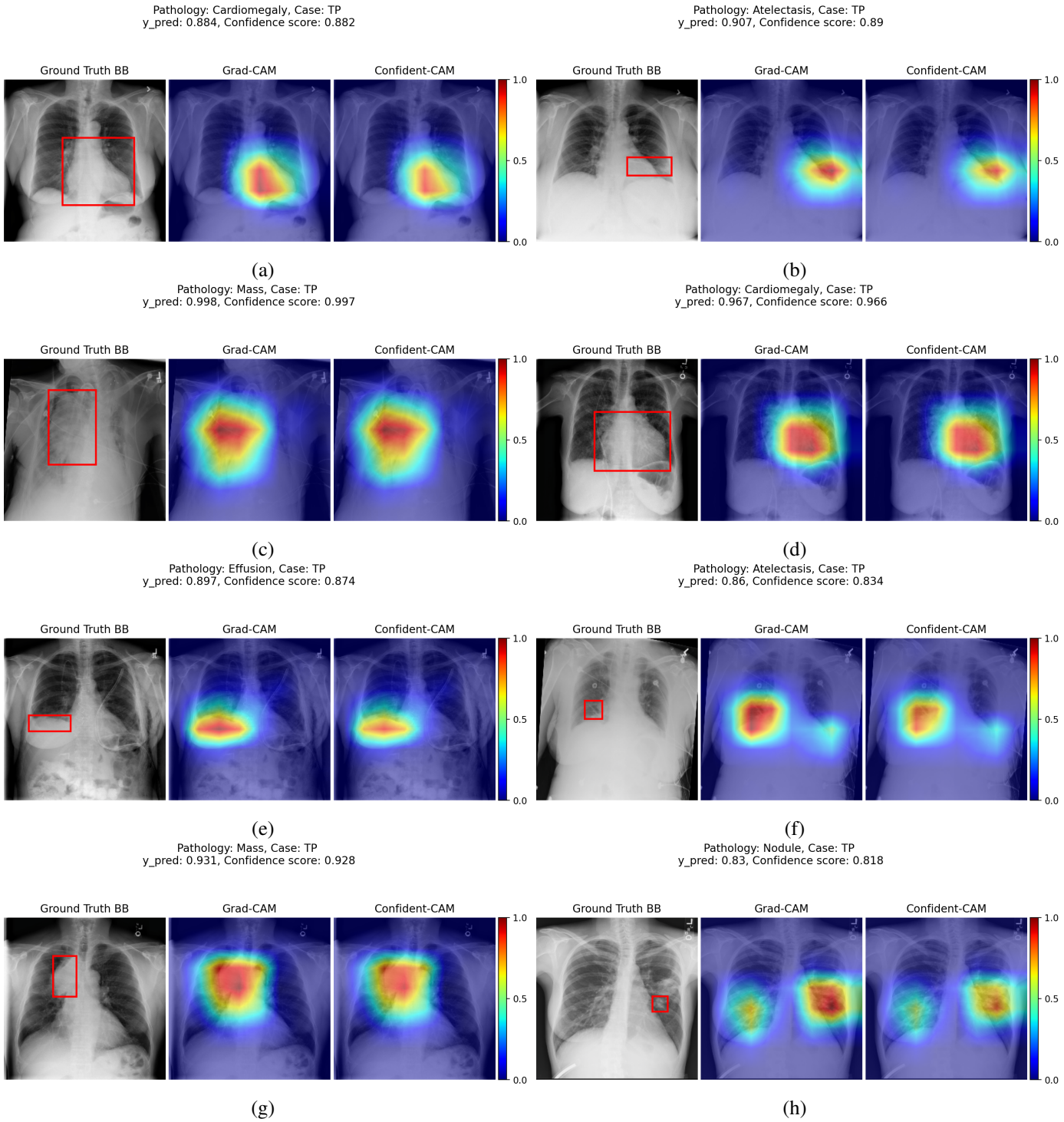



Figure 5.7: Confident-CAM results for several TP predictions with satisfactory Grad-CAMs. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability.

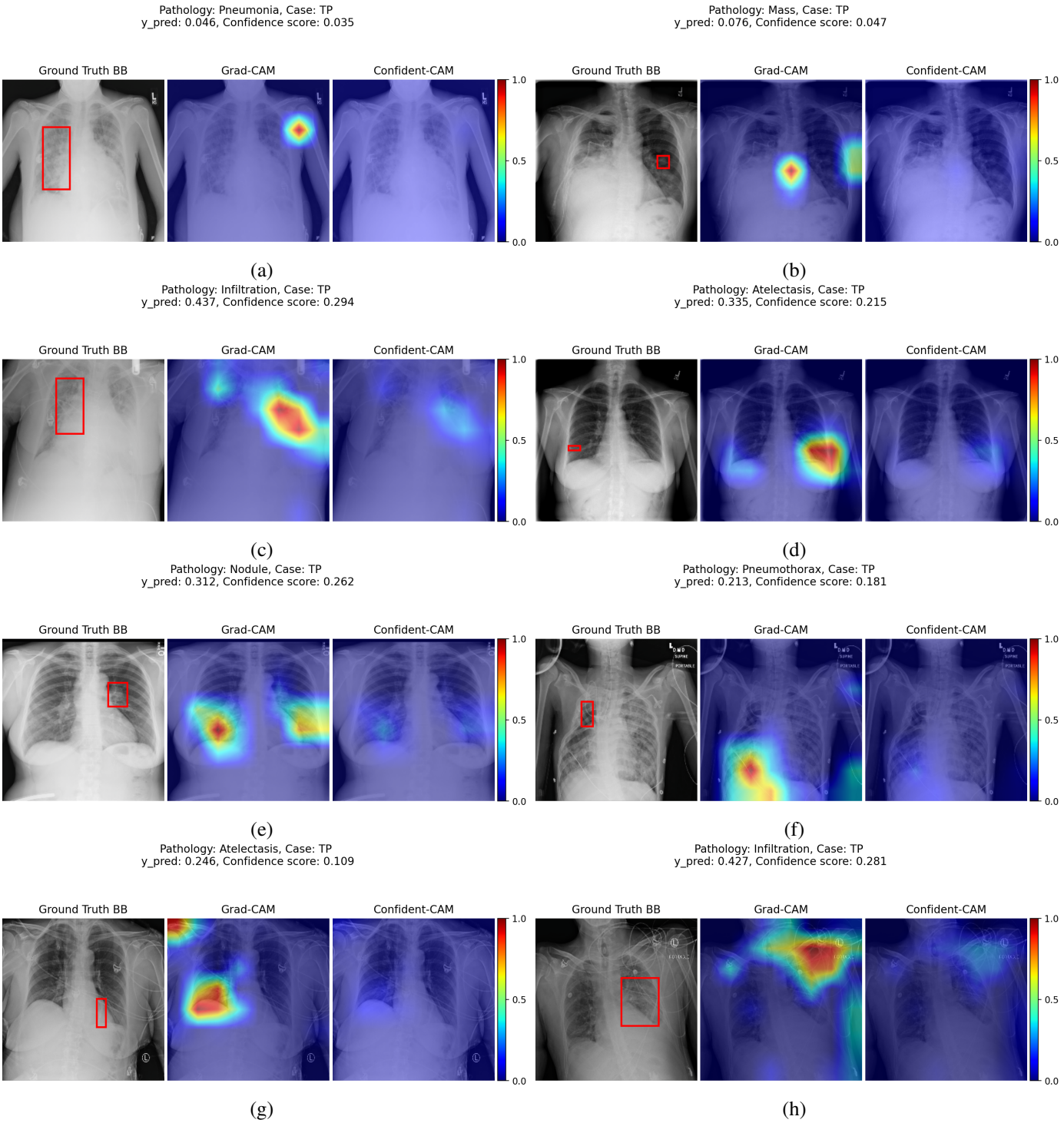
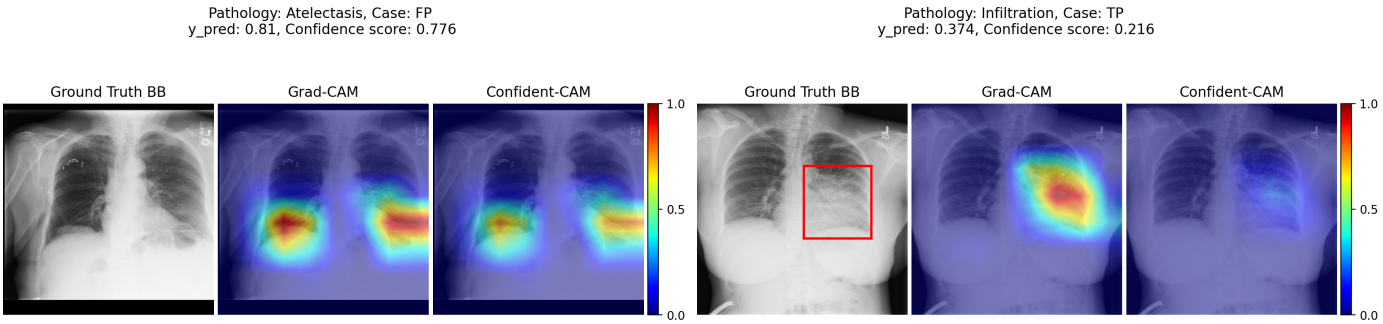



Figure 5.8: Confident-CAM results for several TP predictions with unsatisfactory Grad-CAMs. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability.




(a) (b)

Figure 5.9: Inadequate Confident-CAM results for a FP and a TP prediction. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability.

### 5.4.1 Method Validation

As previously discussed, the results in Table 5.1 show that less represented pathologies are more likely to have lower confidence scores, which may indicate that the proposed score can be informative regarding the sureness of the model's predicted class and corresponding XAI map. However, to further validate the three scenarios established earlier in this section, one must assess the confidence score values for different subsets of instances. In that sense, the test set was split into TP and FP cases, and the mean confidence score was calculated for both of these subsets. Table 5.3 reinforces that FP instances have typically lower confidence scores and will consequently result in low-activated Confident-CAMs.

Additionally, it is also important to verify that the method distinguishes between the correct and incorrect explanations of TP cases using their confidence scores. This specific validation was performed on an independent test set containing 848 out of the available 984 BBs, corresponding to the images predicted as positive by the model. For this, let us assume that each BB contains $N$ pixels, and that each Grad-CAM is binarized by selecting the $N$ highest activated pixels, as shown in Figures 5.10-5.11. The mean SA-IOU, defined in Section 3.5, was calculated for each class in Table 5.4, to measure the overlap between the segmented region and the GT.

The independent test set was split into low-overlap and high-overlap instances, using the mean SA-IOU per class as an adaptive threshold. After that, the corresponding confidence scores of each subset were calculated. Table 5.4 indicates that the TP cases with low overlap between the explanation and the spatial annotation exhibit nearly half of the mean confidence score of the TP cases with high overlap. This may indicate that low confidence scores are mostly obtained for poor explanations (i.e. low overlap), enabling the distinction of inaccurate diagnoses made by the model. This pattern is clear in Figure 5.10.

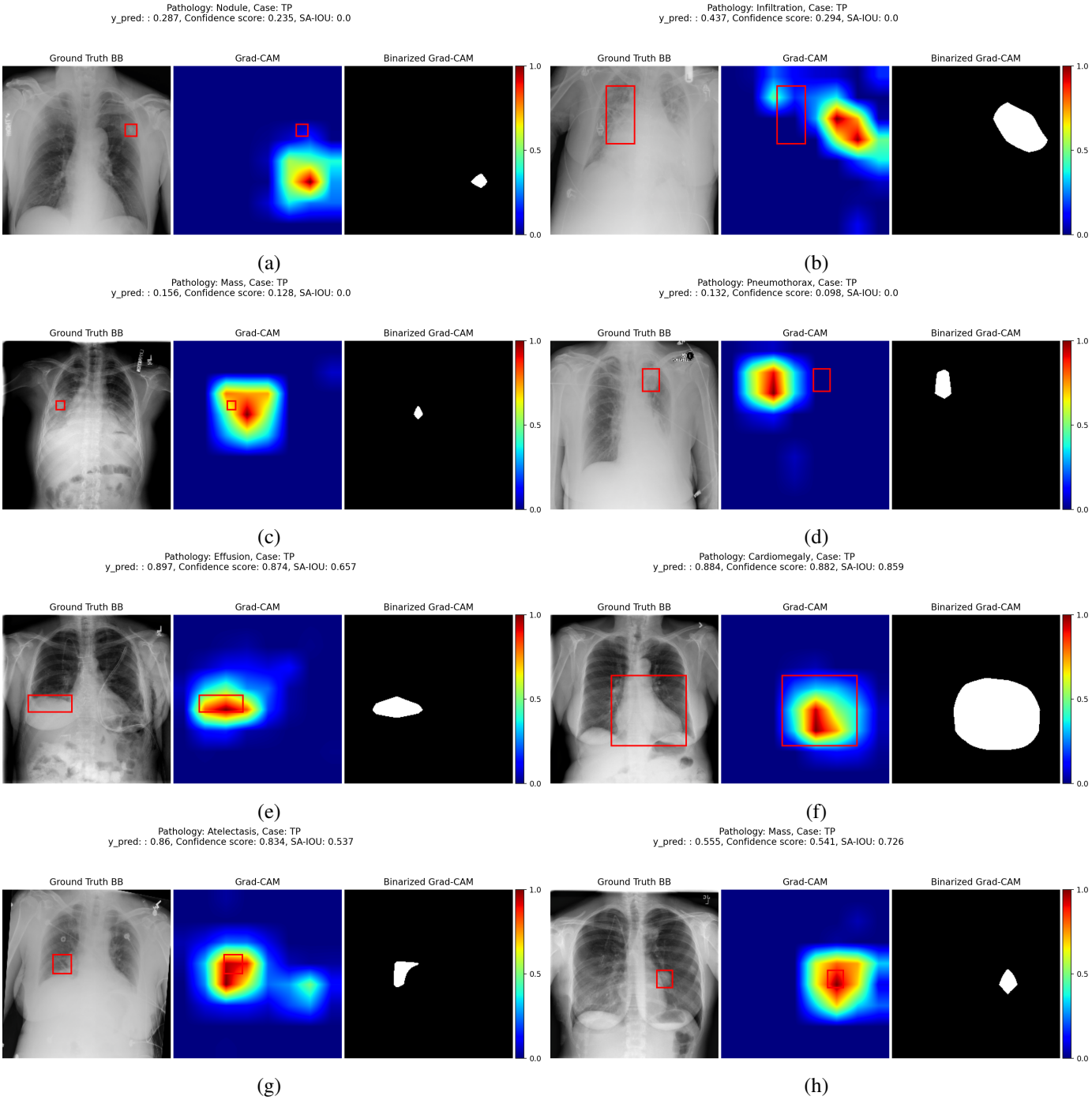



Figure 5.10: Example of TP instances with low and high confidence scores, (a-d) and (e-h) respectively, and their corresponding Grad-CAMs. The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability.

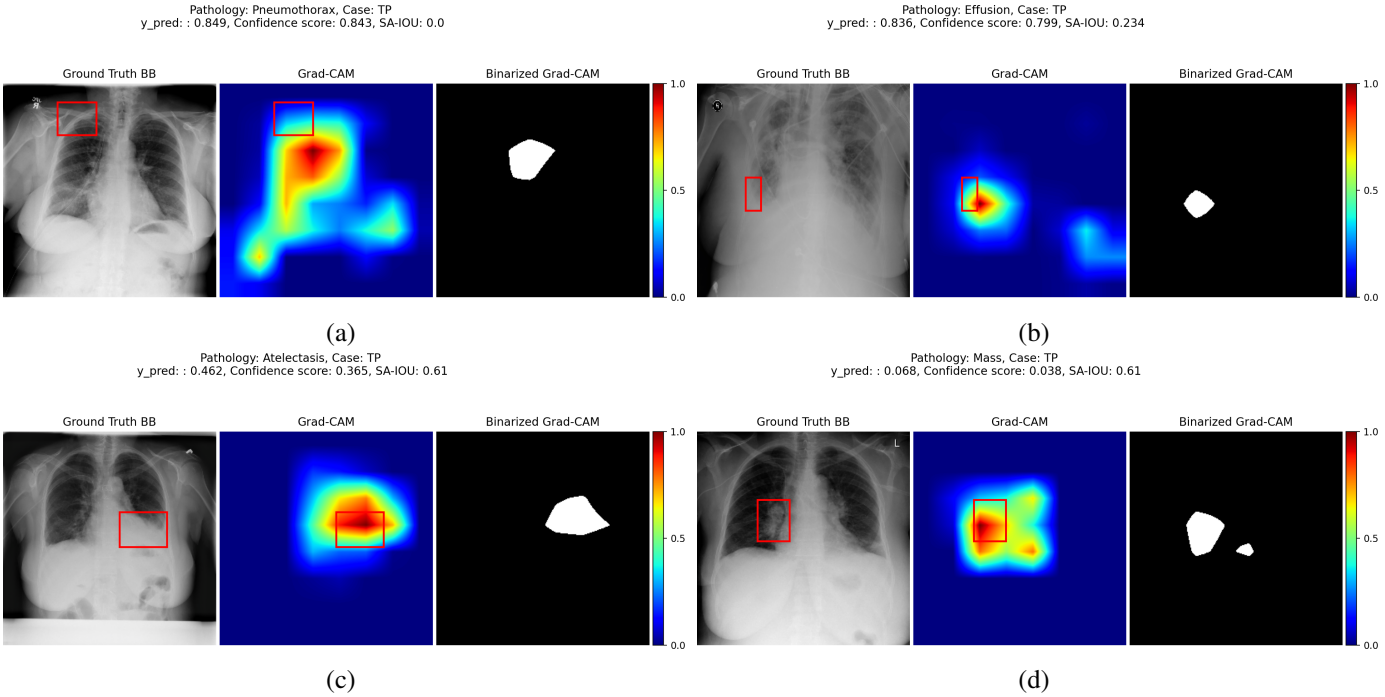



Figure 5.11: Inadequate results for TP instances. High confidence scores with low overlap in (a-b), and vice versa in (c-d). The red BB marks the GT location of the lesion, and *y_pred* is the predicted probability.

At last, it is essential to address the scenarios where the proposed algorithm produces suboptimal results. For instance, the Grad-CAM in Figure 5.9(a) is incorrect, and the confidence score supports the inadequate decision, contrary to what is expected. The Grad-CAM in Figure 5.9(b), on the other hand, highlights the correct anatomical region for the infiltration finding, and yet the confidence score is not high enough to validate that result. These are examples of outliers in this dataset that contribute to the STD values in Table 5.3 and Table 5.4. In other words, although TPs generally have a higher mean confidence score than FPs, there are outliers among the TP instances with notably lower scores. Similarly, some FPs are outliers with confidence scores higher than the mean of their subset. This issue is further accentuated by the very limited number of samples in this test set, which makes the results more susceptible to the influence of outliers.

Figure 5.11 complements this outlier analysis by displaying TP instances with high confidence score and low overlap, or vice versa. In Figure 5.11(a), the model concentrated on the upper right pulmonary region, but focused mostly on the external cord crossing the right lung, misinterpreting it as a pneumothorax finding. The map in Figure 5.11(b), while close to the correct spatial annotation and reasonably acceptable, yields a low SA-IOU score. This is somewhat expected, as Grad-CAMs and other heat map explanations are inherently less precise than BBs, given that the model was not trained using spatial information. Figures 5.11(c-d) present cases where the explanations were also reasonable; however, the confidence scores did not support these interpretations.

Given that a high confidence score is generally associated with a high SA-IOU, heuristic limits were established for both metrics to enable an additional discussion of this topic. This approach

helps quantify the instances where the method falls short. Here, a Confident-CAM is considered to be unsatisfactory based on the following rules: an instance with a high confidence score (greater than 0.7) and low SA-IOU (lower than 0.3), or an instance with a low confidence score (lower than 0.3) and high SA-IOU (greater than 0.7). Only 12.7% of the TP test set matches these conditions, meaning that in most cases the confidence score translates to an appropriate SA-IOU value, and the method succeeds.

Table 5.3: Mean confidence score for instances predicted as positive during inference.

| | **Confidence Score** | | |
|---|---|---|---|
| | **All** | **TPs** | **FPs** |
| **Mean** | 0.209 $\pm$ 0.232 | 0.399 $\pm$ 0.276 | 0.153 $\pm$ 0.142 |

Table 5.4: Confidence score per class for the TP instances of the independent test set, containing a total of 848 BBs (TPs). All the available BBs of the ChestX-ray14 dataset are described in the last column. The final row provides the mean and STD values across these test samples.

| **Pathology Class** | **Mean SA-IOU** | | **Confidence Score** | | | **#TP BBs** | **#BBs** |
|---|---|---|---|---|---|---|---|
| | | | **SA-IOU** $< \overline{SA-IOU}$ | **SA-IOU** $\geq \overline{SA-IOU}$ | | | |
| **Atelectasis** | 0.265 | | 0.351 | 0.542 | | 156 | 180 |
| **Cardiomegaly** | 0.796 | | 0.298 | 0.672 | | 135 | 146 |
| **Effusion** | 0.359 | | 0.481 | 0.735 | | 143 | 153 |
| **Infiltration** | 0.218 | | 0.288 | 0.677 | | 114 | 123 |
| **Mass** | 0.309 | | 0.372 | 0.663 | | 69 | 85 |
| **Nodule** | 0.117 | | 0.433 | 0.632 | | 54 | 79 |
| **Pneumonia** | 0.165 | | 0.022 | 0.137 | | 99 | 120 |
| **Pneumothorax** | 0.176 | | 0.459 | 0.698 | | 78 | 98 |
| **Mean SA-IOU** | **0.334** $\pm$ 0.311 | **Mean C. Score** | **0.335** $\pm$ 0.282 | **0.596** $\pm$ 0.294 | **Total** | 848 | 984 |

## 5.5 Concluding Remarks

This study seeks to improve the interpretation of normalized heat map explanation methods, which are not comparable to each other when it comes to instance-wise prediction confidence. In other words, all maps are plotted in the same 0-1 color scale, regardless of the distance between the predicted probability and the optimal class threshold. As probability by itself cannot provide a metric for the model's degree of confidence, the proposed method suggests the use of a confidence score per instance, per class to measure that distance and evaluate how certain the model is regarding a prediction. The original heat map explanation is then multiplied by this score to get the Confident-CAM.

This method was applied to a multi-label model for the classification of 14 CXR findings and built on top of Grad-CAM explanations, even though it can be used on any other technique whose maps are inherently normalized. The results indicate that instances with a high confidence score

are associated with more accurate maps (concordant Grad-CAMs and Confident-CAMs), while instances with lower confidence scores are associated with inaccurate explanations. There are two possible reasons for this: either the model has predicted the wrong class to begin with, and in this case, there are no lesions to be highlighted, or the model was right about the prediction but highlighted the wrong anatomical region - in both cases the Confident-CAM will not be concordant and will be nearly null.

The confidence scores per class were calculated to validate these conclusions, showing that less represented findings typically yield less certain predictions, as the model is more "unsure" classifying instances that appear less on the training set. TP instances also exhibit a statistically higher confidence score in comparison to FP ones, allowing the distinction between possibly correct and incorrect class predictions, respectively. Finally, for instances predicted as positive, it was verified that higher confidence scores are achieved when a good explanation is provided (i.e. high overlap between the highlighted Grad-CAM regions and the lesions' GT). The opposite was concluded for lower confidence scores, associated with more unsatisfactory maps, and thus discriminating between accurate and inaccurate explanations for correctly predicted classes.

Any high-risk application with significant class imbalance can benefit from the implementation of Confident-CAMs, as they will provide an extra validation step that the users can easily interpret to know if they should trust the initial explanation.

# Chapter 6

# CLARE-XR: Explainable Regression-based Classification of Chest Radiographs with Label Embeddings

**The content of this chapter results from the following publication. The author of this thesis has the main contribution to the conception of the study, method development, and results assessment.**
**Rocha, J.**, Pereira, S. C., Sousa, P., Campilho, A., and Mendonça, A. M. (2024). CLARE-XR: Explainable Regression-based Classification of Chest Radiographs with Label Embeddings. Accepted in Scientific Reports.

