machine learning & knowledge extraction

MDPI

Article

EyeXNet: Enhancing Abnormality Detection and Diagnosis via Eye-Tracking and X-ray Fusion

Chihcheng Hsieh [1,†], André Luís [2,3,†], José Neves [2,3], Isabel Blanco Nobre [4], Sandra Costa Sousa [4], Chun Ouyang [1], Joaquim Jorge [2,3] and Catarina Moreira [1,2,5,*]

[1] School of Information Systems, Queensland University of Technology, Brisbane, QLD 4000, Australia; chihcheng.hsieh@hdr.qut.edu.au (C.H.); c.ouyang@qut.edu.au (C.O.)
[2] Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal; andre.t.luis@tecnico.ulisboa.pt (A.L.); jose.s.neves@tecnico.ulisboa.pt (J.N.); jorgej@tecnico.ulisboa.pt (J.J.)
[3] INESC-ID, 1000-029 Lisbon, Portugal
[4] Grupo Lusíadas, Imagiology Department, 1500-458 Lisbon, Portugal; isabel.blanco.nobre@lusiadas.pt (I.B.N.); sandra.costa.sousa@lusiadas.pt (S.C.S.)
[5] Human Technology Institute, University of Technology Sydney, Sydney, NSW 2007, Australia
* Correspondence: catarina.pintomoreira@uts.edu.au
† These authors contributed equally to this work.

Abstract: Integrating eye gaze data with chest X-ray images in deep learning (DL) has led to contradictory conclusions in the literature. Some authors assert that eye gaze data can enhance prediction accuracy, while others consider eye tracking irrelevant for predictive tasks. We argue that this disagreement lies in how researchers process eye-tracking data as most remain agnostic to the human component and apply the