The previous chapter focused on improving the interpretation of existing XAI maps by building on the current state-of-the-art approach. In contrast, this chapter identifies additional limitations which motivate a novel alternative framework. In detail, CLARE-XR is a novel methodology that, when presented with an X-ray image, identifies the associated pathologies and provides explanations based on the presentation of similar cases. The diagnosis is achieved using a regression model that maps an image into a 2D latent space containing the reference coordinates of all findings. The references are generated once through label embedding, before the regression step, by converting the original binary GT annotations to 2D coordinates. The classification is inferred minding the distance from the coordinates of an inference image to the reference coordinates. Furthermore, as the regressor is trained on a known set of images, the distance from the coordinates of an inference image to the coordinates of the training set images also allows retrieving similar instances, mimicking the common clinical practice of comparing radiographs to confirm diagnoses. This inherently interpretable framework discloses specific classification rules and visual explanations through automatic image retrieval methods, outperforming the multi-label ResNet50 classification baseline across multiple evaluation settings on the NIH ChestX-ray14 dataset.

## 6.1 Introduction

Several studies have addressed the multi-label classification of CXRs, highlighting the challenges posed by the increased number of labels and their interdependencies (Yeh and Li, 2020). As the number of classes grows, their distinction becomes even more intricate, especially when acquiring sufficient training images for rare instances is not possible. State-of-the-art DL approaches are often model-centric, seeking to maximize the classification performance of off-the-shelf architectures on a given dataset. They typically embed images into high-dimensional features to extract relevant information, paired with their binary-encoded GT. Sun et al. (2017) argues that using binary vectors to designate the presence or absence of classes complicates the measurement of label correlation and encourages overfitting due to the extreme nature of the binary representation. Furthermore, the high non-linearity of DL models results in opaque internal decision processes, making them less suitable for the clinical setting. This opacity prevents healthcare professionals from comprehending the reasoning behind specific diagnoses, discouraging the adoption of DL models as a second opinion, while the current XAI methods remain insufficient for high-risk practical applications (Ali et al., 2023).

The objective of this work is to provide a more data-centric approach to the problem, aiming to enhance the model's learning capability and explainability by refining and reshaping the available dataset. In addition to image embedding, this study employs label embedding to encode the GT into a latent space, transforming initial binary vectors into a lower-dimensional Euclidean space. This enables the transition from a standard multi-label classification to a regression problem - a more transparent process where the model predicts continuous coordinates to further elucidate how and why decisions are made by positioning each image's prediction in the reference space. This method provides physicians with instance-specific textual and visual justifications based on both image and label features learned by the model. The contributions of the current chapter are the following:

1. Learning of a two-dimensional (2D) reference space capable of representing the image labels via embeddings.
2. A data-centric regression-based approach for multi-label classification of CXR images.
3. An inherently interpretable set of rules to reach that multi-label classification.
4. A visual explainability approach that exhibits similar cases, based on the image and GT labels' positioning in the reference space.

The next section presents a review of the state-of-the-art publications on the topics of interest. A detailed overview of the dataset and framework is presented in Section 6.3. The results are further discussed in Section 6.4, and the main conclusions are drawn in Section 6.5.

## 6.2 Related Work

In the field of computer vision, there is an extensive body of literature on image embedding (how to represent an image), but significantly less work has focused on label embedding (how to represent a class). Publications on label embedding for image classification are limited, using primarily natural image datasets paired with textual descriptions. Early work on this topic redefined original class labels into vectors, and then created two separate spaces: one for the image features and another for the label attributes (Akata et al., 2016). By embedding each image in the feature space and each class in the attribute space, the "compatibility" between these embeddings can be estimated to make predictions, given that correct classes will ideally rank higher in compatibility than incorrect ones. The work in Yeh and Li (2020) built upon this idea by ranking labels according to their relevance to the input image, in order to position semantically similar labels closer together in the continuous embedding space. Unlike the previous approach that compares two embedding spaces, this model transforms an image into the same space, positioning it near its associated labels. Chen et al. (2019b) follows the same logic, jointly projecting features and labels into a latent space to model their interactions and learn label correlations. Alternatively, one may also leverage label semantic information in spatial attention mechanisms that identify relevant image regions (Huang et al., 2022; Sun et al., 2023a), allowing the model to focus on image feature regions with more pronounced label interdependencies. Some of the previous ideas have been combined into a novel approach (Li et al., 2023). Images and labels are fed into individual encoders, resulting in each image being described with a set of patch embeddings and a set of label embeddings. The method learns to associate image regions with labels by modeling their similarities, offering a tool to visualize the learned label concepts. The novelty of this publication lies in its interpretable outcome, for understanding the interaction between labels and patches. This opened the door for XAI-driven approaches using label embeddings, such as the one in this chapter. Prior to this, only one paper mentioned interpretability as an advantage of using label embeddings to identify meaningful similarities between classes (Sun et al., 2017).

A more recent trend started using label embeddings for multi-label CXR classification, mostly as a way of learning pathology co-occurrence. The works in Chen et al. (2020) and Zhang et al. (2024) explore the dependencies between findings to enhance classification performance. The image feature learning process is guided by the label correlation information derived from the label embeddings. Specifically, the relationship between pathologies is translated into a set of classifier scores that weight the extracted image features to produce the final prediction. The concept of using label embeddings for self-attention operations has also been extended to more complex classification architectures, ViTs, due to their promising results in Wu et al. (2024) and Jiang et al. (2024). Such studies introduce learnable label embeddings to detect and match class-specific image features.

To the best of the author's knowledge, there is only one publication addressing both label embeddings and XAI in CXR classification, though not in conjunction (Sun et al., 2023b). There, the proposed explanations are merely heatmaps highlighting the most relevant image regions for

the predicted class, which is by far the most common XAI approach in this field as seen in the earlier chapters (Yan et al., 2018; Urinbayev et al., 2020; Zhang et al., 2020; Guan and Huang, 2020; Wang et al., 2021).

Chapter 5 addressed various limitations of heat maps, and Grad-CAMs in particular, such as the difficulty in evaluation when no spatial GT is available, leading to questionable XAI performance. One may also add that they are highly sensitive to both model architecture and quality, which can result in a lack of causal accuracy when highlighting certain regions and the omission of key features critical to the prediction (Smilkov et al., 2017). An innovative alternative mimics clinical practice, where radiologists look for cases with similar findings to support and explain their decisions (Silva et al., 2020; Pedrosa et al., 2022). In this approach, image retrieval automatically finds similar cases within the dataset, allowing the end-user to compare these cases to validate the diagnosis. The present chapter integrates label embedding and image retrieval concepts to build a more inherently interpretable framework. The objective is to avoid complex architectures while maintaining classification performance, which is accomplished by favoring lower-dimensional label embeddings, as opposed to the state-of-the-art studies that increase label dimensions. The proposed reference space, representing both the GT label embeddings and the predictions derived from image features, facilitates the understanding of the final decision-making process, as well as the retrieval of similar images to further justify the model's decisions.

## 6.3 Method

### 6.3.1 Dataset Preprocessing

As in Chapter 5, this study also used the ChestX-ray14 dataset. Figure 5.4(a) showed its total number of instances per pathology, which are significantly correlated according to Figure 5.4(b). Additionally, Figure 6.1 presents the histogram of co-occurring pathologies, revealing that the images contain up to nine simultaneous pathologies, further evidencing the high imbalance of the annotations.

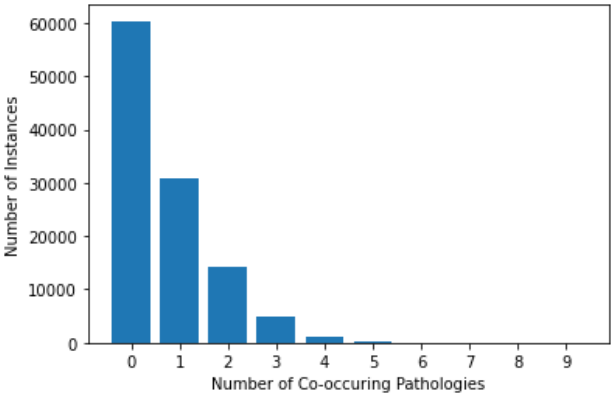



Figure 6.1: Frequency of 0, 1, or more co-occurring pathologies in the original distribution of the ChestX-ray14 dataset.

Regarding preprocessing steps, this study considers a subset of the ChestX-ray14 data consisting of the 110,486 instances with up to 3 co-occurring pathologies, which is the most common scenario. In other words, all images annotated with 4 or more co-occurring findings were discarded due to severe under-representation. Following the same logic, and minding the reduced number of samples for the minority findings (e.g. hernia and pneumonia), the initial 14 pathologies were grouped into the 5 broader categories listed below, inspired by previous work (Irvin et al., 2019). Though pneumothorax is technically classified as a pleural pathology, its radiographic appearance is distinct from other pleural conditions like effusion and pleural thickening. Pneumothorax involves air accumulation in the pleural space, while pleural effusion is the presence of fluid, and pleural thickening indicates scarring of the pleura. These differences lead to contrasting imaging characteristics: pneumothorax appears darker than lung tissue due to air density, whereas pleural effusion and thickening appear lighter or more opaque. Consequently, pneumothorax is categorized separately from other pleural pathologies in this work. The five categories are the following:

- **Cardiac pathologies (*Ca*)**: cardiomegaly.
- **Parenchymal lesions (*Pa*)**: infiltration, fibrosis, atelectasis, mass, nodule, pneumonia, consolidation, and/or edema.
- **Pneumothorax (*Pn*)**: pneumothorax.
- **Pleural lesions (*Pl*)**: effusion, and/or pleural thickening.
- **Other lesions (*Ot*)**: hernia, and/or emphysema.

From here on, these 5 classes were considered in alternative to the 14 pathologies, meaning each original GT annotation was defined as a 5-class vector [*Ca*, *Pa*, *Pn*, *Pl*, *Ot*], where every element is assigned the value 1/0 if the corresponding pathology is present/absent. Table 6.1 shows how many instances per class are annotated as single-pathology or multi-label (i.e. have one, two, or three types of pathologies present). For example, the *Pa* class contains mostly single-pathology cases, the *Pl* class contains mostly multi-label cases with a combination of two types of pathology, and the *Ca* class is quite balanced between the two possibilities. Table 6.2 describes the resulting 26 annotation combinations. The available data was split into five folds, without patient overlap and preserving the original pathology proportions.

Table 6.1: Number of instances per class, annotated with one, two, or three types of pathology (C1, C2, and C3, respectively), in the selected ChestX-ray14 subset.

| Class | C1 Counts | C1 % | C2 Counts | C2 % | C3 Counts | C3 % |
|---|---|---|---|---|---|---|
| **Ca** | 1,093 | 43.8 | 1,089 | 43.6 | 313 | 12.5 |
| **Pa** | 28,440 | 73.2 | 9,476 | 24.4 | 930 | 2.4 |
| **Pn** | 2,194 | 44.9 | 2,089 | 42.7 | 604 | 12.4 |
| **Pl** | 5,332 | 36.8 | 8,360 | 57.7 | 796 | 5.5 |
| **Ot** | 1,003 | 40.5 | 1,100 | 44.4 | 375 | 15.1 |

Table 6.2: Number of instances per combination in the selected ChestX-ray14 subset.

| Combination | Counts | Combination | Counts | Combination | Counts | Combination | Counts |
|---|---|---|---|---|---|---|---|
| **No Finding** | 60,361 | **Ot** | 1,003 | **CaPaPl** | 289 | **CaPn** | 9 |
| **Pa** | 28,440 | **PaOt** | 576 | **PaPnOt** | 200 | **CaPaPn** | 6 |
| **PaPl** | 7,148 | **CaPl** | 532 | **PlOt** | 164 | **CaPnPl** | 3 |
| **Pl** | 5,332 | **CaPa** | 527 | **PaPlOt** | 100 | **CaPnOt** | 2 |
| **Pn** | 2,194 | **PnPl** | 516 | **PnPlOt** | 60 | **CaPaOt** | 2 |
| **PaPn** | 1,225 | **PnOt** | 339 | **CaOt** | 21 | | |
| **Ca** | 1,093 | **PaPnPl** | 333 | **CaPlOt** | 11 | | |

### 6.3.2 Framework

The CLARE-XR methodology intends to *clarify* CXR classification by providing explanations based on the presentation of similar cases along with the final decision. Each input image is passed to a regression model that predicts its corresponding $[x, y]$ coordinates, positioning the radiograph within a latent space that contains the reference coordinates for all labels. The classification of each radiograph is determined by its relative position to these reference points, while the explanation strategy considers its relative position to a known set of training images in the same space to retrieve matching instances.

The framework, outlined in Figure 6.2, organizes the method into three phases: constructing the latent space, positioning the images within this space, and classifying them while generating the corresponding explanations.

The latent space and the regressor model are obtained in Phase I and Phase II of the workflow, which correspond to the training phases. In Phase I, an autoencoder is trained and employed once, before the regression phase, to generate label embeddings that encapsulate the same information as the original GT vectors for the 5 classes (hereon, 5D GT). By transforming 5D GT vectors into 2D vectors, the autoencoder creates a latent space that contains the reference coordinates for all annotations. Note that this phase does not consider any images yet, and looks exclusively at the labels to produce their corresponding new coordinates. The regressor is trained in Phase II, where each image is associated with its new 2D GT resulting from the encoder in Phase I. This way, the regressor learns to map each image into a 2D coordinate in the same space as the references.

Phase III corresponds to the inference stage, which employs the trained regressor to predict the coordinate of an inference image and then attributes specific class(es) to that image by comparing and associating its coordinates to the label reference coordinates. This step combines distance measures and a set of classification rules to reach a final decision, which allow a deeper understanding of why an image is assigned to a certain class. The final decision and its positioning in the latent space are also crucial for the explainability component of the system to retrieve similar cases from a known set of training images.

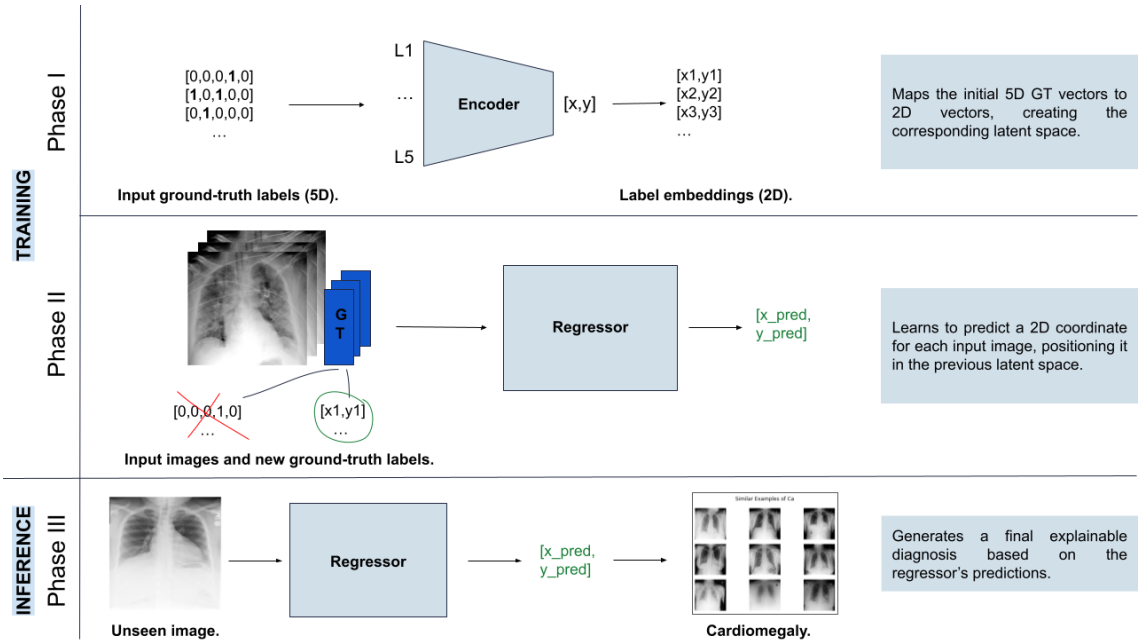



Figure 6.2: Framework of the CLARE-XR methodology.

#### 6.3.2.1 Phase I - Label Autoencoder

The autoencoder phase seeks to create a new way of representing the 5D GT labels. This way, the GT vector of 5 binary elements will be represented as a vector with only 2 real-valued elements, and that will serve as a reference for the following phases. Firstly, a dummy training dataset is generated, containing all possible binary combinations of the 5 classes in Table 6.2, except the *No Finding* annotation. As mentioned previously, only instances with 1, 2, or 3 co-occurring pathologies were considered to exclude rare cases and reduce the complexity of the approach, yielding a total of 25 annotation combinations. These 25 combinations were replicated 1000 times, resulting in a balanced training set of 25,000 instances. The validation and test sets used in this phase are composed of a single fold of the ChestX-ray14 data, keeping the original distribution of *No Finding*, single-class, and multi-label annotations.

This data was fed to an autoencoder architecture - a multi-layer perceptron whose encoder comprises 5 linear layers (each followed by ReLU activation), similar to the decoder. The encoder takes each vector with 5 positions ($L$) and encodes it to a latent size of 2, while the decoder reconstructs that representation back to 5 positions ($\hat{L}$). This process is depicted in Figure 6.3. Ideally, a proficient autoencoder is achieved if the reconstruction loss is nearly null, meaning that the $N$ input and output vectors are equal, and the model is capable of delivering accurate 2D encoded representations of the input vector. The Mean Squared Error (MSE) loss function in Equation 6.1 was used to minimize the difference between the input and reconstructed label data (i.e. $L$ and $\hat{L}$). By the end of this phase, all possible labels are now translated by the encoder to a unique 2D vector, representing the present and absent pathologies, and used as a GT reference for the following regression phase.

$$\text{MSE Loss} = \frac{1}{N}\sum_{i=1}^{N}(L_i - \hat{L}_i)^2 \tag{6.1}$$

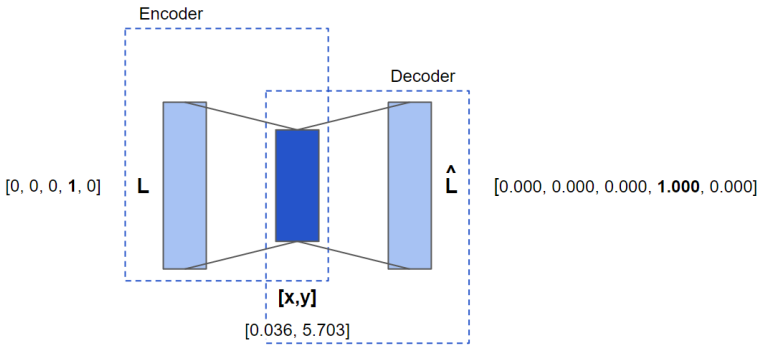



Figure 6.3: Schematic of the proposed autoencoder, exemplified for an input vector $L$ representative of a pleural lesion. The encoder yields the embedding $[x,y]$ and the decoder the reconstructed vector $\hat{L}$.

The autoencoder used in this phase constitutes a non-linear approach for dimensionality reduction. For further evaluation and comparison of this approach, an alternative linear method, Principal Component Analysis (PCA), was also explored for transforming the GT vectors into 2D. Although the primary focus of this work is on the autoencoder, the results obtained from the PCA approach are discussed in detail in Section 6.4.5.

#### 6.3.2.2 Phase II - Embedding-based Regression

The regression phase pairs each available image with its new 2D GT representation resulting from the Phase I encoder. Note that those label embeddings are composed of [$x$,$y$] coordinates, which were min-max normalized to fit in a 0-1 scale. This aspect allows for a more efficient optimization routine of the Phase II regressor.

A ResNet50 architecture was employed here to predict a pair of coordinates $[x_{pred}, y_{pred}]$ per image, modifying the last fully-connected layer to consider these 2 real-valued output nodes. The model was pre-trained on ImageNet weights, so the input images were normalized initially to a 0-1 range and then standardized using ImageNet's mean and STD values. All available radiographs were resized to 256×256 pixels to accelerate the training routine. The training data was also augmented to double the samples through soft random affine transformations (i.e. 5 degrees of rotation and 3 degrees of shear). Finally, a five-fold cross-validation scheme was implemented in all experiments, in which three folds were used for training, one for validation, and one for testing. Note that each label combination has a specific reference embedding coordinate $E$, and so the goal of Phase II is to obtain a predicted embedding $\hat{E}$ as close to that reference as possible for all $N$ images. For example, an image labeled with cardiac and pleural lesions should be positioned in the reference embedding coordinates of the *CaPl* combination. The MSE loss function in Equation 6.2 translates that purpose.

$$\text{MSE Loss} = \frac{1}{N}\sum_{i=1}^{N}(E_i - \hat{E}_i)^2 \tag{6.2}$$

#### 6.3.2.3 Phase III - Multi-label Classification and Explainability

The third phase addresses both the multi-label embedding-based classification and the XAI component in the inference stage of the process. Using the trained regressor from Phase II, each image is placed in the reference space using its predicted coordinates. Phase III then decides which pathology classes are present in an image to reach a diagnosis, based on its location and neighboring reference coordinates. This is achieved by measuring the Euclidean distance of each predicted coordinate to all 26 references obtained in Phase I. An instance-wise list of those distances in ascending order arranges the references from closest to farthest. Several sets of classification rules were proposed and evaluated to reach the most accurate diagnosis, minding that list. These rules can be split into two categories: rules based on the closest reference (top 1) and rules based on the three closest references (top 3). The latter assess if the neighboring references add further valuable insights that potentiate the method's efficiency. The proposed rules, whose pseudocode can be found in Appendix B, are as follows:

- **Rule 1** (top-1 approach): classifying an instance with the label(s) of its closest reference.
- **Rule 2** (top-3 approach): acts as a "control" experiment, assuming that all classes in all top-3 combinations are present in the image.
- **Rule 3** (top-3 approach): narrows down the results by considering only the recurrent classes (i.e. those that appear in at least 2 of the top-3 combinations).
- **Rule 4** (top-3 approach): refines these results by classifying an instance with the label(s) of its closest reference, plus adding the recurrent ones. This attributes a higher importance to the top-1 result (i.e. the most probable outcome), without discarding the surrounding information.

  Note that in all top-3 experiments, the *No Finding* reference is not considered if it the second or third-closest reference, as it cannot be a recurrent pathology. In these cases, that reference was not considered and replaced by the fourth-closest reference. Furthermore, any instance whose top-1 combination is *No Finding* is automatically attributed that decision, given the ease of the model in identifying normal radiographs.

Phase III identifies which rule maximizes classification performance, but it should also obtain coherent visual explanations that support the classification outcome. To fulfill this goal, the proposed XAI strategy takes advantage of the training set to retrieve similar instances with specific findings. By comparing the inference radiograph with others whose GT is known, the physicians can validate if the predicted pathologies are present in the case at hand.

By the end of Phase II, the regressor has positioned the training set images in the latent space, clustering those that share similar features in the same neighborhood. During inference, when

presented with a new unseen image, CLARE-XR assigns the relevant class(es) to it. The XAI component then complements the decision by identifying the training set radiographs that most closely resemble the specific findings in the inference radiograph, given the predicted outcome. This approach retrieves the top-k images closest to the predicted coordinate, selecting the training set images that share the most features learned by the model. Consequently, these examples are expected to be similar to the inference radiograph, revealing the same anatomical cues and lesion types. For an easier understanding, this strategy will be further discussed with examples.

## 6.4 Results and Discussion

This section evaluates the method's ability to classify each radiograph in an explainable manner, discussing its performance based on multiple experimental setups.

### 6.4.1 Baseline Classifier

Before discussing the proposed methodology, it was crucial to establish a baseline that reflects the typical state-of-the-art approach (Kumar et al., 2018; Urinbayev et al., 2020). This baseline is a 5-class multi-label classifier, trained on the same images as the Phase II regressor and under identical conditions regarding data preprocessing/augmentation techniques, cross-validation scheme, etc., to ensure a fair comparison. The classifier is also composed of a ResNet50 pre-trained on ImageNet weights, whose last layer of the network is a fully-connected layer with 5 nodes, followed by sigmoid activation to get a probability prediction per class. The training routine implemented the Adam optimizer with an initial learning rate of $10^{-6}$, a batch size of 16 images, and a total of 15 epochs. A BCE loss function was used to approximate the predicted classes to the GT annotations.

The output probabilities were used to compute the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves, found in Figures B.1-B.2 in Appendix B. These curves served as tools for determining class-specific optimal thresholds under two distinct evaluation criteria. The first aimed to maximize the true positive rate while minimizing the false positive one, whereas the second focused on maximizing the F1-score. Based on the selected class-specific thresholds, the predictions were binarized to evaluate several metrics, including accuracy, precision, recall, and F1-score. The performance of this baseline classifier is disclosed in Table 6.3. Note that the 26 possible combinations of labels are not considered in this analysis; instead, the evaluation follows the common multi-label approach, treating the 5 classes independently. As such, the class-wise values are obtained minding all instances predicted as positive for a given class, regardless of whether they are single-pathology or multi-label combinations.

The results obtained based on the ROC curve indicate a low precision and high recall scenario, leading to low F1-scores. As seen before, this is typical in state-of-the-art classification approaches that favor FP results over FN ones (Hicks et al., 2022). Examining the class-specific values reveals that the F1-score tends to reflect the dataset's distribution. More prevalent classes (e.g. *Pa*, *Pl*) typically achieve higher F1-scores compared to underrepresented ones (e.g. *Ot*, *Ca*). In contrast,

Table 6.3: Classification results of the multi-label baseline classifier. Results obtained for all instances predicted as positive for a given class, regardless of whether they are *No Finding*, single-pathology, or multi-label. The precision, recall, and F1-score values were calculated using class-specific optimal thresholds, optimized based on the ROC curve (*B_ROC*) or the PR curve (*B_PR*).

| **Class** | **AUC** | **Accuracy** | | **Precision** | | **Recall** | | **F1-Score** | |
|---|---|---|---|---|---|---|---|---|---|
| | **B_ROC** | **B_ROC** | **B_PR** | **B_ROC** | **B_PR** | **B_ROC** | **B_PR** | **B_ROC** | **B_PR** |
| **Ca** | 0.879 | 0.803 | 0.971 | 0.084 | 0.348 | 0.785 | 0.347 | 0.152 | 0.348 |
| **Pa** | 0.703 | 0.652 | 0.677 | 0.505 | 0.540 | 0.647 | 0.540 | 0.567 | 0.540 |
| **Pn** | 0.827 | 0.807 | 0.938 | 0.147 | 0.297 | 0.694 | 0.297 | 0.242 | 0.297 |
| **Pl** | 0.832 | 0.777 | 0.862 | 0.340 | 0.472 | 0.743 | 0.471 | 0.466 | 0.471 |
| **Ot** | 0.794 | 0.736 | 0.963 | 0.057 | 0.166 | 0.696 | 0.166 | 0.105 | 0.166 |
| **Mean** | **0.807** | **0.755** | **0.882** | **0.226** | **0.365** | **0.713** | **0.364** | **0.307** | **0.364** |

results derived from the PR curve indicate a more balanced trade-off between precision and recall, deviating from that dependence on class distribution. Overall, the optimization using the PR curve improved accuracy by 12.7% and F1-score by 5.7% compared to the ROC curve optimization, primarily by increasing precision at the expense of recall.

The key distinction between this model and the state of the art that employs ChestX-ray14 lies in the reduced number of instances and class definitions. Specifically, this method uses fewer images and consolidates the original 14 classes from the ChestX-ray14 dataset into 5 broader categories. Despite this, this approach achieves a mean AUC of 0.807, which is comparable to the majority of published results on ChestX-ray14. Typically, mean AUC values for this dataset range from 0.795 to 0.855 (Table 5.2), depending on the specific method and evaluation setting. Additionally, the model developed in Chapter 5, employing the full dataset and all 14 classes, yields a mean AUC of 0.822. This suggests that the data preprocessing steps applied in this chapter do not significantly degrade the model's performance.

To the best of the author's knowledge, no previous publication reported precision and recall values for the ChestX-ray14 dataset. Therefore, a reference is made to a study on multi-label classification using the CheXpert dataset, which is the only one identified to report these metrics (Zhang et al., 2024). That study proposes an end-to-end framework utilizing self-attention to capture global label correlations, which are incorporated into a contrastive loss to learn more discriminative label-level features. As demonstrated in Table 6.4, both the referenced publication and the proposed baseline model exhibit a similar pattern, achieving higher recall than precision.

Table 6.4: Comparison with the results of a state-of-the-art multi-label classification method that discloses the precision, recall, and F1-score. This method (*) is trained and tested on CheXpert, while the proposed baseline and CLARE-XR methods are trained and tested on the ChestX-ray14 dataset. *N/A* stands for Not Available.

| Method | AUC | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Zhang et al. (2024)* | 0.765 | 0.381 | 0.559 | 0.446 |
| **Multi-label Baseline** | **0.807** | **0.226** | **0.713** | **0.307** |
| **CLARE-XR** | **N/A** | **0.355** | **0.543** | **0.390** |

### 6.4.2 Label Autoencoder

The autoencoder was trained using the Adam optimizer, with an initial learning rate of $10^{-3}$, a batch size of 32 vectors, and a total of 20 epochs. Multiple architectures and hyperparameters were initially tested, but ultimately the conditions described above provided the strongest encoder for this specific learning task. The dummy training set was used in this scenario due to the pathology correlation and imbalance verified in the original ChestX-ray14 distribution (Wang et al., 2017a). Using the original set could cause the autoencoder to learn when two or more pathologies co-occur more often and shortcut the learning process based on that information, leading to biased reference embeddings. CLARE-XR prevents that by creating a fully-balanced dummy dataset with all possible combinations with up to 3 pathologies. The number of zeroes surpasses the number of ones given the binary nature of the 5-element vectors, and for this reason, it was not necessary to include the *No Finding* null vector during training. The remaining combinations were enough to learn that specific reconstruction.

The best way to evaluate the quality of the encoder and its resulting 2D representations is to consider the decoder's outputs. In other words, the embeddings are most likely a good representation of the original label vectors if the reconstructed vectors are similar to the input ones. The autoencoder presents a very low mean MSE value ($4.867 \times 10^{-7}$), when using a test set with all possible annotation combinations (only instances with 0, 1, 2, or 3 pathologies are considered). This indicates that $L$ and $\hat{L}$ are similar. The previous metrics were also employed in this context to show the perfect reconstruction ability of the autoencoder, in terms of precision, recall, and F1-score (1.000). This was expected due to the limited number of combinations to reconstruct.

It is now fundamental to inspect the resulting encoder's representations and position them in their new latent space. Each of the 26 annotation combinations in Table 6.2 is handled independently by the encoder and assigned a distinct 2D coordinate shown in Figure 6.4. The 5 single-class and *No Finding* embeddings are marked with a colored dot, while the remaining 20 multi-label embeddings are marked with black crosses. This plot serves as a guide for the following phases. Note that a multi-label annotation is not guaranteed to be near its single-class references, and that this is not a requirement, since the autoencoder has no notion of what a pathology is or what are the common features across combinations (it merely encodes the vectors).

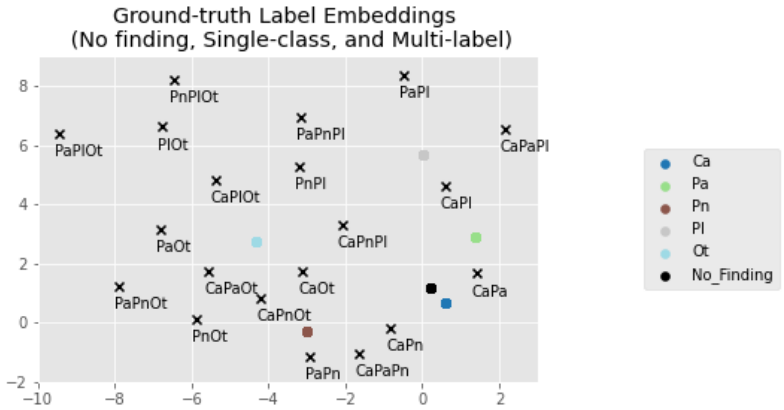



Figure 6.4: Reference label embeddings obtained by the Phase I encoder. The *No Finding* and single-class GT coordinates are plotted in colored dots, while the remaining multi-label combination coordinates are marked with black crosses.

### 6.4.3 Embedding-based Regression

The regressor was trained using the Adam optimizer, with an initial learning rate of $10^{-6}$, a batch size of 16 images, and a total of 10 epochs. Multiple architectures and hyperparameters were initially tested, but ultimately the conditions described above provided the strongest encoder for this specific learning task. The regression model is optimized to decrease the distance between the predicted embedding coordinates and the GT coordinates, being initialized with the baseline classifier's weights to be fine-tuned. It achieved an overall MSE of $2.015 \pm 0.014$, and while the combination with the lowest MSE matches the most represented combination (*No Finding*), the opposite cannot be concluded. On the other hand, note that a high MSE value is indicative of a poor prediction, but a low MSE does not guarantee a perfect prediction - e.g. a *Ca* instance can be plotted on top of the *No Finding* reference and still obtain a low MSE.

Given the trained model, one may visualize the predicted coordinates per combination on top of the reference space. Figure 6.5 exemplifies the scattering of the predicted embeddings for the *No Finding*, *PaOt*, *Ot* and *Pn* combinations. The remaining plots can be found in Figures B.3-B.5 in Appendix B. In agreement with the MSE analysis, the majority of the 60 thousand *No Finding* instances form a dense cloud near the correct reference coordinate, meaning that the model identifies such cases with ease. Note that the predictions do not need to match the reference coordinate exactly; they only need to fall within the area where the nearest coordinate remains the reference point. The remaining single-class and multi-label combinations maintain this tendency to group around the corresponding reference, but their predictions also spread out near other references that have at least one of their individual classes, or even to the *No Finding* one. An example of that is the *PaOt* dispersion in Figure 6.5(b), whose coordinates mostly gather around the *PaOt*, *Ot* and *Pa* references. These patterns are more defined when dealing with highly represented combinations, but even the under-represented ones are not randomly placed and present such predisposition.

There are two single-class combinations worth discussing given their particularities - *Ot* and *Pn* in Figures 6.5(c) and 6.5(d), respectively. The *Ot* class aggregates lesions that do not fit any

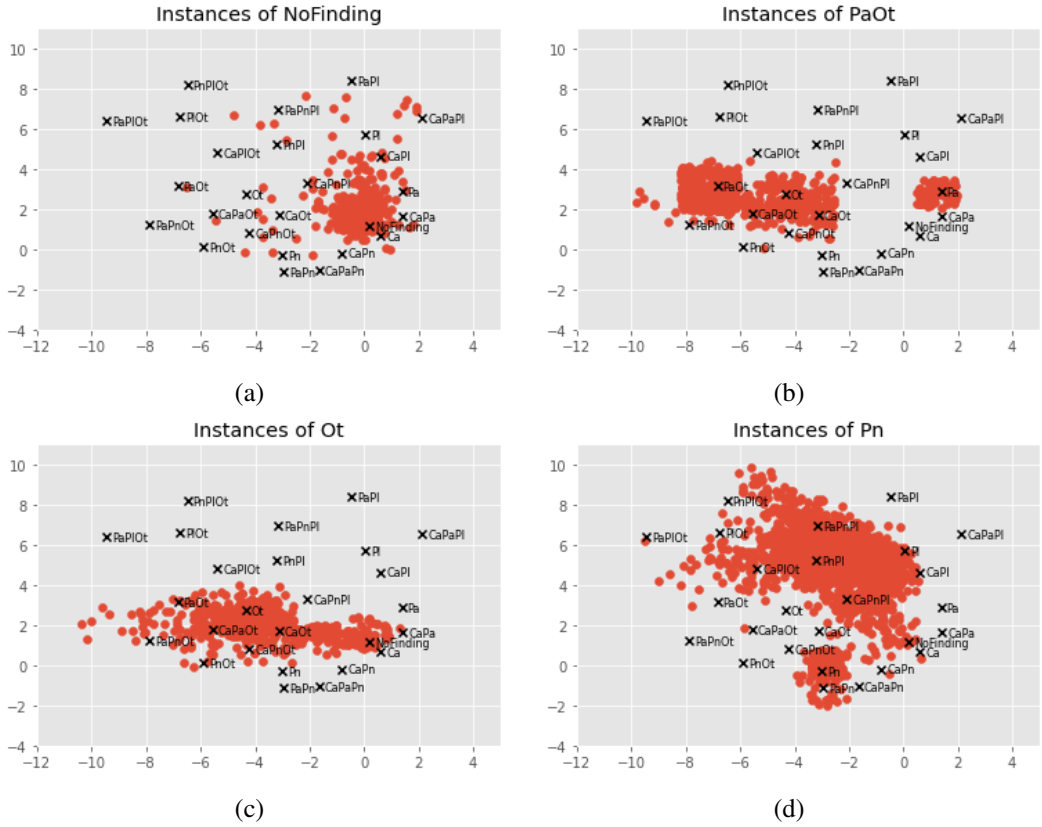



Figure 6.5: Predicted embedding coordinates obtained by the Phase II regressor for different combinations. The GT references are marked with black crosses. All predicted coordinates whose GT matches the combination in the title are marked with red dots.

of the other classes, meaning that their associated pathologies (hernia and/or emphysema) do not share common characteristics. Such aspect makes it difficult for the model to learn their specific features and recognize their presence in an image. This is aggravated by the fact that *Ot* is the less represented single-class combination, resulting in a broader dispersion cloud. Even so, the predicted coordinates are located in areas whose closest reference combinations contain the *Ot* class. The *Pn* case is also interesting because even though it constitutes an independent class, pneumothorax can be interpreted as a pleural lesion (*Pl* class). That is reflected in the plot, where a percentage of the predictions is located in between the *Pn* and *Pl* references, i.e. in a region whose multi-label references contain both *Pn* and *Pl*. This may indicate the partially shared features between the two classes. The same is not verified in the *Pl* scatter plot in Figure B.5, perhaps due to a higher representation that facilitates learning the specific *Pl* features.

### 6.4.4 Multi-label Classification

This phase of the methodology followed the same evaluation strategy as the multi-label baseline. Table 6.5 shows the results obtained for all classes, where the distance between each prediction and the closest references is taken into consideration to reach a final decision, depending on the

considered rule. Evaluating the precision-recall trade-off using the F1-score was prioritized to identify the most suitable approach for this particular classification task. This choice stems from the limitations of the accuracy metric when applied to imbalanced datasets. Consequently, the F1-score served as the key metric for assessing and determining the performance of the proposed method.

Table 6.5: Classification results of the CLARE-XR method. Results obtained for all instances predicted as positive for a given class, regardless of whether they are *No Finding*, single-pathology, or multi-label.The precision, recall, and F1-score metrics were calculated based on the predictions obtained by each rule (*R*).

| **Class** | **Accuracy** | | **Precision** | | **Recall** | | **F1-Score** | |
|---|---|---|---|---|---|---|---|---|
| | **R1** | **R2** | **R1** | **R2** | **R1** | **R2** | **R1** | **R2** |
| **Ca** | 0.700 | 0.475 | 0.025 | 0.026 | 0.328 | 0.622 | 0.047 | 0.051 |
| **Pa** | 0.634 | 0.631 | 0.477 | 0.481 | 0.418 | 0.644 | 0.446 | 0.551 |
| **Pn** | 0.943 | 0.891 | 0.355 | 0.224 | 0.347 | 0.596 | 0.351 | 0.326 |
| **Pl** | 0.872 | 0.794 | 0.510 | 0.369 | 0.680 | 0.808 | 0.538 | 0.507 |
| **Ot** | 0.974 | 0.961 | 0.424 | 0.314 | 0.473 | 0.609 | 0.447 | 0.414 |
| **Mean** | **0.825** | **0.750** | **0.358** | **0.283** | **0.449** | **0.656** | **0.375** | **0.370** |

| **Class** | **Accuracy** | | **Precision** | | **Recall** | | **F1-Score** | |
|---|---|---|---|---|---|---|---|---|
| | **R3** | **R4** | **R3** | **R4** | **R3** | **R4** | **R3** | **R4** |
| **Ca** | 0.558 | 0.499 | 0.028 | 0.027 | 0.541 | 0.608 | 0.052 | 0.052 |
| **Pa** | 0.630 | 0.629 | 0.470 | 0.469 | 0.426 | 0.434 | 0.447 | 0.451 |
| **Pn** | 0.956 | 0.943 | 0.500 | 0.382 | 0.320 | 0.454 | 0.390 | 0.415 |
| **Pl** | 0.878 | 0.870 | 0.529 | 0.504 | 0.663 | 0.688 | 0.588 | 0.582 |
| **Ot** | 0.975 | 0.971 | 0.446 | 0.395 | 0.499 | 0.529 | 0.471 | 0.452 |
| **Mean** | **0.799** | **0.782** | **0.394** | **0.355** | **0.490** | **0.543** | **0.390** | **0.390** |

Table 6.6: Classification results of the baseline, fixing the recall or precision obtained by the best setting of the CLARE-XR methodology (Rule 4, *R4*).

| **Class** | **Recall (fixed by R4)** | **Precision** | **F1-score** | **Precision (fixed by R4)** | **Recall** | **F1-score** |
|---|---|---|---|---|---|---|
| **Ca** | 0.608 | 0.152 | 0.243 | 0.027 | 0.990 | 0.053 |
| **Pa** | 0.434 | 0.568 | 0.492 | 0.469 | 0.752 | 0.577 |
| **Pn** | 0.454 | 0.230 | 0.305 | 0.382 | 0.195 | 0.258 |
| **Pl** | 0.688 | 0.361 | 0.473 | 0.504 | 0.426 | 0.462 |
| **Ot** | 0.529 | 0.083 | 0.144 | 0.395 | 0.002 | 0.004 |
| **MEAN** | **0.543** | **0.279** | **0.332** | **0.355** | **0.473** | **0.271** |

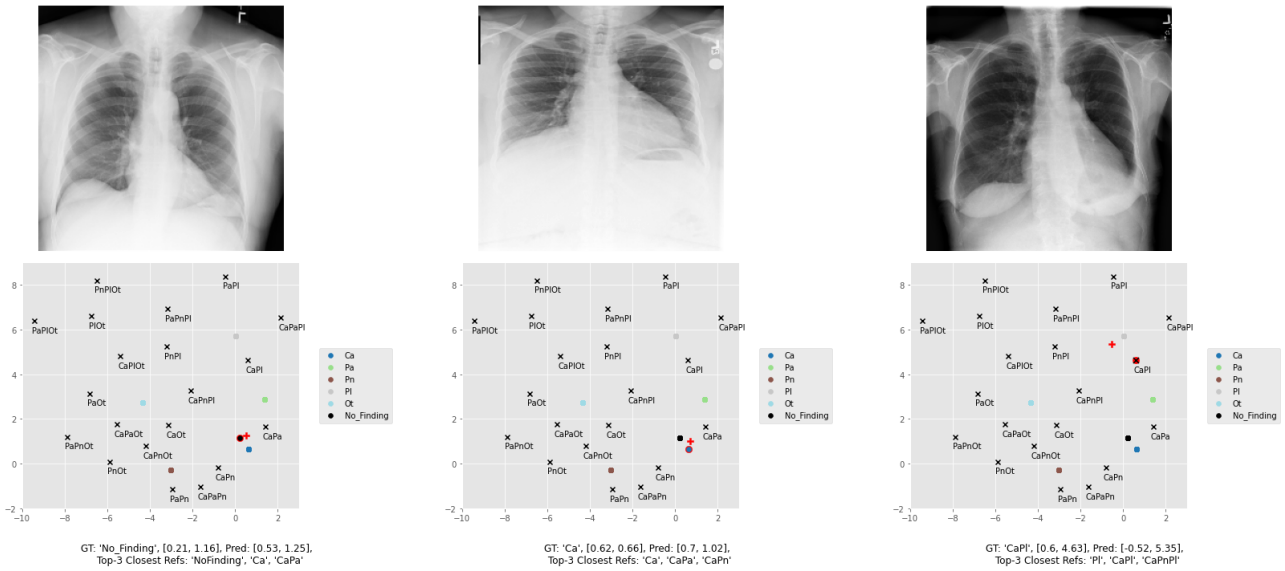
Rule 2 constitutes a "control" experiment in which all classes in the top-3 closest references are included in the final decision. As expected, it produces the most similar results to the multi-label baseline, achieving very low precision and higher recall, and thus promoting FP results over FN ones. Ideally, the remaining approaches should balance both metrics to fight this trade-off

and increase the overall F1-score. All proposed rules are successful at surpassing the precision and F1 values obtained by the multi-label classifier, partially compromising the recall. This is valid even for the straightforward top-1 approach in Rule 1. Both Rules 3 and 4 maximize the F1-score, revealing that the top-3 approaches are beneficial in this context. Considering the neighbor references adds useful information regarding recurrent pathologies and reduces the error when compared to the "hard" classification approaches of the baselines or top-1 rule. Rule 4 not only maximizes the F1 value, but also increases the recall by 5.3% in comparison to Rule 3, with a precision decrease of only 3.9%. For this reason, it was selected as the best approach for the task at hand.

Compared to the baseline, applying Rule 4 resulted in a 12.9% increase in precision and an 8.3% increase in F1-score with ROC-based optimization, or a 17.9% increase in recall and a 2.6% increase in F1-score with PR-based optimization. The CLARE-XR was the most efficient in classifying the presence of *Pn*, *Pl* and *Ot* lesions, consistently achieving higher F1-scores compared to the baseline.

While these results already highlight the superiority of the proposed method, additional experiments were conducted to enable a more thorough and equitable comparison with the baseline model, minding specifically the precision-recall trade-off. To compare the performance of both approaches at identical precision levels, the precision values achieved by Rule 4 were fixed, and the corresponding recall values for the baseline were obtained. This allows a direct comparison of the recall between the CLARE-XR method and the baseline under equivalent precision conditions. Similarly, the recall values achieved by Rule 4 were fixed, and the corresponding precision values for the baseline were measured, enabling a comparison of precision under identical recall conditions. These results are displayed in Table 6.6, and indicate that CLARE-XR consistently outperformed the baseline across these additional evaluation scenarios as well.

In conclusion, the relative superiority of the proposed model leads to believe that CLARE-XR is effective in associating certain regions of the reference space to specific pathologies, and that this improvement is not simply due to the shift from the typical multi-label classification approach to an embedding-based multi-class one. Additionally, a typical multi-class approach would be limited to the already-existing combinations (i.e. its outcome will always select a combination previously seen in the training set), while the multi-label baseline and the proposed method are more flexible in that aspect (i.e. it is possible to get an output combination that was not present in the training set).




(a) (b) (c)

(d) (e) (f)

Figure 6.6: Inference image examples and their respective predicted coordinates in the reference space (red plus sign). The *No Finding* and single-class references are plotted in colored dots, while the remaining multi-label combination coordinates are marked with black crosses. The corresponding GT reference is highlighted in red.

Table 6.7: GT labels and final decision based on the proposed rules, for multiple inference images.

| Figure | GT | Rule 1 | Rule 2 | Rule 3 | Rule 4 |
|---|---|---|---|---|---|
| **6.6(a)** | **No Finding** | No Finding | No Finding | No Finding | **No Finding** |
| **6.6(b)** | **Ca** | Ca | CaPaPn | Ca | **Ca** |
| **6.6(c)** | **CaPl** | Pl | CaPnPl | CaPl | **CaPl** |
| **6.6(d)** | **PaPl** | Pl | CaPaPl | Pl | **Pl** |
| **6.6(e)** | **PnPl** | CaPnPl | CaPnPlOt | CaPl | **CaPnPl** |
| **6.6(f)** | **PaOt** | No Finding | No Finding | No Finding | **No Finding** |

Figure 6.6 shows the predicted coordinates for multiple inference images, whose classification outcomes are discriminated in Table 6.7. In specific, Figures 6.6(a)-(c) present a *No Finding*, a single-class, and a multi-label instance in which the proposed Rule 4 correctly classified the images. These confirm that in general, the model tends to locate the prediction in the vicinity of the corresponding reference. Note that Figure 6.6(a) was automatically classified as *No Finding* because it is the nearest reference, while in the following two examples of *Ca* and *CaPl*, the outcome weighted both the first position and the recurrent pathologies in the top 3. Figures 6.6(d)-(e) are cases in which CLARE-XR is partially correct. Figure 6.6(d) corresponds to a *PaPl* instance classified as *Pl* (one TP class and a FN one). A possible justification for that is the *Pl* lesions being more evident in this image in comparison to the *Pa* ones. Figure 6.6(e) corresponds to a *PnPl* instance classified as *CaPnPl* (two TP classes and a FP one). This last example is essential to understand why Rule 4 is preferable to Rule 3, which would get a *CaPl* outcome instead and miss one of the GT classes (*Pn*). In the medical context, a higher recall prevents FN diagnoses such as this one. Finally, in Figure 6.6(f) the model fails to diagnose a *PaOt* instance, classifying it as *No Finding*.

To further evaluate the results obtained with Rule 4, the final predictions were analyzed per combination, rather than evaluating each class independently as in Table 6.5. Considering the five binary elements in the prediction and GT vectors, six error categories were established. These categories represent cases where a prediction is incorrect in 0, 1, 2, 3, 4, or all 5 elements. For example, a prediction of [1,1,0,0,0] compared to a GT of [1,0,1,0,0] falls into error category 2, as it includes one FN and one FP. Table 6.8 displays the number of instances per combination that fall into each error category, along with the resulting weighted mean error. This analysis identifies the most significant error category for each combination and reveals that there is no direct relationship between the frequency of each combination and its mean error. The most common error categories are 1 and 0, also represented in Figure 6.7, indicating that the model typically makes either one false positive/negative per prediction (45.7% of cases), or produces entirely accurate predictions (35.6% of cases). When analyzing the most frequent error category in the last two columns of Table 6.8 (i.e. category 1, where a single element is misclassified), it is clear that the model generally favors false positives over false negatives. This observation aligns with the results in Table 6.5. However, this trend varies depending on the specific combination, whose histograms

are presented in Figures B.6-B.8 in Appendix B.

Table 6.8: Classification results of the CLARE-XR method per combination, ordered by their frequency and using Rule 4. *N/A* stands for Not Available.

| **Combination** | **Error Category Count** | | | | | | **Weighted Mean Error** | **Error Category 1 Count** | |
|---|---|---|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **5** | | **FN** | **FP** |
| **NoFinding** | 37844 | 19203 | 1731 | 1439 | 144 | 0 | 0.457 | N/A | 19203 |
| **Pa** | 0 | 23584 | 610 | 3671 | 504 | 71 | 1.343 | 9725 | 13859 |
| **PaPl** | 367 | 895 | 5569 | 253 | 51 | 13 | 1.827 | 772 | 123 |
| **Pl** | 475 | 3456 | 371 | 1030 | 0 | 0 | 1.367 | 747 | 2709 |
| **Pn** | 50 | 1091 | 556 | 455 | 42 | 0 | 1.703 | 742 | 349 |
| **PaPn** | 54 | 41 | 681 | 303 | 146 | 0 | 2.364 | 22 | 19 |
| **Ca** | 25 | 1060 | 8 | 0 | 0 | 0 | 0.984 | 566 | 494 |
| **Ot** | 1 | 551 | 189 | 158 | 99 | 5 | 1.819 | 363 | 188 |
| **PaOt** | 45 | 19 | 315 | 56 | 65 | 76 | 2.530 | 0 | 19 |
| **CaPl** | 106 | 33 | 393 | 0 | 0 | 0 | 1.539 | 18 | 15 |
| **CaPa** | 331 | 5 | 188 | 3 | 0 | 0 | 0.740 | 2 | 3 |
| **PnPl** | 11 | 162 | 225 | 61 | 56 | 1 | 1.984 | 29 | 133 |
| **PnOt** | 0 | 118 | 150 | 61 | 10 | 0 | 1.891 | 3 | 115 |
| **PaPnPl** | 1 | 18 | 110 | 166 | 23 | 15 | 2.712 | 16 | 2 |
| **CaPaPl** | 3 | 230 | 19 | 37 | 0 | 0 | 1.311 | 230 | 0 |
| **PaPnOt** | 14 | 32 | 70 | 73 | 11 | 0 | 2.175 | 18 | 14 |
| **PlOt** | 3 | 17 | 89 | 46 | 9 | 0 | 2.250 | 7 | 10 |
| **PaPlOt** | 0 | 3 | 8 | 56 | 33 | 0 | 3.190 | 3 | 0 |
| **PnPlOt** | 1 | 5 | 22 | 29 | 2 | 1 | 2.483 | 2 | 3 |
| **CaOt** | 1 | 3 | 17 | 0 | 0 | 0 | 1.762 | 0 | 3 |
| **CaPlOt** | 0 | 1 | 3 | 4 | 3 | 0 | 2.818 | 1 | 0 |
| **CaPn** | 0 | 4 | 5 | 0 | 0 | 0 | 1.556 | 4 | 0 |
| **CaPaPn** | 0 | 1 | 3 | 2 | 0 | 0 | 2.167 | 1 | 0 |
| **CaPnPl** | 1 | 0 | 1 | 1 | 0 | 0 | 1.667 | 0 | 0 |
| **CaPnOt** | 1 | 0 | 1 | 0 | 0 | 0 | 1.000 | 0 | 0 |
| **CaPaOt** | 0 | 0 | 0 | 2 | 0 | 0 | 3.000 | 0 | 0 |
| **Sum** | **39334** | **50532** | **11334** | **7906** | **1198** | **182** | | **13271** | **37261** |

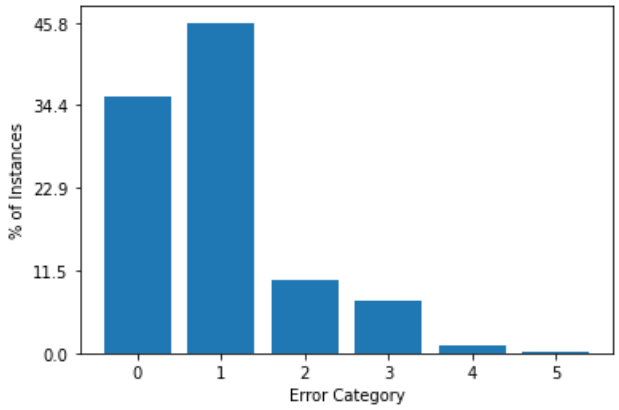



Figure 6.7: Frequency of instances per error category.

### 6.4.5 PCA *vs.* Label Autoencoder

The present section revisits the three phases of the work, substituting the Phase I autoencoder with PCA to obtain the 2D representation of the GT vectors. As introduced in Section 6.3.2.1, the goal was to evaluate whether a simpler linear method, such as PCA, could serve as an effective alternative to the autoencoder, offering a less complex and demanding solution for the first stage of the framework.

This alternative dimensionality reduction technique receives the original binary vectors, representing all 26 combinations, and produces the map shown in Figure 6.8. At first glance, in comparison to the autoencoder's references in Figure 6.4, the PCA-generated coordinates are more symmetrical and less equidistant from one another. This occurs because PCA focuses on maximizing variance along the first principal component, in the horizontal axis, resulting in a more uneven distribution of points across the two dimensions. The second principal component divides the 26 combinations into seven vertical regions: *Ca* is placed in $y = 0.894$, multi-label references with *Ca* and another class in $y = 0.671$, multi-label references with *Ca* and two other classes in $y = 0.447$, *No Finding* in $y = 0.000$, the remaining single-class references in $y = -0.224$, multi-label references of two classes without *Ca* in $y = -0.447$, and multi-label references of three classes without *Ca* in $y = -0.671$. The *No Finding* reference is positioned at the center of the plot, while the others are distributed around it, creating a sense of centrality in their arrangement.

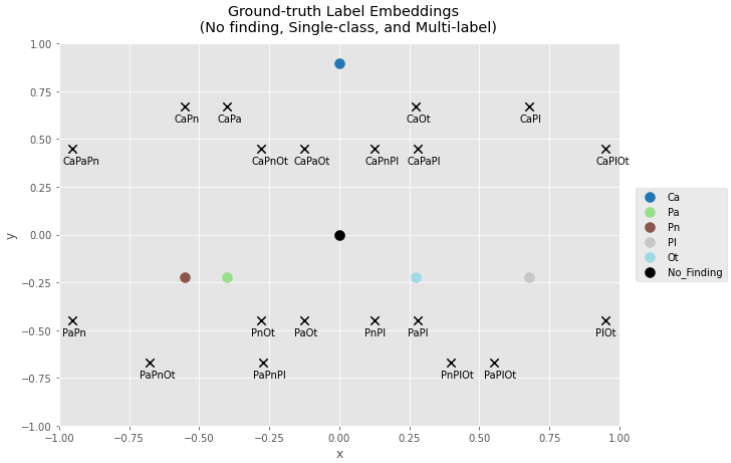



Figure 6.8: Reference label embeddings obtained by the Phase I PCA. The *No Finding* and single-class GT coordinates are plotted in colored dots, while the remaining multi-label combination coordinates are marked with black crosses.

The Phase II regressor was trained under the same conditions described earlier. The only difference is that it used the 2D GT derived from PCA, rather than the 2D GT produced by the encoder. The results indicate an overall MSE of $0.046 \pm 0.002$ after cross-validation. This value cannot be compared with the one obtained in the previous setting, as the autoencoder and the PCA operate on significantly different scales.

The established rules were once again applied to determine the final decision for each instance and evaluate the classification based on these alternative GT representations. Table 6.9 displays the results for this experimental setting using the previously considered evaluation metrics. Rule 4 benefited the classification when using the encoder-based representations, by considering both the top-1 and the recurrent classes. Here, Rule 2 achieves the highest F1-score among the four, which assumes all top-3 references are present in each instance. The latter results in lower F1-scores for three out of the five classes (*Pn*, *Pl* and *Ot*) in comparison to the best rule of the autoencoder approach. In any case, the PCA experiment was not capable of outperforming the previous experiment's F1-score.

Table 6.9: Classification results of the CLARE-XR method, using the PCA-generated references. Results obtained for all instances predicted as positive for a given class, regardless of whether they are *No Finding*, single-pathology, or multi-label. *R* stands for Rule.

| **Class** | **Accuracy** | | **Precision** | | **Recall** | | **F1-Score** | |
|---|---|---|---|---|---|---|---|---|
| | **R1** | **R2** | **R1** | **R2** | **R1** | **R2** | **R1** | **R2** |
| **Ca** | 0.979 | 0.979 | 0.550 | 0.539 | 0.425 | 0.433 | 0.480 | 0.480 |
| **Pa** | 0.696 | 0.687 | 0.613 | 0.557 | 0.370 | 0.539 | 0.461 | 0.548 |
| **Pn** | 0.927 | 0.701 | 0.209 | 0.107 | 0.234 | 0.789 | 0.221 | 0.189 |
| **Pl** | 0.875 | 0.876 | 0.565 | 0.522 | 0.209 | 0.616 | 0.305 | 0.565 |
| **Ot** | 0.859 | 0.668 | 0.067 | 0.045 | 0.411 | 0.680 | 0.116 | 0.084 |
| **Mean** | **0.867** | **0.782** | **0.401** | **0.354** | **0.330** | **0.611** | **0.316** | **0.373** |

| **Class** | **Accuracy** | | **Precision** | | **Recall** | | **F1-Score** | |
|---|---|---|---|---|---|---|---|---|
| | **R3** | **R4** | **R3** | **R4** | **R3** | **R4** | **R3** | **R4** |
| **Ca** | 0.979 | 0.979 | 0.549 | 0.549 | 0.426 | 0.426 | 0.479 | 0.480 |
| **Pa** | 0.667 | 0.695 | 0.583 | 0.600 | 0.182 | 0.395 | 0.278 | 0.476 |
| **Pn** | 0.876 | 0.845 | 0.160 | 0.150 | 0.423 | 0.536 | 0.232 | 0.234 |
| **Pl** | 0.893 | 0.892 | 0.613 | 0.596 | 0.502 | 0.544 | 0.552 | 0.569 |
| **Ot** | 0.853 | 0.785 | 0.067 | 0.054 | 0.428 | 0.523 | 0.116 | 0.098 |
| **Mean** | **0.854** | **0.839** | **0.394** | **0.390** | **0.392** | **0.485** | **0.331** | **0.371** |

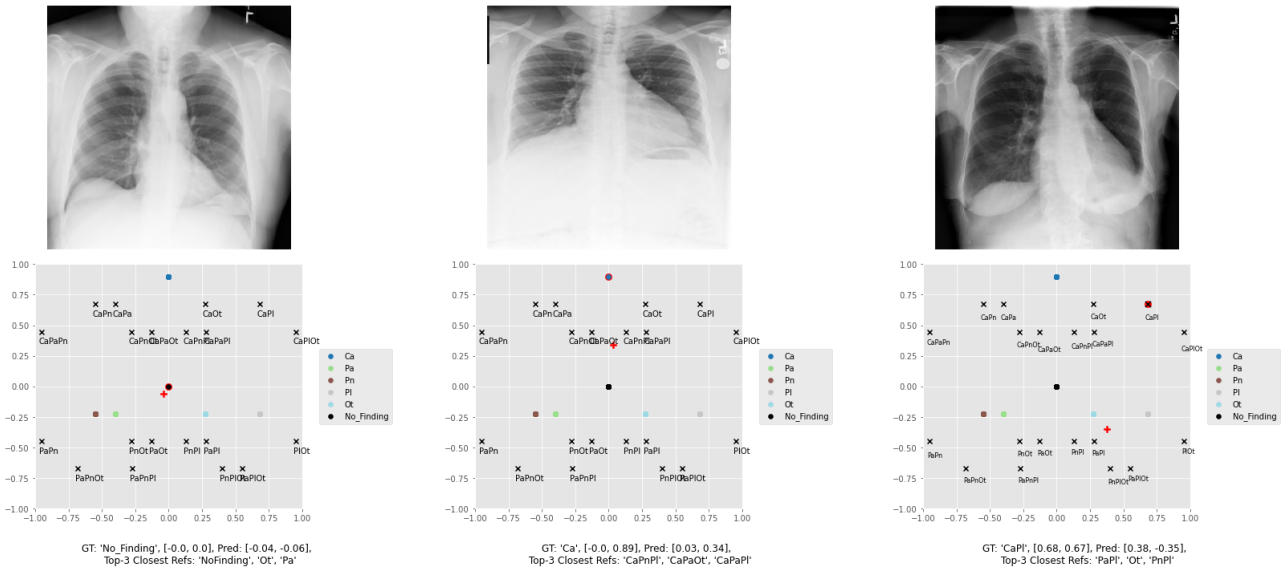
Figure 6.9 illustrates the corresponding results for the same instances shown in Figure 6.6. The central positioning of the *No Finding* coordinates simplifies the regressor's task of placing the "normal" images near this reference and leads to mostly accurate decisions. Figure 6.9(a) illustrates this, where both approaches agree on the nearest reference. Similarly, they both place the radiograph in Figure 6.9(f) near *No Finding*, despite its GT being *PaOt*. However, in most instances, the autoencoder and PCA produce distinct outcomes, as observed in Figures 6.9(b-e). Table 6.10 presents the possible decisions for the instances in Figure 6.9. It becomes clear in Figures 6.9(d,e) that the only way to include all GT classes in the diagnosis is by applying Rule 2, which in turn also results in a higher number of FPs and leads to a less accurate decision.

To further illustrate why the autoencoder outperforms PCA, Figure 6.10 compares the two outcomes for the same instance of *Ca*. Although the predicted coordinate is better positioned in

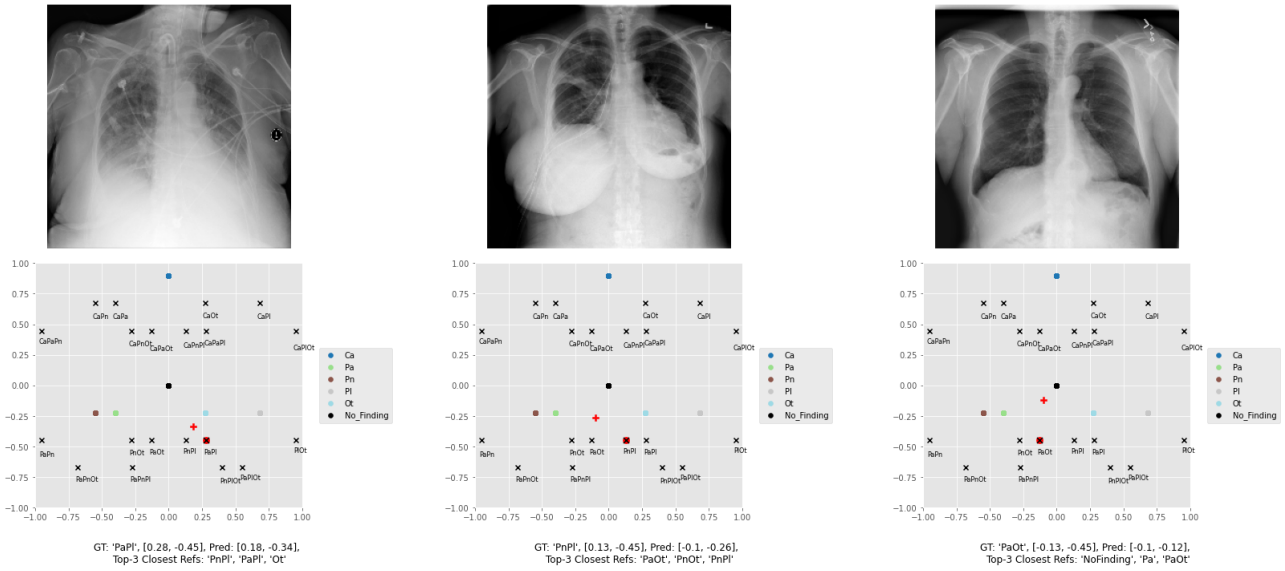
the PCA map (i.e. closer to the *Ca* reference), the autoencoder's Rule 4 leads to *CaPa* decision, resulting in one TP and one FP. In contrast, the PCA's Rule 2 produces a *CaPnPlOt* decision, which includes one TP and three FPs. This aligns with the previous conclusions.

In sum, these results may be attributed to the linear nature of PCA, which was unable to capture the complex relationships necessary to produce more insightful label embeddings. The autoencoder's non-linear transformation of the data, on the other hand, allowed it to learn the intricate relationships between labels to get better 2D representations, even if the resulting reference positioning is less visually intuitive. In light of these conclusions, the PCA experimental setup was not pursued further.

The following section explores the explainability component, following the results obtained in Sections 6.4.2-6.4.4. It discusses how the diagnostic information is disclosed to the clinicians, as well as the corresponding explanations, highlighting the advantages of CLARE-XR in this aspect.

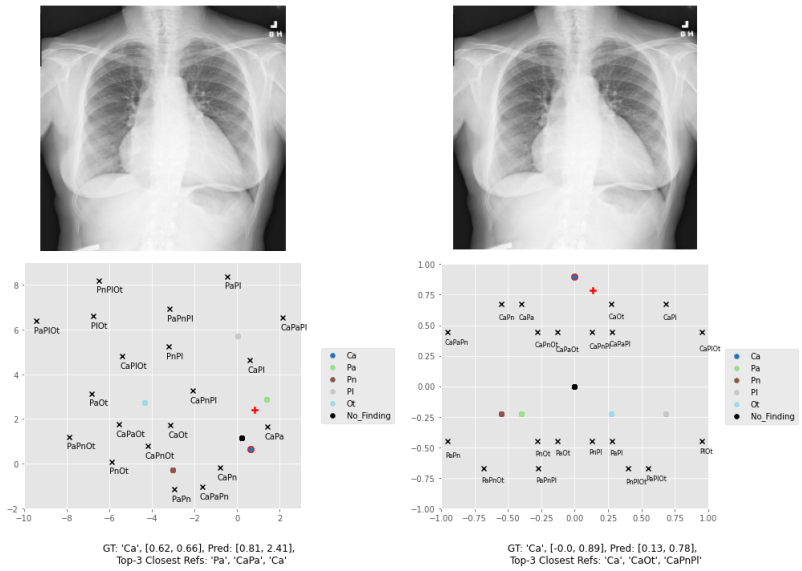



(a) (b) (c)




(d) (e) (f)

Figure 6.9: Inference image examples and their respective predicted coordinates in the reference space obtained with PCA (red plus sign). The *No Finding* and single-class references are plotted in colored dots, while the remaining multi-label combination coordinates are marked with black crosses. The corresponding GT reference is highlighted in red.




(a) (b)

Figure 6.10: Inference image and its respective predicted coordinate (red plus sign) in the reference space obtained with the autoencoder and PCA, respectively. The *No Finding* and single-class references are plotted in colored dots, while the remaining multi-label combination coordinates are marked with black crosses. The corresponding GT reference is highlighted in red.

Table 6.10: GT labels and final decision based on the proposed rules and the PCA encoding, for multiple inference images.

| Figure | GT | Rule 1 | Rule 2 | Rule 3 | Rule 4 |
|---|---|---|---|---|---|
| **6.9(a)** | **No Finding** | No Finding | **No Finding** | No Finding | No Finding |
| **6.9(b)** | **Ca** | CaPnPl | **CaPaPnPlOt** | CaPaPl | CaPaPnPl |
| **6.9(c)** | **CaPl** | PaPl | **PaPnPlOt** | Pl | PaPl |
| **6.9(d)** | **PaPl** | PnPl | **PaPnPlOt** | Pl | PnPl |
| **6.9(e)** | **PnPl** | PaOt | **PaPnPlOt** | PnOt | PaPnOt |
| **6.9(f)** | **PaOt** | No Finding | **No Finding** | No Finding | No Finding |

### 6.4.6 Explainability

It is important to discuss the information that will be disclosed to the clinicians, keeping in mind the goal of providing clear interpretable justifications along with the final decision. Suppose that the specialist inquires the system regarding a certain inference radiograph. The system will disclose the top-3 most likely combinations, letting them know that that image is attributed to a region in which those classes are predominant. The specialist is shown the final decision, with a brief textual explanation of how it was achieved based on Rule 4. E.g., the following text would be provided for the image in Figure 6.6(c): "`This radiograph is likely to include findings associated with cardiac and pleural lesions, resulting in a` *`CaPl`* `diagnosis. The system has reached this decision because` *`Pl`* `is the most probable finding`

present in the image, and *Ca* is a recurrent finding in images with similar characteristics. Here is the complete list of possible lesions present: *Pl*, *CaPl*, *CaPnPl*."

Additionally, a strategy was proposed to get complementary visual explanations based on the scatter plots, to validate the proposed diagnosis. Suppose that the model is provided the same inference radiograph in Figure 6.6(c) with a *CaPl* prediction. The first step of the proposed XAI approach is to look at all the training set samples with that GT and position them in the reference space (red dots in Figure 6.11). Then, images from the training set with coordinates closest to the predicted coordinate of the inference radiograph are selected and displayed as the most similar examples to that radiograph. The resulting images are shown in Figure 6.12, and exhibit comparable characteristics to the instances in Figure 6.6, in the sense that they are represented by analogous intensities and/or shapes (e.g. of the lesions, lungs, clavicles...).

In short, the XAI approach generates a personalized explanation for each inference radiograph, providing examples that match that specific finding and patient. It is also possible to retrieve more than just 9 images if more examples are needed. Such functionality can be beneficial for non-specialist physicians, less experienced radiologists and/or educational purposes. Overall, the XAI stage of the work focused on justifying *how* that decision was achieved (using rule to classify each case) and *why* the model decided so (showing similar examples that support the conclusion, based on the features learned by the model and the consequent positioning of the images in the reference space).

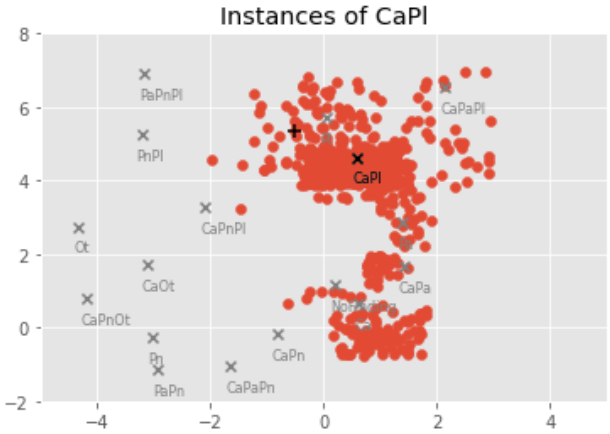



Figure 6.11: Embedding coordinates obtained by the Phase II regressor for the training instances annotated with *CaPl* (red dots). The GT reference is marked with the black cross, and the predicted coordinate for the current inference image is marked with a black plus sign.




Figure 6.12: Visual explanations for the inference image examples in Figure 6.6, represented on the left.

### 6.4.7 Limitations

This study initially applied the same concepts and methods to the original 14 classes of the ChestX-ray14 dataset, corresponding to a total of 470 possible label combinations with up to 3 co-occurring pathologies. However, only 369 of the 470 combinations were actually represented in the dataset, and 260 of those were represented by less than 30 images. For this reason, it was not possible to get statistically significant results, and it was necessary to reduce the complexity of the exercise by considering the 5 broader categories. This scenario ensures that all 26 possible combinations are represented in the dataset, with only 8 combinations being under-represented.

This affects mostly the *Ca* class, present in 7 of those combinations. The class's lowest F1-scores in Table 6.5, as well as the scarcity of examples for the explanations (e.g. Figure 6.12(e)), are a consequence of such limitation.

While this dataset is not ideal, the obtained results are justifiable and ensure the proof of concept of the CLARE-XR methodology. The obtained results cannot be directly compared to any state-of-the-art publications because no other publication considered the same 5 broader categories, nor disclosed the precision, recall, and F1-score metrics on ChestX-ray14. Although there is room for improvement, given the lower precision and moderate recall values, Table 6.4 indicates that both the baseline and CLARE-XR align with the state of the art, sharing identical limitations. Nevertheless, these results should be interpreted with caution, as that approach employs the significantly larger CheXpert dataset.

## 6.5 Concluding Remarks

This work introduced the novel CLARE-XR approach that involves label embedding prior to the common image embedding, to enhance both predictive performance and model transparency. CLARE-XR automatically converts the original annotations into a continuous 2D embedded space, capturing the hidden structure of the labels. This transformation allows the multi-label classification to be framed as a regression task, where an image is mapped to a 2D coordinate instead of binary label vectors, positioning it within the embedded reference space. In this space, images with similar features are placed near each other, contextualizing neighboring information. The primary advantage of this method lies in its inherently interpretable framework, based on the lower-dimensional representation of the GT labels. This allows the definition of a clear set of classification rules and consequently provides physicians with justifications for the final diagnosis. Furthermore, visual explanations are offered through label embedding-based automatic image retrieval methods, showing tailored similar examples of the findings. This approach aligns with common clinical practices, where comparing radiographs is a standard method for confirming diagnoses. The results show that this framework outperforms the multi-label classification architecture across various evaluation settings on the CXR14 dataset. CLARE-XR lays the groundwork as a proof of concept, striving to balance predictive quality and explainability to foster safer and more responsible AI applications in healthcare.

# Chapter 7

# Conclusions and Future Work

Thoracic pathologies are highly prevalent worldwide, with increasing severity due to aging populations and lifestyle factors. Prevention and early treatment are crucial, which has led to extensive research into CAD systems to act as a second opinion. This thesis tackles key challenges in the field, addressing limitations of the state-of-the-art visual XAI methods, which stem e.g. from bias-inducing artifacts and the inherent difficulties in interpreting visual explanations.

Chapter 2 emphasized how DL models have achieved unprecedented performance in extracting information from CXR data, backed by more powerful computational resources and optimized training routines. However, their integration into clinical practice remains far from fully established, given significant limitations. One of the most pressing limitations is the lack of sufficient quality and quantity of training data, leading to potential biases in the learning process. These biases are often obscured by another major limitation — DL's inherent lack of explainability — which makes them difficult to detect and address. Chapter 3 concluded that an effective XAI approach in healthcare must address the unique needs of the domain and its users to provide actionable and reliable explanations. Accordingly, current XAI evaluation frameworks emphasize the predictive and descriptive accuracy of the explanations, their relevance, and expert involvement. Engaging clinicians early in the development of explainability guidelines is crucial for fostering trust and ensuring that the models' outputs are communicated in a practical and user-friendly manner. Keeping that in mind, three studies were conducted to improve the explainability outcomes of different tasks, described below.

The work in Chapter 4 tackles image artifacts in the periphery of the images, which contribute to biased predictions and misleading explanations. The STERN methodology addresses this issue by limiting the image to the thoracic ROI before classifying it as normal or abnormal, thus enhancing both model accuracy and XAI performance in the CheXpert dataset. This process directs the model to focus on relevant diagnostic features by using a novel STERN loss function. The methodology's non-isotropic delimitation of the anatomical area has been shown to produce cleaner intensity profiles and more uniform images concerning thorax positioning compared to the original images. Overall, the advantages of STERN are evident in its architecture, which operates end-to-end solely on image-level annotations. The method delivers results comparable to those

achieved with YOLO-cropped images but with reduced computational cost and no need for spatial GT labels.

Chapter 5 discusses the limitations of current explainability heat maps, namely Grad-CAMs. Among others, the normalization of heat maps can obscure the relative importance of different regions within an image, as well as distort the relative importance of regions across multiple images, making it difficult to compare and interpret their highlighted features. Confident-CAM improves the interpretation of the Grad-CAM results by measuring the distance between the output probability and the corresponding class threshold, using a confidence score to generate nearly null maps when the initial explanations are most likely incorrect. The results reveal that high-confidence predictions are associated with more accurate explanation maps (concordant Grad-CAMs and Confident-CAMs), while lower-confidence predictions often lead to less reliable explanations. This discrepancy can be attributed to two factors: either the model initially predicted the wrong class, and in this case, there are no lesions to highlight, or the model correctly identified the class but failed to focus on the correct anatomical region. In both scenarios, the Confident-CAM will be inconsistent with the previous explanation and nearly null. The study validated the proposed algorithm in a multi-label CXR classification task, targeting 14 radiological findings from the ChestX-ray14 dataset with significant class imbalance, and confirming that confidence scores are effective in distinguishing between likely accurate and inaccurate explanations.

In opposition to Chapter 5, where the research builds on existing XAI methods, Chapter 6 introduces an entirely new approach for a multi-label classification framework specifically designed for enhanced explainability. Given a CXR image as input, CLARE-XR identifies associated pathologies and provides explanations based on the presentation of similar cases. The novelty of this work lies in a regression model that maps images into a 2D latent space, containing the reference coordinates of all findings. These references are created by embedding the binary ground truth annotations into 2D coordinates before the regression step, using an encoder. Images with similar features are positioned close to each other in this space, facilitating contextual understanding of neighboring information. CLARE-XR stands out for its inherent interpretability, as it establishes a clear set of classification rules, thereby offering physicians textual justifications for the final diagnosis. Additionally, it provides visual explanations by retrieving similar instances from the training set to compare with the inference radiograph and confirm the decision. This approach offers a valuable alternative to the often inadequate low-resolution Grad-CAMs, while simultaneously improving classification performance in ChestX-ray14.

In sum, this PhD thesis presented several innovative approaches for a CXR-based CAD system, designed to prioritize abnormal cases and identify specific findings in the images. The core contribution of this work is the development of a novel, robust XAI framework, CLARE-XR, whose proof of concept improves transparency by providing clearer, interpretable justifications for the model's recommendations. This, in turn, empowers healthcare professionals to better understand, validate, and trust its outputs.

The three studies emphasize the need for larger, higher-quality datasets to support their validity. In particular, the development of extensive, international public datasets that accurately

capture the diversity and complexity of specific findings could stimulate more robust algorithms and fairer performance comparisons. Data collectors should prioritize minimizing annotation errors and increasing the representation of rarer conditions to ensure a more comprehensive dataset coverage.

Future work should begin by exploring the modularity of the CLARE-XR framework and evaluating alternative techniques for each phase. Similarly to the studies conducted for the encoder, the regression phase could potentially benefit from exploring alternative approaches to replace the current deep neural network. Additionally, the training strategy for each phase can be further refined, e.g. by employing contrastive learning to create a latent space where similar samples and/or references are closer together and dissimilar ones are farther apart. Moreover, the interaction between the encoder and the regressor encourages further research. Currently, these phases are trained separately, but future work could focus on simultaneously optimizing the positioning of references and image coordinates, enabling the encoder and regressor to be trained together for possibly better results.

# Publications

The studies carried out during the development of this PhD thesis resulted in the following scientific publications, ordered chronologically.

### Scientific Journals

- **Rocha, J.**, Mendonça, A. M., and Campilho, A. (2021). A Review on Deep Learning Methods for Chest X-Ray based Abnormality Detection and Thoracic Pathology Classification. U. Porto Journal of Engineering, 7(4), 16-32.
- Pereira, S. C., **Rocha, J.**, Campilho, A., Sousa, P., and Mendonça, A. M. (2023). Lightweight multi-scale classification of chest radiographs via size-specific batch normalization. Computer Methods and Programs in Biomedicine, 236, 107558.
- **Rocha, J.**, Pereira, S. C., Pedrosa, J., Campilho, A., and Mendonça, A. M. (2024). STERN: Attention-driven Spatial Transformer Network for abnormality detection in chest X-ray images. Artificial Intelligence in Medicine, 147, 102737.
- Pereira, S. C., **Rocha, J.**, Gaudio, A., Smailagic, A., and Mendonça, A. M. (2024). Addressing Chest Radiograph Projection Bias in Deep Classification Models. In Medical Imaging with Deep Learning (pp. 1199-1210). PMLR.
- Pereira, S. C., **Rocha, J.**, Campilho, A., and Mendonça, A. M. (2024). Distribution-Based Detection of Radiographic Changes in Pneumonia Patterns: A COVID-19 Case Study. Heliyon, 10(16).
- **Rocha, J.**, Pereira, S. C., Sousa, P., Campilho, A., and Mendonça, A. M. (2024). CLARE-XR: explainable regression-based classification of chest radiographs with label embeddings. Scientific Reports, 14(1), 31024.

### Conference Proceedings

- **Rocha, J.**, Pereira, S., Campilho, A., and Mendonça, A. M. (2021). Segmentation of COVID-19 lesions in CT Images. In 2021 IEEE EMBS international conference on biomedical and health informatics (BHI) (pp. 1-4). IEEE.

- Sousa, M. Q. E., Pedrosa, J., **Rocha, J.**, Pereira, S. C., Mendonça, A. M., and Campilho, A. (2021). Chest Radiography Few-Shot Image Synthesis for Automated Pathology Screening Applications. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 1791-1798). IEEE.

- **Rocha, J.**, Pereira, S. C., Pedrosa, J., Campilho, A., and Mendonça, A. M. (2022). Attention-driven spatial transformer network for abnormality detection in chest x-ray images. In 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS) (pp. 252-257). IEEE.

- Belo, R., **Rocha, J.**, Mendonça, A. M., and Campilho, A. (2023). An active learning approach for support device detection in chest radiography images. In Fifteenth International Conference on Machine Vision (ICMV 2022) (Vol. 12701, pp. 271-278). SPIE.

- Belo, R., **Rocha, J.**, and Pedrosa, J. (2023). Leveraging Longitudinal Data for Cardiomegaly and Change Detection in Chest Radiography. In Iberoamerican Congress on Pattern Recognition (pp. 434-448). Cham: Springer Nature Switzerland.

- **Rocha, J.**, Pereira, S. C., Campilho, A., and Mendonça, A. M. (2023). Confident-CAM: Improving Heat Map Interpretation in Chest X-Ray Image Classification. In 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 4116-4123). IEEE.

- Pereira, P., **Rocha, J.**, Pedrosa, J., and Mendonça, A. M. (2024). Evaluating Visual Explainability in Chest X-Ray Pathology Detection. In 2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON) (pp. 1116-1121). IEEE.

- Pereira, S. C., Pedrosa, J., **Rocha, J.**, Sousa, P., Campilho, A., and Mendonça, A. M. (2024, December). DeepClean – Contrastive Learning Towards Quality Assessment in Large-Scale CXR Data Sets. In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 4253-4260). IEEE.

# References


Sudil Hasitha Piyath Abeyagunasekera, Yuvin Perera, Kenneth Chamara, Udari Kaushalya, Prasanna Sumathipala, and Oshada Senaweera. LISA : Enhance the explainability of medical images unifying current XAI techniques. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pages 1–9, April 2022. doi: 10.1109/I2CT54291.2022.9824840.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31:9505–9515, 2018.

Heba M. Afify, Kamel K. Mohammed, and Aboul Ella Hassanien. Novel prediction model on OSCC histopathological images via deep transfer learning combined with Grad-CAM interpretation. *Biomedical Signal Processing and Control*, 83:104704, May 2023. ISSN 1746-8094. doi: 10.1016/j.bspc.2023.104704. URL https://www.sciencedirect.com/science/article/pii/S1746809423001374.

Md Manjurul Ahsan, Kishor Datta Gupta, Mohammad Maminur Islam, Sajib Sen, Md. Lutfar Rahman, and Mohammad Shakhawat Hossain. COVID-19 Symptoms Detection Based on NasNetMobile with Explainable AI Using Various Imaging Modalities. *Machine Learning and Knowledge Extraction*, 2(4):490–504, October 2020. ISSN 2504-4990. doi: 10.3390/make2040027. URL https://www.mdpi.com/2504-4990/2/4/27.

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-Embedding for Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (7):1425–1438, July 2016. ISSN 1939-3539. doi: 10.1109/TPAMI.2015.2487986. URL https://ieeexplore.ieee.org/abstract/document/7293699?casa_token=YKB45yBQlq0AAAAA:MStCYF8gGH68pxQlLcyWNzCeS0-QY8t0szGweqHPtdwf0i3pXoLwC2LLi6bPLxfMTCNiutsBpw. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Algarve Hospital and University Centre. Incidences. 2018. URL https://epos.myesr.org/posterimage/esr/ecr2018/143153/mediagallery/757035.

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99:101805, November 2023. ISSN 1566-2535. doi: 10.1016/j.inffus.2023.101805. URL https://www.sciencedirect.com/science/article/pii/S1566253523001148.

Johannes Allgaier, Lena Mulansky, Rachel Lea Draelos, and Rüdiger Pryss. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artificial Intelligence in Medicine*, 143:102616, September 2023. ISSN

0933-3657. doi: 10.1016/j.artmed.2023.102616. URL https://www.sciencedirect.com/science/article/pii/S0933365723001306.

Ahad Alloqmani, Yoosef B., Asif Khan, and Fawaz Alsolami. Deep Learning based Anomaly Detection in Images: Insights, Challenges and Recommendations. *International Journal of Advanced Computer Science and Applications*, 12, January 2021. doi: 10.14569/IJACSA.2021.0120428.

American Cancer Society. Global Cancer Facts & Figures. (5th Edition), 2024. URL https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/global-cancer-facts-and-figures/global-cancer-facts-and-figures-2024.pdf.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, and others. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

Ivo M. Baltruschat, Leonhard Steinmeister, Harald Ittrich, Gerhard Adam, Hannes Nickisch, Axel Saalbach, Jens von Berg, Michael Grass, and Tobias Knopp. When Does Bone Suppression And Lung Field Segmentation Improve Chest X-Ray Disease Classification? In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1362–1366, April 2019. doi: 10.1109/ISBI.2019.8759510. ISSN: 1945-8452.

Oren Barkan, Omri Armstrong, Amir Hertz, Avi Caciularu, Ori Katz, Itzik Malkiel, and Noam Koenigstein. GAM: Explainable Visual Similarity and Classification via Gradient Activation Maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 68–77, Virtual Event Queensland Australia, October 2021. ACM. ISBN 978-1-4503-8446-9. doi: 10.1145/3459637.3482430. URL https://dl.acm.org/doi/10.1145/3459637.3482430.

Arindam Baruah, Rakesh Kumar, and Meenu Gupta. Traffic Sign Recognition Using Deep Learning Neural Network and Spatial Transformer. In *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pages 1–8, May 2023. doi: 10.1109/ACCAI58221.2023.10199560.

Subrato Bharati, Prajoy Podder, and M. Rubaiyat Hossain Mondal. Hybrid deep learning for detecting lung diseases from X-ray images. *Informatics in Medicine Unlocked*, 20:100391, January 2020. ISSN 2352-9148. doi: 10.1016/j.imu.2020.100391. URL https://www.sciencedirect.com/science/article/pii/S2352914820300290.

British Heart Foundation. Global Heart & Circulatory Diseases Factsheet. 2024. URL https://www.bhf.org.uk/-/media/files/for-professionals/research/heart-statistics/bhf-cvd-statistics-global-factsheet.pdf?rev=e61c05db17e9439a8c2e4720f6ca0a19&hash=6350DE1B2A19D939431D876311077C7B.

David A Broniatowski. Psychological foundations of explainability and interpretability in artificial intelligence. Technical Report NIST IR 8367, National Institute of Standards and Technology (U.S.), Gaithersburg, MD, April 2021. URL https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf.

Luca Brunese, Francesco Mercaldo, Alfonso Reginelli, and Antonella Santone. Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays. *Computer Methods and Programs in Biomedicine*, 196:105608, November 2020. ISSN 01692607. doi: 10.1016/j.cmpb.2020.105608. URL https://linkinghub.elsevier.com/retrieve/pii/S0169260720314413.

Felix Busch, Jakob Nikolas Kather, Christian Johner, Marina Moser, Daniel Truhn, Lisa C. Adams, and Keno K. Bressem. Navigating the European Union Artificial Intelligence Act for Healthcare. *npj Digital Medicine*, 7(1):1–6, August 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01213-6. URL https://www.nature.com/articles/s41746-024-01213-6. Publisher: Nature Publishing Group.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, December 2020. ISSN 13618415. doi: 10.1016/j.media.2020.101797. URL https://linkinghub.elsevier.com/retrieve/pii/S1361841520301614.

Guillaume Chassagnon, Maria Vakalopoulou, Nikos Paragios, and Marie-Pierre Revel. Artificial intelligence applications for thoracic imaging. *European Journal of Radiology*, 123:108774, February 2020. ISSN 0720048X. doi: 10.1016/j.ejrad.2019.108774. URL https://linkinghub.elsevier.com/retrieve/pii/S0720048X19304243.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, March 2018. doi: 10.1109/WACV.2018.00097. URL https://ieeexplore.ieee.org/document/8354201.

Raheel Chaudhry, Adekunle E. Omole, and Bruno Bordoni. Anatomy, Thorax, Lungs. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2017. URL http://www.ncbi.nlm.nih.gov/books/NBK470197/.

Bingzhi Chen, Jinxing Li, Xiaobao Guo, and Guangming Lu. DualCheXNet: dual asymmetric feature learning for thoracic disease classification in chest X-rays. *Biomedical Signal Processing and Control*, 53:101554, August 2019a. ISSN 1746-8094. doi: 10.1016/j.bspc.2019.04.031. URL https://www.sciencedirect.com/science/article/pii/S1746809419301260.

Bingzhi Chen, Jinxing Li, Guangming Lu, Hongbing Yu, and David Zhang. Label Co-Occurrence Learning With Graph Convolutional Networks for Multi-Label Chest X-Ray Image Classification. *IEEE Journal of Biomedical and Health Informatics*, 24(8):2292–2302, August 2020. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2020.2967084. URL https://ieeexplore.ieee.org/document/8961143/.

Chen Chen, Haobo Wang, Weiwei Liu, Xingyuan Zhao, Tianlei Hu, and Gang Chen. Two-Stage Label Embedding via Neural Factorization Machine for Multi-Label Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3304–3311, July 2019b. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.33013304. URL https://ojs.aaai.org/index.php/AAAI/article/view/4203.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, July 2021. ISSN

2522-5839. doi: 10.1038/s42256-021-00338-7. URL http://www.nature.com/articles/s42256-021-00338-7.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, March 2016. ISSN 1067-5027, 1527-974X. doi: 10.1093/jamia/ocv080. URL https://academic.oup.com/jamia/article-lookup/doi/10.1093/jamia/ocv080.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. page 8, August 2009. doi: 10.1109/CVPR.2009.5206848.

Xin Deng, Enpeng Liu, Shengxi Li, Yiping Duan, and Mai Xu. Interpretable Multi-Modal Image Registration Network Based on Disentangled Convolutional Sparse Coding. *IEEE Transactions on Image Processing*, 32:1078–1091, 2023. ISSN 1941-0042. doi: 10.1109/TIP.2023.3240024. Conference Name: IEEE Transactions on Image Processing.

Shivang Desai, Ahmad Baghal, Thidathip Wongsurawat, Shaymaa Al-Shukri, Kim Gates, Phillip Farmer, Michael Rutherford, Geri D. Blake, Tracy Nolan, Thomas Powell, Kevin Sexton, William Bennett, and Fred Prior. Chest Imaging with Clinical and Genomic Correlates Representing a Rural COVID-19 Positive Population, 2020. URL https://www.cancerimagingarchive.net/collection/covid-19-ar/.

Tribikram Dhar, Nilanjan Dey, Surekha Borra, and R. Simon Sherratt. Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust. *IEEE Transactions on Technology and Society*, 4(1):68–75, March 2023. ISSN 2637-6415. doi: 10.1109/TTS.2023.3234203. URL https://ieeexplore.ieee.org/document/10005626/?arnumber=10005626. Conference Name: IEEE Transactions on Technology and Society.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608 [cs, stat]*, 2017. URL https://arxiv.org/abs/1702.08608.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929.

Rachel Lea Draelos and Lawrence Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks, November 2021. URL http://arxiv.org/abs/2011.08891. arXiv:2011.08891 [cs, eess].

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

Carmen Esposito, Gregory A. Landrum, Nadine Schneider, Nikolaus Stiefl, and Sereina Riniker. GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning. *Journal of Chemical Information and Modeling*, 61(6):2623–2640, June 2021. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c00160. URL https://doi.org/10.1021/acs.jcim.1c00160. Publisher: American Chemical Society.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, February 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature21056. URL http://www.nature.com/articles/nature21056.

Yeli Feng, Hui Seong Teh, and Yiyu Cai. Deep Learning for Chest Radiology: A Review. *Current Radiology Reports*, 7(8):24, August 2019. ISSN 2167-4825. doi: 10.1007/s40134-019-0333-9. URL http://link.springer.com/10.1007/s40134-019-0333-9.

Keigo Furutani, Yasushi Hirano, and Shoji Kido. Segmentation of lung region from chest x-ray images using U-net. In *International Forum on Medical Imaging in Asia 2019*, volume 11050, pages 165–169. SPIE, March 2019. doi: 10.1117/12.2521594. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11050/1105010/Segmentation-of-lung-region-from-chest-x-ray-images-using/10.1117/12.2521594.full.

Christian Garbin, Pranav Rajpurkar, Jeremy Irvin, Matthew P. Lungren, and Oge Marques. Structured dataset documentation: a datasheet for CheXpert, May 2021. URL http://arxiv.org/abs/2105.03020. Number: arXiv:2105.03020 arXiv:2105.03020 [cs, eess].

Beatriz Garcia Santa Cruz, Matías Nicolás Bossa, Jan Sölter, and Andreas Dominik Husch. Public Covid-19 X-ray datasets and their impact on model bias – A systematic review of a significant problem. *Medical Image Analysis*, 74:102225, December 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102225. URL https://www.sciencedirect.com/science/article/pii/S136184152100270X.

J Raymond Geis, Adrian P Brady, Carol C Wu, Jack Spencer, Erik Ranschaert, Jacob L Jaremko, Steve G Langer, Andrea Borondy Kitts, Judy Birch, William F Shields, and others. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *Canadian Association of Radiologists Journal*, 70(4):329–334, 2019.

Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11):1544–1547, November 2018. ISSN 2168-6106. doi: 10.1001/jamainternmed.2018.3763. URL https://doi.org/10.1001/jamainternmed.2018.3763.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

John K. Gohagan, Philip C. Prorok, Richard B. Hayes, and Barnett-S. Kramer. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, organization, and status. *Controlled Clinical Trials*, 21(6):251S–272S, December 2000. ISSN 01972456. doi: 10.1016/S0197-2456(00)00097-0. URL https://linkinghub.elsevier.com/retrieve/pii/S0197245600000970.

Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Qingji Guan and Yaping Huang. Multi-label chest X-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters*, 130:259–266, February 2020. ISSN

01678655. doi: 10.1016/j.patrec.2018.10.027. URL https://linkinghub.elsevier.com/retrieve/pii/S0167865518308559.

Shuxuan Guo, Li Liu, Wei Wang, Songyang Lao, and Liang Wang. An attention model based on spatial transformers for scene recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3757–3762, December 2016. doi: 10.1109/ICPR.2016.7900219.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. pages 770–778, 2016. URL http://arxiv.org/abs/1512.03385.

Lukas-Valentin Herm, Kai Heinrich, Jonas Wanner, and Christian Janiesch. Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management*, 69:102538, April 2023. ISSN 0268-4012. doi: 10.1016/j.ijinfomgt.2022.102538. URL https://www.sciencedirect.com/science/article/pii/S026840122200072X.

Steven A. Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence, April 2021. URL https://www.medrxiv.org/content/10.1101/2021.04.07.21254975v2. Pages: 2021.04.07.21254975.

Steven A. Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1):5979, April 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-09954-8. URL https://www.nature.com/articles/s41598-022-09954-8. Publisher: Nature Publishing Group.

Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4), July 2019. ISSN 1942-4787, 1942-4795. doi: 10.1002/widm.1312.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.243. URL https://ieeexplore.ieee.org/document/8099726/.

Tianxi Huang, Dongdong Wu, Guiduo Duan, and Hao Huang. Multi-label Image Classification Model via Label Correlation Matrix. *Journal of Physics: Conference Series*, 2216(1):012107, March 2022. ISSN 1742-6596. doi: 10.1088/1742-6596/2216/1/012107. URL https://dx.doi.org/10.1088/1742-6596/2216/1/012107. Publisher: IOP Publishing.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597, July 2019. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v33i01.3301590. URL https://aaai.org/ojs/index.php/AAAI/article/view/3834.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/hash/33ceb07bf4eeb3da587e268d663aba1a-Abstract.html.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J. Wáng, Pu-Xuan Lu, and George Thoma. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6):475–477, December 2014. ISSN 2223-4292. doi: 10.3978/j.issn.2223-4292.2014.11.20.

Ebenezer Jangam and Chandra Sekhara Rao Annavarapu. A stacked ensemble for the detection of COVID-19 with high recall and accuracy. *Computers in Biology and Medicine*, 135:104608, August 2021. ISSN 00104825. doi: 10.1016/j.compbiomed.2021.104608. URL https://linkinghub.elsevier.com/retrieve/pii/S0010482521004029.

Jaehyup Jeong, Bosoung Jeoun, Yeonju Park, and Bohyung Han. An Optimized Ensemble Framework for Multi-Label Classification on Long-Tailed Chest X-ray Data. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2731–2738, Paris, France, October 2023. IEEE. ISBN 9798350307443. doi: 10.1109/ICCVW60793.2023.00289. URL https://ieeexplore.ieee.org/document/10350895/.

Xun Jia, Lei Ren, and Jing Cai. Clinical implementation of AI technologies will require interpretable AI models. *Medical Physics*, 47(1):1–4, January 2020. ISSN 0094-2405, 2473-4209. doi: 10.1002/mp.13891. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13891.

Xiaoben Jiang, Yu Zhu, Yatong Liu, Gan Cai, and Hao Fang. TransDD: A transformer-based dual-path decoder for improving the performance of thoracic diseases classification using chest X-ray. *Biomedical Signal Processing and Control*, 91:105937, May 2024. ISSN 17468094. doi: 10.1016/j.bspc.2023.105937. URL https://linkinghub.elsevier.com/retrieve/pii/S1746809423013708.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, December 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. URL http://www.nature.com/articles/s41597-019-0322-0.

José Raniery Ferreira Junior, Diego Armando Cardona Cardenas, Ramon Alfredo Moreno, Marina de Fátima de Sá Rebelo, José Eduardo Krieger, and Marco Antonio Gutierrez. A general fully automated deep-learning method to detect cardiomegaly in chest x-rays. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, pages 537–542. SPIE, February 2021. doi: 10.1117/12.2581980. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11597/115972B/A-general-fully-automated-deep-learning-method-to-detect-cardiomegaly/10.1117/12.2581980.full.

Anwesh Kabiraj, Tanushree Meena, Kalyan Tadepalli, and Sudipta Roy. An explainable weakly supervised model for multi-disease detection and localization from thoracic X-rays. *Applied Soft Computing*, 166:112139, November 2024. ISSN 1568-4946. doi: 10.1016/j.asoc.2024.112139. URL https://www.sciencedirect.com/science/article/pii/S156849462400913X.

K. Kallianos, J. Mongan, S. Antani, T. Henry, A. Taylor, J. Abuya, and M. Kohli. How far have we come? Artificial intelligence for chest radiograph interpretation. *Clinical Radiology*, 74 (5):338–345, May 2019. ISSN 00099260. doi: 10.1016/j.crad.2018.12.015. URL https://linkinghub.elsevier.com/retrieve/pii/S0009926019300194.

Abdulkadir Karacı. VGGCOV19-NET: automatic detection of COVID-19 cases from X-ray images using modified VGG19 CNN architecture and YOLO algorithm. *Neural Computing and Applications*, 34(10):8253–8274, May 2022. ISSN 0941-0643, 1433-3058. doi: 10.1007/s00521-022-06918-x. URL https://link.springer.com/10.1007/s00521-022-06918-x.

Md Rezaul Karim, Till Döhmen, Dietrich Rebholz-Schuhmann, Stefan Decker, Michael Cochez, and Oya Beyan. DeepCOVIDExplainer: Explainable COVID-19 Diagnosis Based on Chest X-ray Images. *arXiv:2004.04582 [cs, eess]*, June 2020. URL http://arxiv.org/abs/2004.04582.

K. Karthik and S. Sowmya Kamath. MSDNet: a deep neural ensemble model for abnormality detection and classification of plain radiographs. *Journal of Ambient Intelligence and Humanized Computing*, April 2022. ISSN 1868-5145. doi: 10.1007/s12652-022-03835-8. URL https://doi.org/10.1007/s12652-022-03835-8.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y.L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1131.e9, February 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.02.010. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867418301545.

Bahador Khaleghi. The How of Explainable AI: Post-modelling Explainability, August 2019. URL https://towardsdatascience.com/the-how-of-explainable-ai-post-modelling-explainability-8b4cbc7adf5f. Publication Title: Medium.

Euyoung Kim, Soochahn Lee, and Kyoung Mu Lee. Abnormality Detection in Chest X-Ray via Residual-Saliency From Normal Generation. *IEEE Access*, 11:21799–21810, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3251350. Conference Name: IEEE Access.

Shiva prasad Koyyada and Thipendra P. Singh. An explainable artificial intelligence model for identifying local indicators and detecting lung disease from chest X-ray images. *Healthcare Analytics*, 4:100206, December 2023. ISSN 2772-4425. doi: 10.1016/j.health.2023.100206. URL https://www.sciencedirect.com/science/article/pii/S2772442523000734.

Aurimas Kudzinskas and Avery L. Callahan. Anatomy, Thorax. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023. URL http://www.ncbi.nlm.nih.gov/books/NBK557710/.

Pulkit Kumar, Monika Grewal, and Muktabh Mayank Srivastava. Boosted Cascaded Convnets for Multilabel Classification of Thoracic Diseases in Chest Radiographs. In Aurélio

Campilho, Fakhri Karray, and Bart ter Haar Romeny, editors, *Image Analysis and Recognition*, volume 10882, pages 546–552. Springer International Publishing, Cham, 2018. ISBN 978-3-319-92999-6 978-3-319-93000-8. doi: 10.1007/978-3-319-93000-8_62. URL http://link.springer.com/10.1007/978-3-319-93000-8_62.

June-Goo Lee, Sanghoon Jun, Young-Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep Learning in Medical Imaging: General Overview. *Korean Journal of Radiology*, 18(4):570, 2017. ISSN 1229-6929, 2005-8330. doi: 10.3348/kjr.2017.18.4.570. URL https://www.kjronline.org/DOIx.php?id=10.3348/kjr.2017.18.4.570.

Matthew C. H. Lee, Ozan Oktay, Andreas Schuh, Michiel Schaap, and Ben Glocker. Image-and-Spatial Transformer Networks for Structure-Guided Image Registration. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 337–345, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32245-8. doi: 10.1007/978-3-030-32245-8_38.

Miaoge Li, Dongsheng Wang, Xinyang Liu, Zequn Zeng, Ruiying Lu, Bo Chen, and Mingyuan Zhou. PatchCT: Aligning Patch Set and Label Set with Conditional Transport for Multi-Label Image Classification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15302–15312, Paris, France, October 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.01408. URL https://ieeexplore.ieee.org/document/10378262/.

Xiuyuan Li, Fengchao Li, Jiangang Yu, and Guowen An. A high-precision underwater object detection based on joint self-supervised deblurring and improved spatial transformer network, March 2022. URL http://arxiv.org/abs/2203.04822. arXiv:2203.04822 [cs].

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1):18, January 2021. ISSN 1099-4300. doi: 10.3390/e23010018. URL https://www.mdpi.com/1099-4300/23/1/18. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Zachary C. Lipton. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, June 2018. ISSN 1542-7730. doi: 10.1145/3236386.3241340. URL https://dl.acm.org/doi/10.1145/3236386.3241340.

P. J. G. Lisboa, S. Saralajew, A. Vellido, R. Fernández-Domenech, and T. Villmann. The coming of age of interpretable and explainable machine learning models. *Neurocomputing*, 535:25–39, May 2023. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.02.040. URL https://www.sciencedirect.com/science/article/pii/S0925231223001893.

Dichao Liu, Yu Wang, and Jien Kato. Supervised Spatial Transformer Networks for Attention Learning in Fine-grained Action Recognition. In *VISIGRAPP (4: VISAPP)*, pages 311–318, 2019a.

Han Liu, Lei Wang, Yandong Nan, Faguang Jin, Qi Wang, and Jiantao Pu. SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. *Computerized Medical Imaging and Graphics*, 75:66–73, July 2019b. ISSN 0895-6111. doi: 10.1016/j.compmedimag.2019.05.005. URL https://www.sciencedirect.com/science/article/pii/S0895611118306177.

Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, and Yizhou Yu. Align, Attend and Locate: Chest X-Ray Diagnosis via Contrast Induced Attention Network With Limited Supervision. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10631–10640, Seoul, Korea (South), October 2019c. IEEE. ISBN 978-1-72814-803-8. doi: 10.1109/ICCV.2019.01073. URL https://ieeexplore.ieee.org/document/9010639/.

Helena Liz, Javier Huertas-Tato, Manuel Sánchez-Montañés, Javier Del Ser, and David Camacho. Deep learning for understanding multilabel imbalanced Chest X-ray datasets. *Future Generation Computer Systems*, 144:291–306, July 2023. ISSN 0167-739X. doi: 10.1016/j.future.2023.03.005. URL https://www.sciencedirect.com/science/article/pii/S0167739X23000821.

Alex John London. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, 49(1):15–21, January 2019. ISSN 00930334. doi: 10.1002/hast.973. URL http://doi.wiley.com/10.1002/hast.973.

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Shreya Lunkad. Break open the Black Box: Explainable AI frameworks to increase transparency, December 2023. URL https://medium.com/@shreyalunkad/break-open-the-black-box-explainable-ai-frameworks-to-increase-transparency-a07b171ee10e.

Navid Mahabadi, Alberto A. Goizueta, and Bruno Bordoni. Anatomy, Thorax, Lung Pleura And Mediastinum. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2018. URL http://www.ncbi.nlm.nih.gov/books/NBK519048/.

Yifan Mao, Fei-Fei Xue, Ruixuan Wang, Jianguo Zhang, Wei-Shi Zheng, and Hongmei Liu. Abnormality Detection in Chest X-Ray Images Using Uncertainty Prediction Autoencoders. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 529–538, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59725-2. doi: 10.1007/978-3-030-59725-2_51.

Jose Antonio Marmolejo-Saucedo and Utku Kose. Numerical Grad-Cam Based Explainable Convolutional Neural Network for Brain Tumor Diagnosis. *Mobile Networks and Applications*, August 2022. ISSN 1572-8153. doi: 10.1007/s11036-022-02021-6. URL https://doi.org/10.1007/s11036-022-02021-6.

Morgan P. McBee, Omer A. Awan, Andrew T. Colucci, Comeron W. Ghobadi, Nadja Kadom, Akash P. Kansagra, Srini Tridandapani, and William F. Auffermann. Deep Learning in Radiology. *Academic Radiology*, 25(11):1472–1480, November 2018. ISSN 1076-6332. doi:

10.1016/j.acra.2018.02.018. URL https://www.sciencedirect.com/science/article/pii/S1076633218301041.

Akhil Meethal, Marco Pedersoli, Soufiane Belharbi, and Eric Granger. Convolutional STN for Weakly Supervised Object Localization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10157–10164, January 2021. doi: 10.1109/ICPR48806.2021.9412029. ISSN: 1051-4651.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, February 2018. ISSN 1051-2004. doi: 10.1016/j.dsp.2017.10.011. URL https://www.sciencedirect.com/science/article/pii/S1051200417302385.

Mehdi Moradi, Ken C. L. Wong, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Quality controlled segmentation to aid disease detection. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 362–368. SPIE, March 2020. doi: 10.1117/12.2549426. URL https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11314/113141K/Quality-controlled-segmentation-to-aid-disease-detection/10.1117/12.2549426.full.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, October 2019. doi: 10.1073/pnas.1900654116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1900654116. Publisher: Proceedings of the National Academy of Sciences.

Takahiro Nakao, Shouhei Hanaoka, Yukihiro Nomura, Masaki Murata, Tomomi Takenaga, Soichiro Miki, Takeyuki Watadani, Takeharu Yoshikawa, Naoto Hayashi, and Osamu Abe. Unsupervised Deep Anomaly Detection in Chest Radiographs. *Journal of Digital Imaging*, 34(2):418–427, April 2021. ISSN 1618-727X. doi: 10.1007/s10278-020-00413-2. URL https://doi.org/10.1007/s10278-020-00413-2.

Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9(1):429, July 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01498-w. URL https://www.nature.com/articles/s41597-022-01498-w. Publisher: Nature Publishing Group.

Nillmani, Neeraj Sharma, Luca Saba, Narendra N. Khanna, Mannudeep K. Kalra, Mostafa M. Fouda, and Jasjit S. Suri. Segmentation-Based Classification Deep Learning Model Embedded with Explainable AI for COVID-19 Detection in Chest X-ray Scans. *Diagnostics*, 12(9):2132, September 2022. ISSN 2075-4418. doi: 10.3390/diagnostics12092132. URL https://www.mdpi.com/2075-4418/12/9/2132. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

Lauren Oakden-Rayner. Half a million x-rays! First impressions of the Stanford and MIT chest x-ray datasets, February 2019. URL https://laurenoakdenrayner.com/2019/02/25/half-a-million-x-rays-first-impressions-of-the-stanford-and-mit-chest-x-ray-datasets/.

Omer Onder, Yasin Yarasir, Aynur Azizova, Gamze Durhan, Mehmet Ruhi Onur, and Orhan Macit Ariyurek. Errors, discrepancies and underlying bias in radiology with case examples: a pictorial review. *Insights into Imaging*, 12(1):51, April 2021. ISSN 1869-4101. doi: 10.1186/s13244-021-00986-8. URL https://doi.org/10.1186/s13244-021-00986-8.

OpenStax. OpenStax AnatPhys fig.19.2 - Heart Position in Thorax - English labels | AnatomyTOOL, 2024. URL https://anatomytool.org/content/openstax-anatphys-fig192-heart-position-thorax-english-labels.

B Padmaja, M Madhubala, M Nagaraju, Nandhan Varma Somalaraju, Meghana Kovuri, and Krishnaveni Sriramwar. Chest X-Ray Image Analysis for Respiratory Disease Prediction using Grad-CAM. In *2023 2nd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, pages 1–7, February 2023. doi: 10.1109/DELCON57910.2023.10127464.

Daniel H. Pak, Andrés Caballero, Wei Sun, and James S. Duncan. Efficient Aortic Valve Multilabel Segmentation Using a Spatial Transformer Network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1738–1742, April 2020. doi: 10.1109/ISBI45749.2020.9098378. ISSN: 1945-8452.

João Pedrosa, Pedro Sousa, Joana Silva, Ana Maria Mendonça, and Aurélio Campilho. Lesion-Based Chest Radiography Image Retrieval for Explainability in Pathology Detection. In Armando J. Pinho, Petia Georgieva, Luís F. Teixeira, and Joan Andreu Sánchez, editors, *Pattern Recognition and Image Analysis*, pages 81–94, Cham, 2022. Springer International Publishing. ISBN 978-3-031-04881-4. doi: 10.1007/978-3-031-04881-4_7.

Sérgio Pereira, Raphael Meier, Richard McKinley, Roland Wiest, Victor Alves, Carlos A. Silva, and Mauricio Reyes. Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation. *Medical Image Analysis*, 44:228–244, February 2018. ISSN 1361-8415. doi: 10.1016/j.media.2017.12.009. URL https://www.sciencedirect.com/science/article/pii/S1361841517301901.

Ardimas Andi Purwita and Nunung Nurul Qomariyah. Impact of Aligning Saliency Maps on COVID-19 Disease Detection Using Chest X-Ray Images. In *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 396–401, July 2021. doi: 10.1109/ISITIA52817.2021.9502264. URL https://ieeexplore.ieee.org/abstract/document/9502264?casa_token=XcAmnBguHjkAAAAA:MN2Usc2o1yF5zVQwjnR85mzjsxLHsaKhWQSvoZh6Ce8M4kkan7pK5PJG03Jvp84KWLg73mu86w.

Wendi Qu, Indranil Balki, Mauro Mendez, John Valen, Jacob Levman, and Pascal N. Tyrrell. Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging. *International Journal of Computer Assisted Radiology and Surgery*, 15(12):2041–2048, December 2020. ISSN 1861-6429. doi: 10.1007/s11548-020-02260-6. URL https://doi.org/10.1007/s11548-020-02260-6.

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

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *arXiv:1506.02640 [cs]*, May 2016. URL http://arxiv.org/abs/1506.02640.

Ibraheem Rehman and Afzal Rehman. Anatomy, Thorax, Heart. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2017. URL http://www.ncbi.nlm.nih.gov/books/NBK470256/.

Pantelis Rempakos, Michalis Vrigkas, Marina E. Plissiti, and Christophoros Nikou. Spatial Transformer Generative Adversarial Network for Image Super-Resolution. In Gian Luca Foresti, Andrea Fusiello, and Edwin Hancock, editors, *Image Analysis and Processing – ICIAP 2023*, Lecture Notes in Computer Science, pages 399–411, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43148-7. doi: 10.1007/978-3-031-43148-7_34.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A. Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M. Summers, and Roland Wiest. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiology: Artificial Intelligence*, 2(3):e190043, May 2020. ISSN 2638-6100. doi: 10.1148/ryai.2020190043.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, August 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778. URL https://dl.acm.org/doi/10.1145/2939672.2939778.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 (1), April 2018. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v32i1.11491. URL https://ojs.aaai.org/index.php/AAAI/article/view/11491.

Joana Rocha, Sofia Cardoso Pereira, Joao Pedrosa, Aurelio Campilho, and Ana Maria Mendonca. Attention-driven Spatial Transformer Network for Abnormality Detection in Chest X-Ray Images. In *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 252–257, Shenzen, China, July 2022. IEEE. ISBN 978-1-66546-770-4. doi: 10.1109/CBMS55023.2022.00051. URL https://ieeexplore.ieee.org/document/9867115/.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x.

Sungweon Ryoo and Hee Jin Kim. Activities of the Korean Institute of Tuberculosis. *Osong Public Health and Research Perspectives*, 5:S43–S49, December 2014. ISSN 22109099. doi: 10.1016/j.phrp.2014.10.007. URL http://linkinghub.elsevier.com/retrieve/pii/S2210909914001192.

Luca Saba, Mainak Biswas, Venkatanareshbabu Kuppili, Elisa Cuadrado Godia, Harman S. Suri, Damodar Reddy Edla, Tomaž Omerzu, John R. Laird, Narendra N. Khanna, Sophie Mavrogeni, Athanasios Protogerou, Petros P. Sfikakis, Vijay Viswanathan, George D. Kitas, Andrew Nicolaides, Ajay Gupta, and Jasjit S. Suri. The present and future of deep learning in radiology. *European Journal of Radiology*, 114:14–24, May 2019. ISSN 0720-048X. doi:

10.1016/j.ejrad.2019.02.038. URL https://www.sciencedirect.com/science/article/pii/S0720048X19300919.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv:1708.08296 [cs, stat]*, August 2017. URL http://arxiv.org/abs/1708.08296.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven Q. H. Truong, Chanh D. T. Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G. Blankenberg, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. Benchmarking saliency methods for chest X-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, October 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00536-x. URL https://www.nature.com/articles/s42256-022-00536-x. Publisher: Nature Publishing Group.

Sam Sattarzadeh, Mahesh Sudhakar, Konstantinos N. Plataniotis, Jongseong Jang, Yeonjeong Jeong, and Hyunwoo Kim. Integrated Grad-Cam: Sensitivity-Aware Visual Explanation of Deep Convolutional Networks Via Integrated Gradient-Based Scoring. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1775–1779, June 2021. doi: 10.1109/ICASSP39728.2021.9415064. ISSN: 2379-190X.

Alfred Schöttl. Improving the Interpretability of GradCAMs in Deep Classification Networks. *Procedia Computer Science*, 200:620–628, January 2022. ISSN 1877-0509. doi: 10.1016/j.procs.2022.01.260. URL https://www.sciencedirect.com/science/article/pii/S1877050922002691.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, October 2017. doi: 10.1109/ICCV.2017.74. URL https://ieeexplore.ieee.org/document/8237336. ISSN: 2380-7504.

Koosha Sharifani and Mahyar Amini. Machine Learning and Deep Learning: A Review of Methods and Applications, 2023. URL https://papers.ssrn.com/abstract=4458723.

N. J. Shaw, M. Hendry, and O. B. Eden. Inter-Observer Variation in Interpretation of Chest X-Rays. *Scottish Medical Journal*, 35(5):140–141, October 1990. ISSN 0036-9330, 2045-6441. doi: 10.1177/003693309003500505. URL http://journals.sagepub.com/doi/10.1177/003693309003500505.

Ivaxi Sheth and Samira Ebrahimi Kahou. Overcoming Interpretability and Accuracy Trade-off in Medical Imaging. April 2023. URL https://openreview.net/forum?id=BSf6JALJoc.

Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule. *American Journal of Roentgenology*, 174(1):71–74, January 2000. ISSN 0361-803X. doi: 10.2214/ajr.174.1.1740071. URL https://www.ajronline.org/doi/full/10.2214/ajr.174.1.1740071. Publisher: American Roentgen Ray Society.

Nina Shvetsova, Bart Bakker, Irina Fedulova, Heinrich Schulz, and Dmitry V. Dylov. Anomaly Detection in Medical Imaging With Deep Perceptual Autoencoders. *IEEE Access*, 9:118571–118583, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3107163. Conference Name: IEEE Access.

Wilson Silva, Alexander Poellinger, Jaime S. Cardoso, and Mauricio Reyes. Interpretability-Guided Content-Based Medical Image Retrieval. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 305–314, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59710-8. doi: 10.1007/978-3-030-59710-8_30.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, April 2015. URL http://arxiv.org/abs/1409.1556.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. URL http://arxiv.org/abs/1312.6034. arXiv:1312.6034 [cs].

Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging*, 6(6):52, June 2020. ISSN 2313-433X. doi: 10.3390/jimaging6060052. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8321083/.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise, June 2017. URL http://arxiv.org/abs/1706.03825. arXiv:1706.03825 [cs, stat].

P Sprawls Jr. Physical Principles of Medical Imaging, 2nd Ed. *Madison, Wis: Medical Physics*, 1993.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net, April 2015. URL http://arxiv.org/abs/1412.6806. arXiv:1412.6806 [cs].

St Vincent's University Hospital. Left lower lobe pneumonia – lateral CXR. 2024. URL http://www.svuhradiology.ie/case-study/left-lower-lobe-pneumonia-lateral-cxr/.

Martin Stano, Wanda Benesova, and Lukas S Martak. Explainable 3D convolutional neural network using GMM encoding. In *Twelfth International Conference on Machine Vision (ICMV 2019)*, volume 11433, page 114331U. International Society for Optics and Photonics, 2020.

Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5):e1379, 2020. ISSN 1942-4795. doi: 10.1002/widm.1379. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1379. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1379.

Dengdi Sun, Leilei Ma, Zhuanlian Ding, and Bin Luo. An Attention-Driven Multi-label Image Classification with Semantic Embedding and Graph Convolutional Networks. *Cognitive Computation*, 15(4):1308–1319, July 2023a. ISSN 1866-9964. doi: 10.1007/s12559-021-09977-9. URL https://doi.org/10.1007/s12559-021-09977-9.

Xu Sun, Bingzhen Wei, Xuancheng Ren, and Shuming Ma. Label Embedding Network: Learning Label Representation for Soft Training of Deep Networks, October 2017. URL http://arxiv.org/abs/1710.10393. arXiv:1710.10393 [cs].

Zexuan Sun, Linhao Qu, Jiazheng Luo, Zhijian Song, and Manning Wang. Label correlation transformer for automated chest X-ray diagnosis with reliable interpretability. *La radiologia medica*, 128(6):726–733, May 2023b. ISSN 1826-6983. doi: 10.1007/s11547-023-01647-0. URL https://link.springer.com/10.1007/s11547-023-01647-0.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR, July 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html. ISSN: 2640-3498.

Amin Tafti and Doug W. Byerly. X-ray Radiographic Patient Positioning. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2022. URL http://www.ncbi.nlm.nih.gov/books/NBK565865/.

Yu-Xing Tang, You-Bao Tang, Yifan Peng, Ke Yan, Mohammadhadi Bagheri, Bernadette A. Redd, Catherine J. Brandon, Zhiyong Lu, Mei Han, Jing Xiao, and Ronald M. Summers. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digital Medicine*, 3(1):70, December 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-0273-z. URL http://www.nature.com/articles/s41746-020-0273-z.

Lucas O. Teixeira, Rodolfo M. Pereira, Diego Bertolini, Luiz S. Oliveira, Loris Nanni, George D. C. Cavalcanti, and Yandre M. G. Costa. Impact of Lung Segmentation on the Diagnosis and Explanation of COVID-19 in Chest X-ray Images. *Sensors*, 21(21):7116, January 2021. ISSN 1424-8220. doi: 10.3390/s21217116. URL https://www.mdpi.com/1424-8220/21/21/7116. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute.

Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D. Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, Edmund Yick Mun Wong, Charumathi Sabanayagam, Mani Baskaran, Farah Ibrahim, Ngiap Chuan Tan, Eric A. Finkelstein, Ecosse L. Lamoureux, Ian Y. Wong, Neil M. Bressler, Sobha Sivaprasad, Rohit Varma, Jost B. Jonas, Ming Guang He, Ching-Yu Cheng, Gemmy Chui Ming Cheung, Tin Aung, Wynne Hsu, Mong Li Lee, and Tien Yin Wong. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*, 318(22):2211, December 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.18152. URL http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.18152.

Erico Tjoa and Cuntai Guan. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, November 2021. ISSN 2162-2388. doi: 10.1109/tnnls.2020.3027314. URL https://doi.org/10.1109/TNNLS.2020.3027314.

Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, pages 359–380. PMLR, October 2019. URL https://proceedings.mlr.press/v106/tonekaboni19a.html. ISSN: 2640-3498.

Maximilian E. Tschuchnig and Michael Gadermayr. Anomaly Detection in Medical Imaging - A Mini Review. In Peter Haber, Thomas J. Lampoltshammer, Helmut Leopold, and Manfred Mayr, editors, *Data Science – Analytics and Applications*, pages 33–38, Wiesbaden, 2022. Springer Fachmedien. ISBN 978-3-658-36295-9. doi: 10.1007/978-3-658-36295-9_5.

Kudaibergen Urinbayev, Yerassyl Orazbek, Yernur Nurambek, Almas Mirzakhmetov, and Huseyin Atakan Varol. End-to-End Deep Diagnosis of X-ray Images. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2182–2185, July 2020. doi: 10.1109/EMBC44109.2020.9175208. URL https://ieeexplore.ieee.org/abstract/document/9175208. ISSN: 2694-0604.

Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, Marisa Caparrós, Germán González, and Jose María Salinas. BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients, June 2020. URL https://ieee-dataport.org/open-access/bimcv-covid-19-large-annotated-dataset-rx-and-ct-images-covid-19-patients. arXiv:2006.01174 [cs, eess].

Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, December 2021. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.05.009. URL https://www.sciencedirect.com/science/article/pii/S1566253521001093.

Haidong Wang, Mohsen Naghavi, Christine Allen, Ryan M Barber, and Zulfiqar Bhutta et. al. Global, regional,and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death,1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *388*, pages 1459–1544, 2016. ISSN 01406736. URL https://linkinghub.elsevier.com/retrieve/pii/S0140673616310121. Place: The Lancet.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. pages 24–25, 2020a. URL https://openaccess.thecvf.com/content_CVPRW_2020/html/w1/Wang_Score-CAM_Score-Weighted_Visual_Explanations_for_Convolutional_Neural_Networks_CVPRW_2020_paper.html.

Hongyu Wang, Shanshan Wang, Zibo Qin, Yanning Zhang, Ruijiang Li, and Yong Xia. Triple attention learning for classification of 14 thoracic diseases using chest radiography. *Medical Image Analysis*, 67:101846, January 2021. ISSN 13618415. doi: 10.1016/j.media.2020.101846. URL https://linkinghub.elsevier.com/retrieve/pii/S1361841520302103.

Linda Wang, Zhong Qiu Lin, and Alexander Wong. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports*, 10(1):19549, November 2020b. ISSN 2045-2322. doi: 10.1038/s41598-020-76550-z. URL https://www.nature.com/articles/s41598-020-76550-z. Publisher: Nature Publishing Group.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, Honolulu, HI, July 2017a. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.369. URL http://ieeexplore.ieee.org/document/8099852/.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. NIH Chest X-rays. https://www.kaggle.com/nih-chest-xrays/data, 2017b. Publication Title: https://www.kaggle.com/nih-chest-xrays/data.

World Health Organization. Pneumonia in children, 2022. URL https://www.who.int/news-room/fact-sheets/detail/pneumonia.

World Health Organization. Global cancer burden growing, amidst mounting need for services. *Press Release no. 345*, 2024. URL https://www.iarc.who.int/wp-content/uploads/2024/02/pr345_E.pdf.

Xin Wu, Yue Feng, Hong Xu, Zhuosheng Lin, Shengke Li, Shihan Qiu, QiChao Liu, and Yuangang Ma. CheXNet: Combing Transformer and CNN for Thorax Disease Diagnosis from Chest X-ray Images. In Qingshan Liu, Hanzi Wang, Zhanyu Ma, Weishi Zheng, Hongbin Zha, Xilin Chen, Liang Wang, and Rongrong Ji, editors, *Pattern Recognition and Computer Vision*, volume 14437, pages 73–84. Springer Nature Singapore, Singapore, 2024. ISBN 978-981-9985-57-9 978-981-9985-58-6. doi: 10.1007/978-981-99-8558-6_7. URL https://link.springer.com/10.1007/978-981-99-8558-6_7. Series Title: Lecture Notes in Computer Science.

Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA, April 2020. ACM. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376807. URL https://dl.acm.org/doi/10.1145/3313831.3376807.

Shimin Xiong, Bin Li, and Shiao Zhu. DCGNN: a single-stage 3D object detection network based on density clustering and graph neural network. *Complex & Intelligent Systems*, 9(3):3399–3408, June 2023. ISSN 2198-6053. doi: 10.1007/s40747-022-00926-z. URL https://doi.org/10.1007/s40747-022-00926-z.

Shuaijing Xu, Xiaoyilei Yang, Junqi Guo, Hao Wu, Guangzhi Zhang, and Rongfang Bie. Cxnet-M3: A Deep Quintuplet Network for Multi-Lesion Classification in Chest X-Ray Images Via Multi-Label Supervision. *IEEE Access*, 8:98693–98704, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2996217. URL https://ieeexplore.ieee.org/abstract/document/9097871. Conference Name: IEEE Access.

M J Yaffe and J A Rowlands. X-ray detectors for digital radiography. *Physics in Medicine and Biology*, 42(1):1–39, January 1997. ISSN 0031-9155, 1361-6560. doi: 10.1088/0031-9155/42/1/001. URL https://iopscience.iop.org/article/10.1088/0031-9155/42/1/001.

Toshinori Yamauchi and Masayoshi Ishikawa. Spatial Sensitive GRAD-CAM: Visual Explanations for Object Detection by Incorporating Spatial Sensitivity. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 256–260, October 2022. doi: 10.1109/ICIP46576.2022.9897350. ISSN: 2381-8549.

Chaochao Yan, Jiawen Yao, Ruoyu Li, Zheng Xu, and Junzhou Huang. Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-rays. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 103–110, Washington DC USA, August 2018. ACM. ISBN 978-1-4503-5794-4. doi: 10.1145/3233547.3233573. URL https://dl.acm.org/doi/10.1145/3233547.3233573.

Howard Yang. Towards a comparable metric for AI model interpretability — Part 2, July 2022. URL https://medium.com/dsaid-govtech/towards-a-comparable-metric-for-ai-model-interpretability-part-2-423e4fc2b232.

Koichiro Yasaka and Osamu Abe. Deep learning and artificial intelligence in radiology: Current applications and future directions. *PLOS Medicine*, 15(11):e1002707, November 2018. ISSN 1549-1676. doi: 10.1371/journal.pmed.1002707. URL https://dx.plos.org/10.1371/journal.pmed.1002707.

E.J. Yates, L.C. Yates, and H. Harvey. Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. *Clinical Radiology*, 73(9):827–831, September 2018. ISSN 00099260. doi: 10.1016/j.crad.2018.05.015. URL https://linkinghub.elsevier.com/retrieve/pii/S000992601830206X.

Mei-Chen Yeh and Yi-Nan Li. Multilabel Deep Visual-Semantic Embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(6):1530–1536, June 2020. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2019.2911065. URL https://ieeexplore.ieee.org/document/8691414/.

Aina R Zehnder, Anna J Pedrosa Carrasco, and Simon N Etkind. Factors associated with hospitalisations of patients with chronic heart failure approaching the end of life: A systematic review. *Palliative Medicine*, 36(10):1452–1468, December 2022. ISSN 0269-2163. doi: 10.1177/02692163221123422. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9749018/.

Cheng Zhang, Francine Chen, and Yan-Ying Chen. Thoracic Disease Identification and Localization using Distance Learning and Region Verification. *arXiv:2006.04203 [cs]*, August 2020. URL http://arxiv.org/abs/2006.04203.

Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, 40(6):e13244, 2023a. ISSN 1468-0394. doi: 10.1111/exsy.13244. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13244. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.13244.

Dandan Zhang, Hongyu Wang, Jiahui Deng, Tonghui Wang, Cong Shen, and Jun Feng. CAMS-Net: An attention-guided feature selection network for rib segmentation in chest X-rays. *Computers in Biology and Medicine*, 156:106702, April 2023b. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2023.106702. URL https://www.sciencedirect.com/science/article/pii/S0010482523001671.

Kai Zhang, Wei Liang, Peng Cao, Xiaoli Liu, Jinzhu Yang, and Osmar Zaiane. Label correlation guided discriminative label feature learning for multi-label chest image classification. *Computer Methods and Programs in Biomedicine*, 245:108032, March 2024. ISSN 01692607. doi: 10.1016/j.cmpb.2024.108032. URL https://linkinghub.elsevier.com/retrieve/pii/S0169260724000282.

Yunyan Zhang, Daphne Hong, Daniel McClement, Olayinka Oladosu, Glen Pridham, and Garth Slaney. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods*, 353:109098, April 2021. ISSN 0165-0270. doi: 10.1016/j.jneumeth.2021.109098. URL https://www.sciencedirect.com/science/article/pii/S0165027021000339.

Qingda Zhao, Fuquan Zhang, and Linjuan Ma. Lung Detection Based on YOLOv3. In Jeng-Shyang Pan, Valentina Emilia Balas, and Chien-Ming Chen, editors, *Advances in Intelligent Data Analysis and Applications*, Smart Innovation, Systems and Technologies, pages 111–116, Singapore, 2022. Springer. ISBN 9789811650369. doi: 10.1007/978-981-16-5036-9_12.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.319. URL http://ieeexplore.ieee.org/document/7780688/.

Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G. van Leeuwen, and Keelin Murphy. Deep learning for chest X-ray analysis: A survey. *Medical Image Analysis*, 72:102125, August 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102125. URL https://www.sciencedirect.com/science/article/pii/S1361841521001717.

# Appendix A

# Evaluation Metrics

The following metrics were used across the multiple studies in this PhD thesis to evaluate the performance of the proposed methods. They compare the probability score, which indicate the likelihood of a sample belonging to a specific class, against the corresponding GT annotation.

The author has adopted a binary setting, where each instance is classified as either belonging to a class (positive) or not (negative). This decision is determined by applying a probability threshold for each class. As a result, a model yields True Positive (TP) or True Negative (TN) results when the decision is correct, and False Positive (FP) or False Negative (FN) results when it is not. Accuracy measures the percentage of correct classifications, precision measures the proportion of TPs out of all instances predicted as positive, while recall (or sensitivity) measures the proportion of TPs out of all actual positive instances. The F1-score is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives. These metrics are formulated next:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \tag{A.1}$$

$$\text{Precision} = TP/(TP+FP) \tag{A.2}$$

$$\text{Recall} = TP/(TP+FN) \tag{A.3}$$

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \tag{A.4}$$

Drawing a Receiver Operating Characteristic (ROC) curve implies plotting the TP rate on the vertical axis and the FP rate on the horizontal axis, using continuous threshold values on the predicted probability scores (Figure A.1). The curve shows the trade-off between the recall and specificity ($1 - \text{FP rate}$), with the Area Under the ROC Curve (AUC) value indicating overall model performance. A perfect model has an AUC of 1, while a random model yields an AUC of 0.5.

Additionally, one calculates the precision and recall at various probability thresholds to plot a Precision-Recall (PR) curve. Precision is plotted on the vertical axis and recall on the horizontal axis. This curve illustrates the trade-off between precision and recall, with the area under the curve

(PR-AUC) providing insight into the model's performance, especially on imbalanced datasets. A higher PR-AUC indicates better performance.

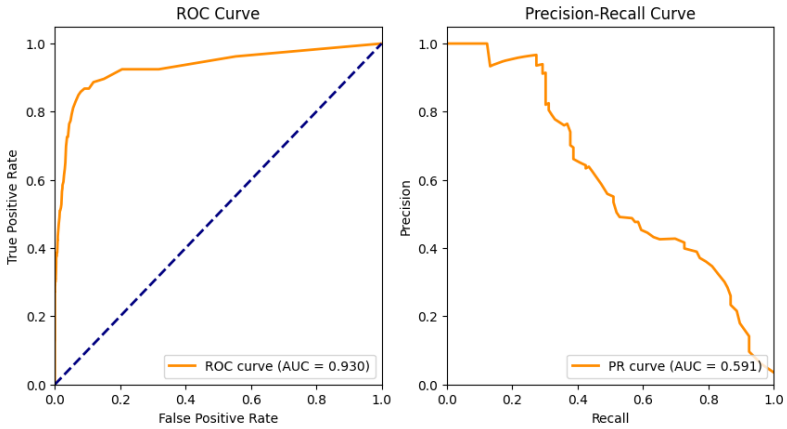



Figure A.1: Examples of ROC and PR curves, respectively.

# Appendix B

# Supplementary Material of CLARE-XR

**Rule 1** (Top-1 approach)

**for** each inference instance **do**
    Get the closest reference coordinate.
    Attribute that top-1 reference's label to the current instance, as a final decision.
**end for**

**Rule 2** (Top-3 approach)

**for** each inference instance **do**
    Get the 3 closest reference coordinates.
    **if** *No Finding* appears in 1st position **then**
        Attribute the *No Finding* label to the current instance, as a final decision.
    **else**
        **if** *No Finding* appears in 2nd or 3rd position **then**
            Delete the *No Finding* label.
            Add the 4th closest reference coordinate to the top 3.
        **end if**
        Attribute all class labels present in the top 3 to the current instance, as a final decision.
    **end if**
**end for**

**Rule 3** (Top-3 approach)

**for** each inference instance **do**
    Get the 3 closest reference coordinates.
    **if** *No Finding* appears in 1st position **then**
        Attribute the *No Finding* label to the current instance, as a final decision.
    **else**
        **if** *No Finding* appears in 2nd or 3rd position **then**
            Delete the *No Finding* label.
            Add the 4th closest reference coordinate to the top 3.
        **end if**
    **end if**
    Attribute the class labels present in at least 2 of the top-3 closest references to the current instance, as a final decision.
**end for**

**Rule 4** (Top-3 approach)

**for** each inference instance **do**
    Get the 3 closest reference coordinates.
    **if** *No Finding* appears in 1st position **then**
        Attribute the *No Finding* label to the current instance, as a final decision.
    **else**
        **if** *No Finding* appears in 2nd or 3rd position **then**
            Delete the *No Finding* label.
            Add the 4th closest reference coordinate to the top 3.
        **end if**
        Attribute the class labels present in the top 1, plus the ones present in at least 2 of the top-3 closest references to the current instance, as a final decision.
    **end if**
**end for**

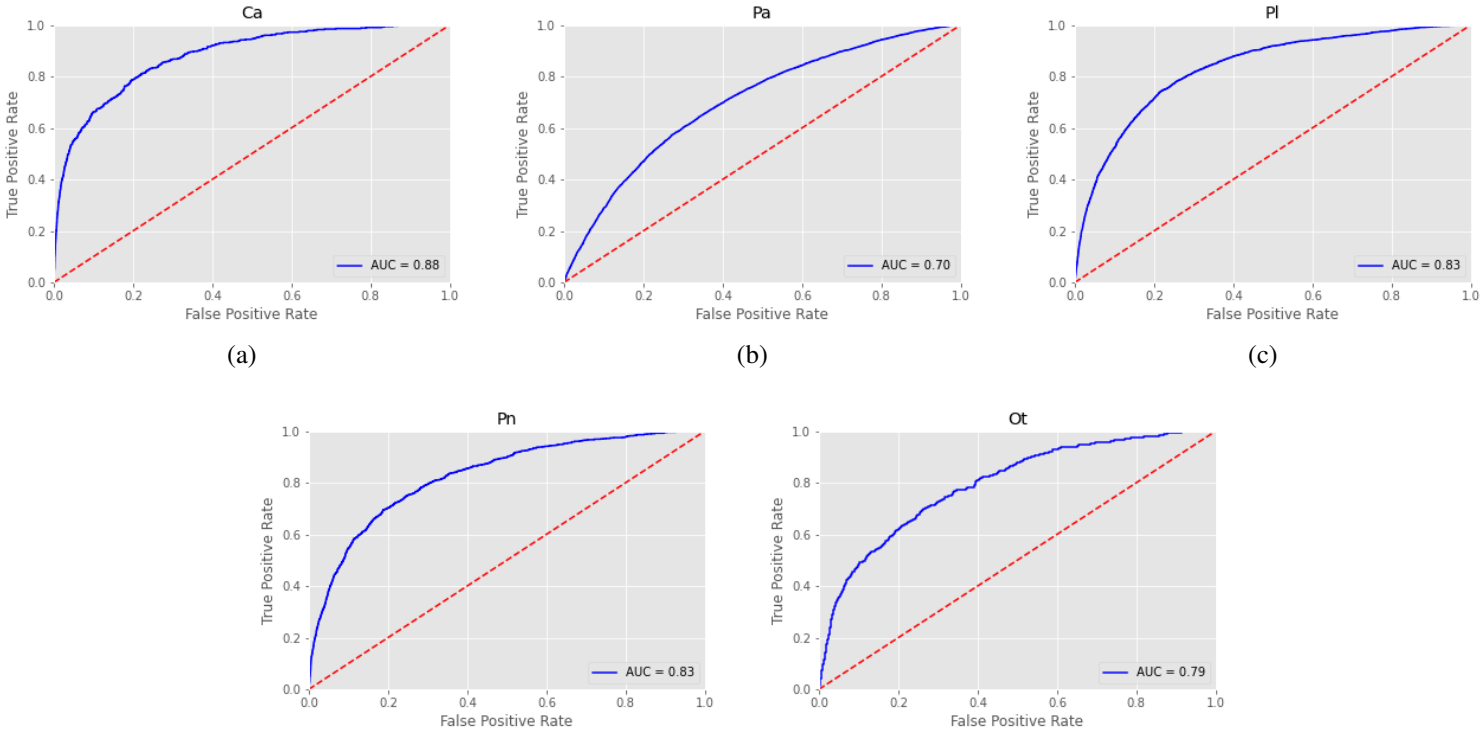



Figure B.1: ROC curves of the multi-label baseline classifier.

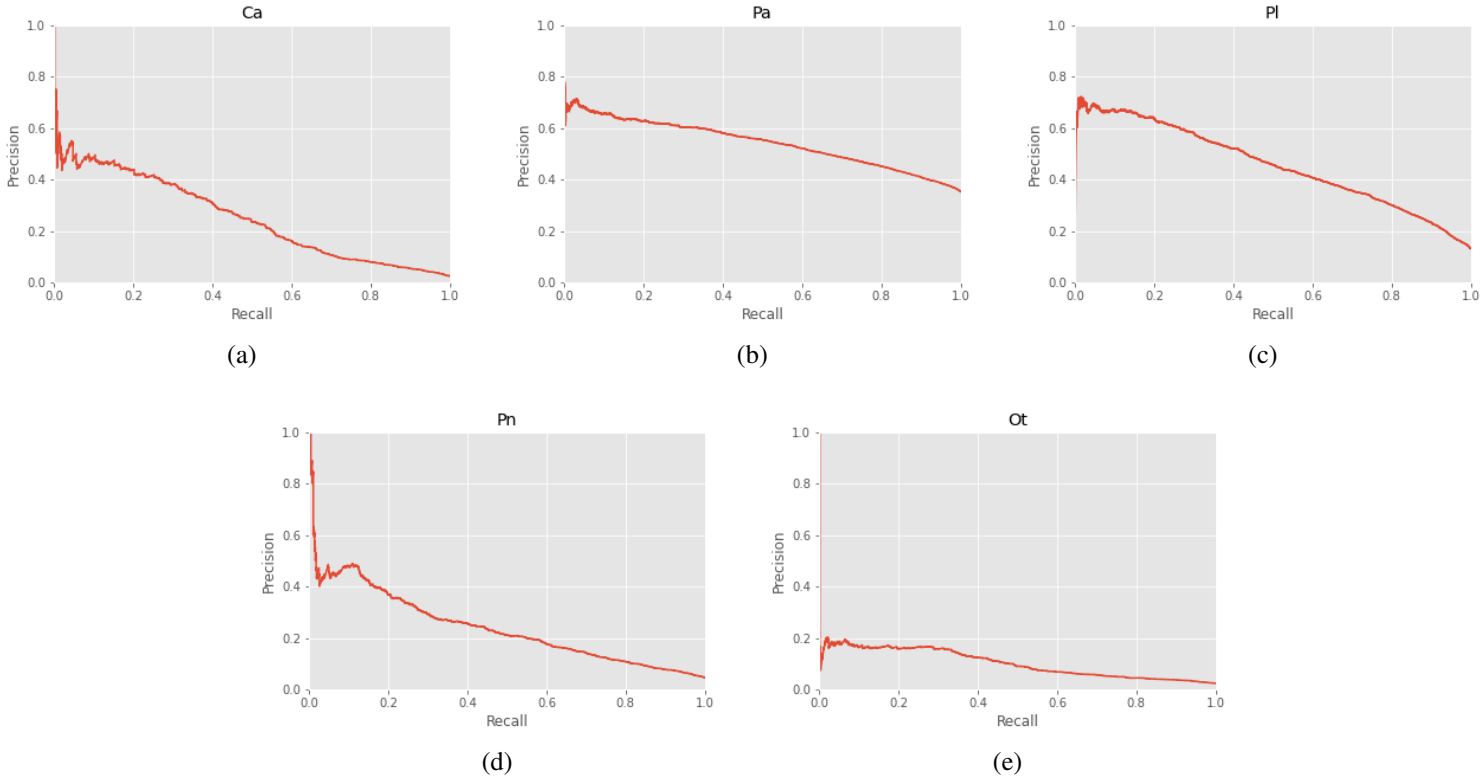



Figure B.2: PR curves of the multi-label baseline classifier.

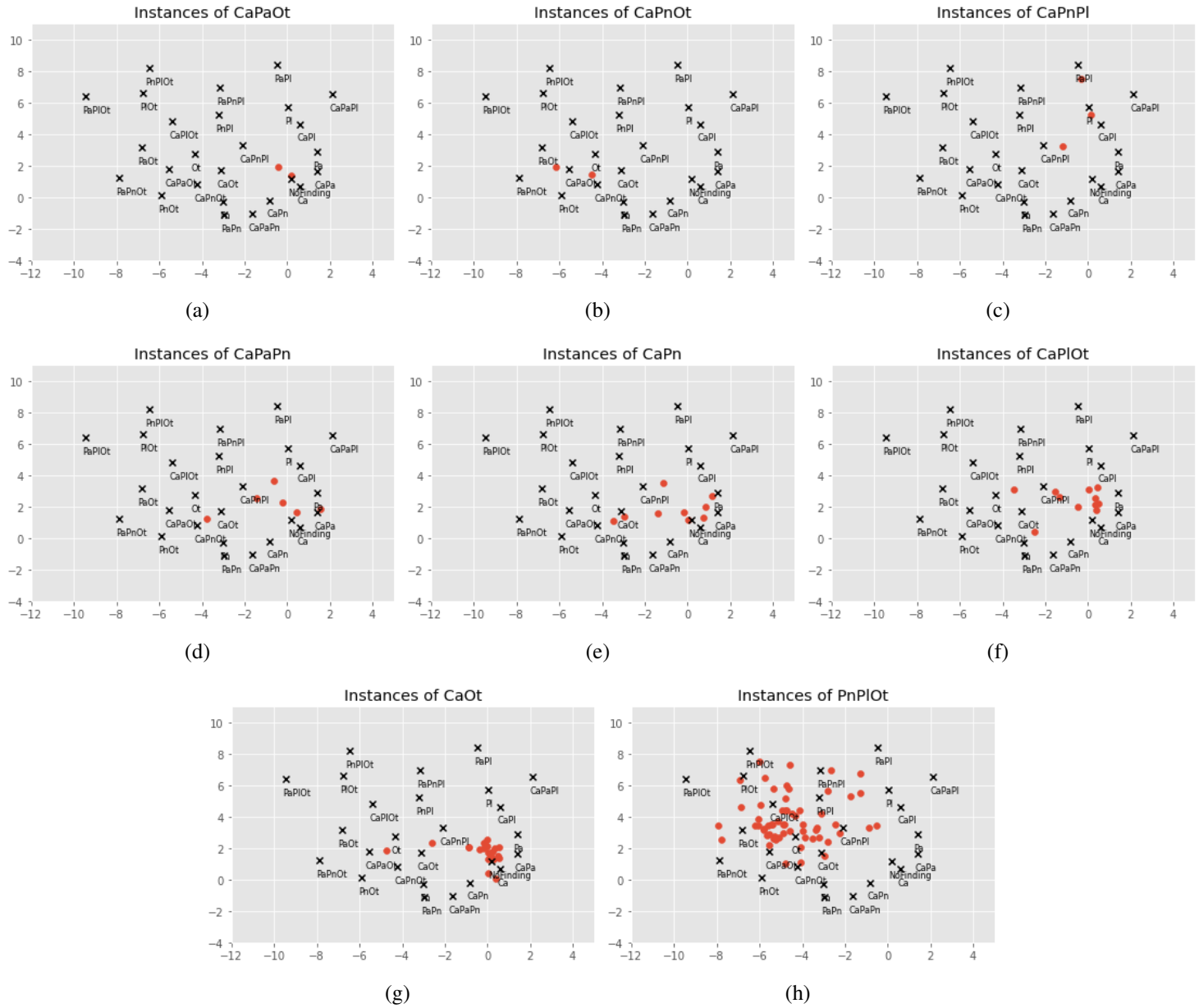



Figure B.3: Predicted embedding coordinates obtained by the Phase II regressor for under-represented combinations. The GT references are marked with black crosses. All predicted coordinates whose GT matches the combination in the title are marked with red dots.

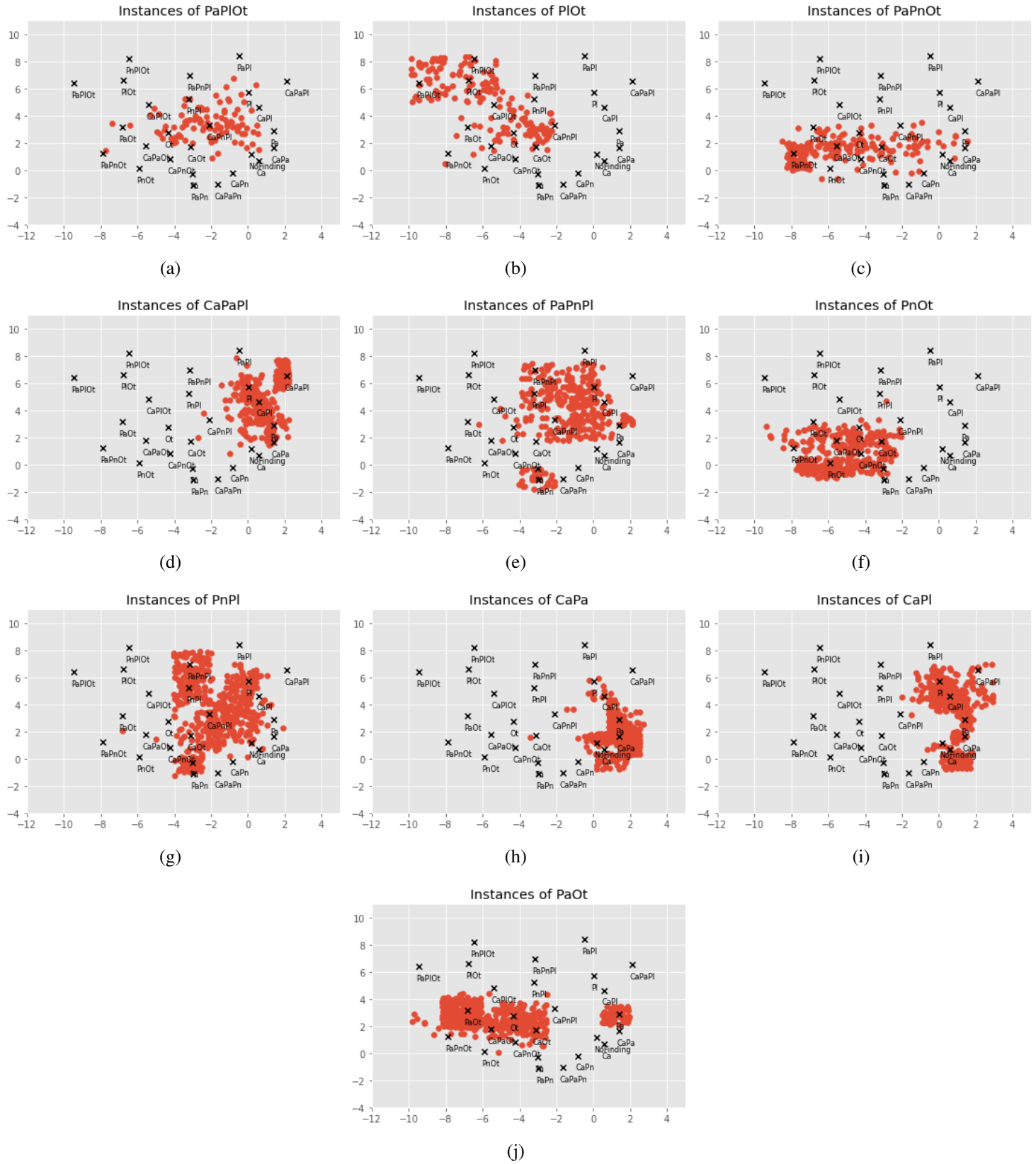



Figure B.4: Predicted embedding coordinates obtained by the Phase II regressor for intermediately-represented combinations. The GT references are marked with black crosses. All predicted coordinates whose GT matches the combination in the title are marked with red dots.

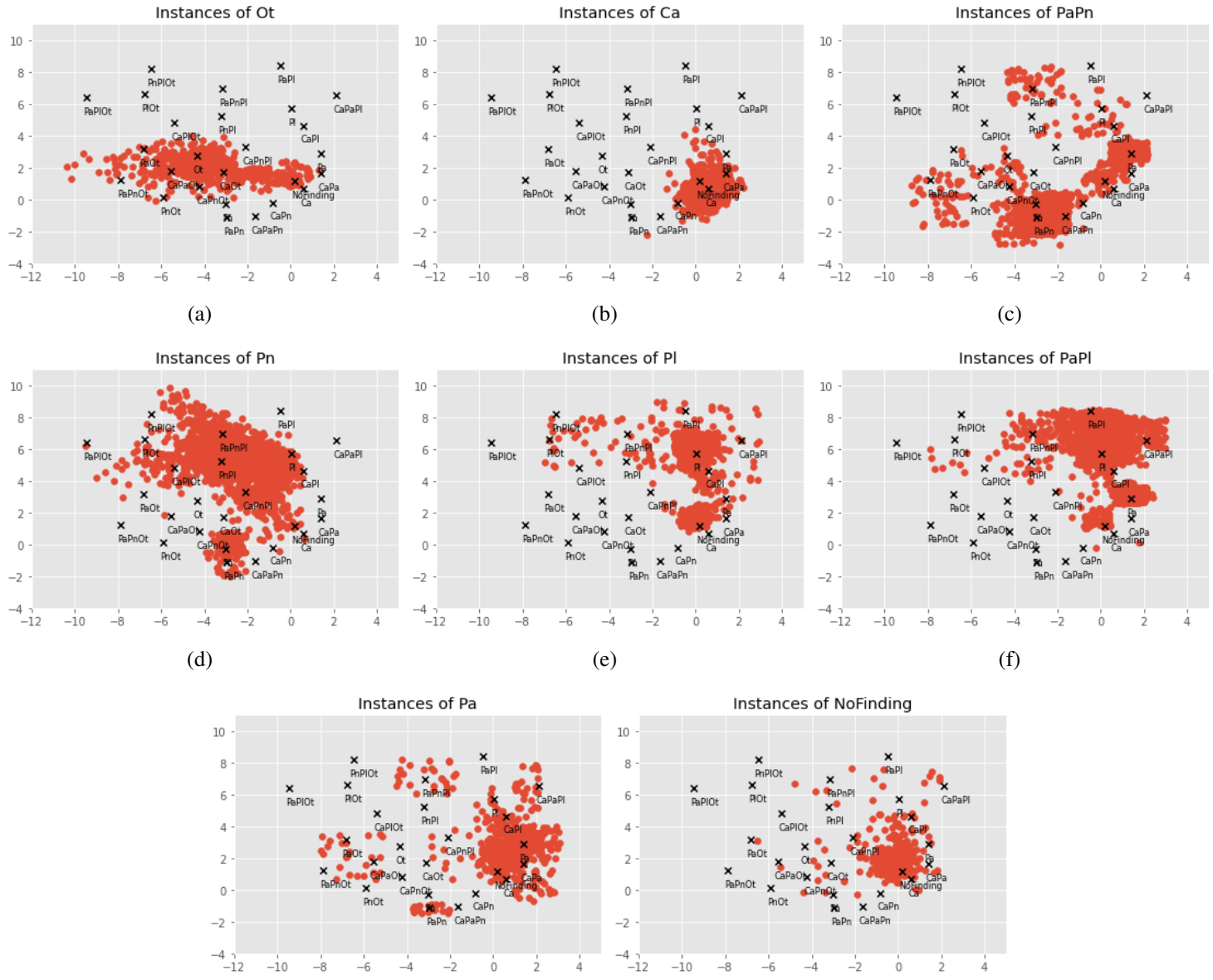
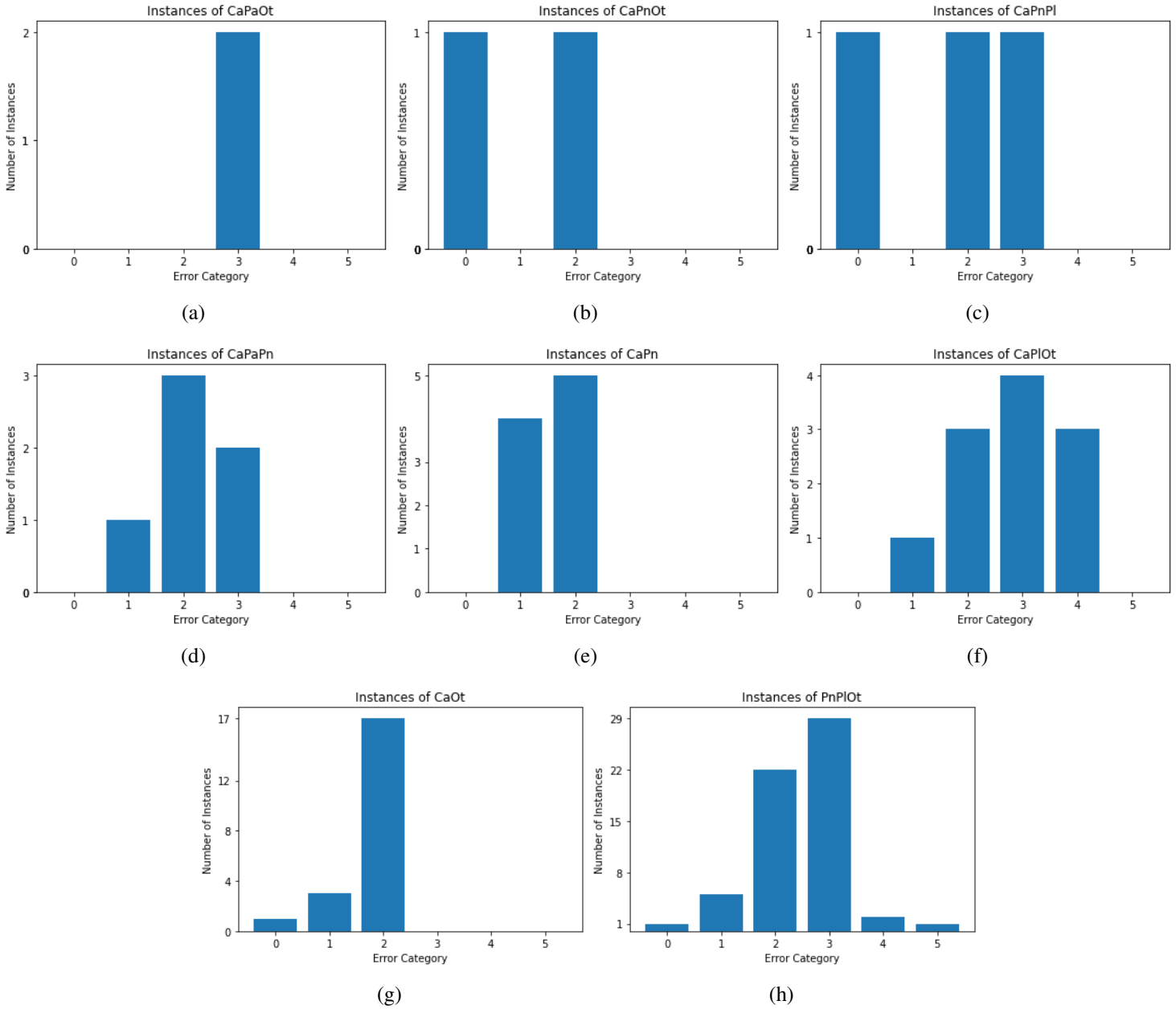



Figure B.5: Predicted embedding coordinates obtained by the Phase II regressor for highly-represented combinations. The GT references are marked with black crosses. All predicted coordinates whose GT matches the combination in the title are marked with red dots.




(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure B.6: Frequency of instances per error category for under-represented combinations.

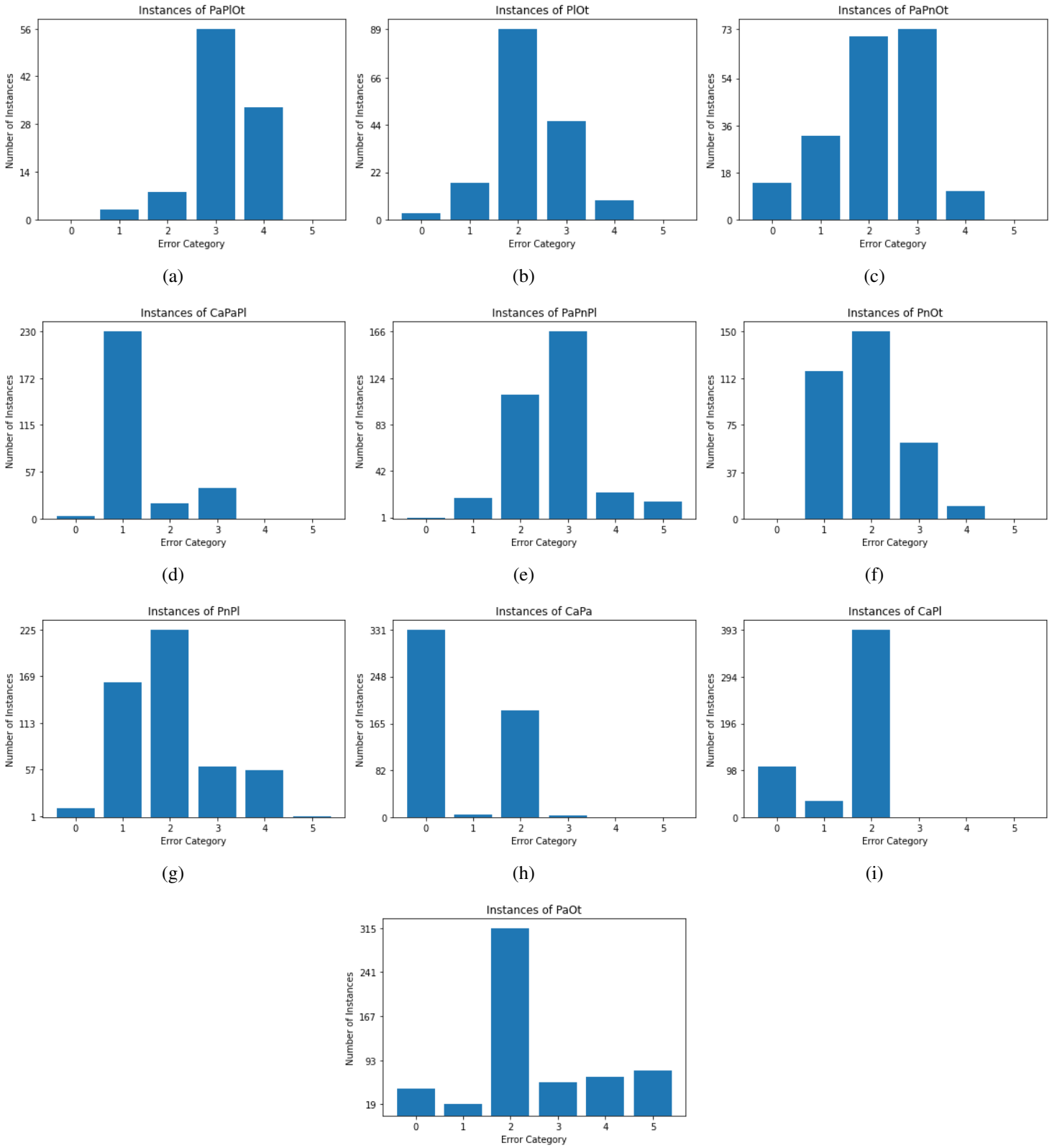



Figure B.7: Frequency of instances per error category for intermediately-represented combinations.

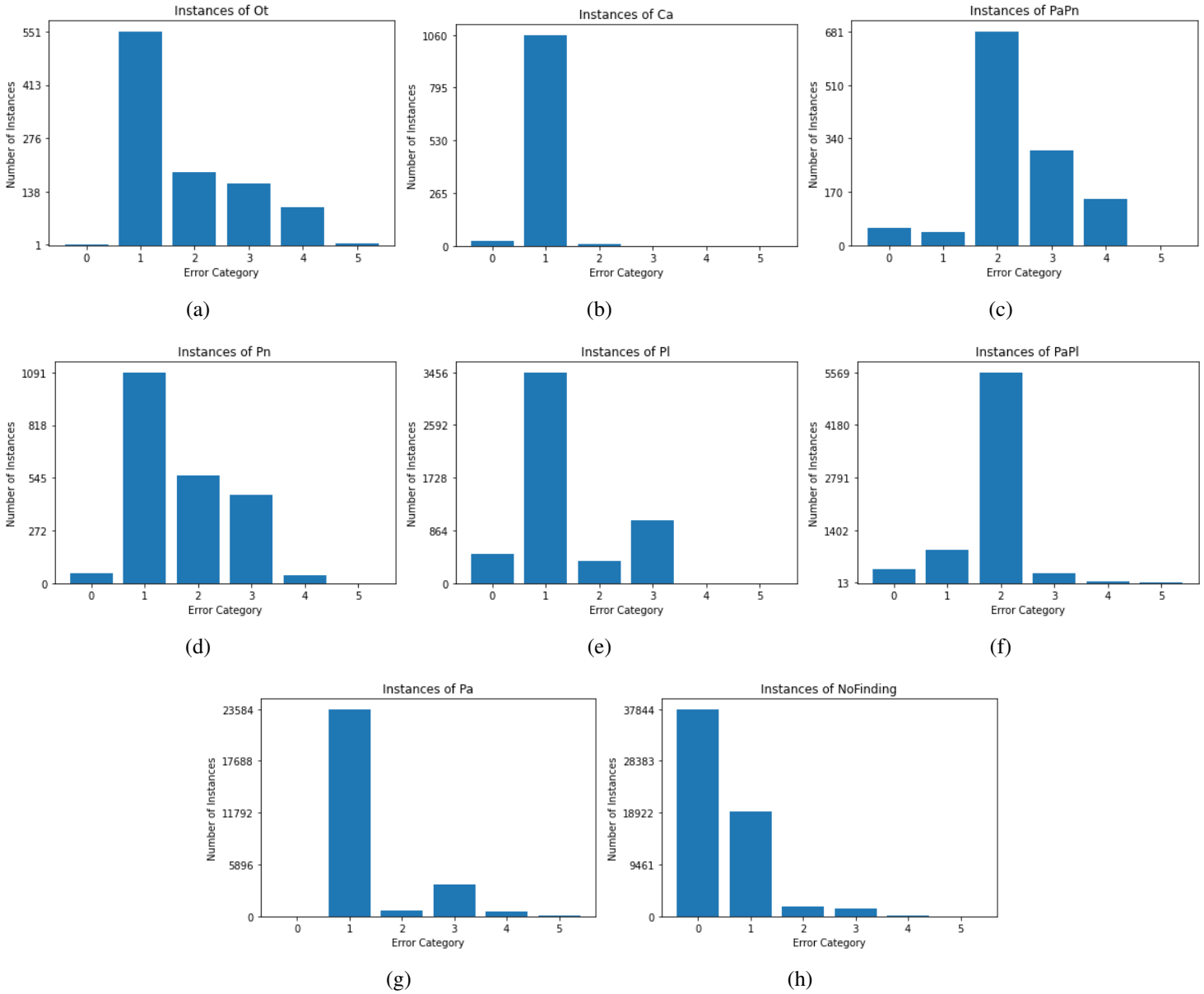



Figure B.8: Frequency of instances per error category for highly-represented combinations.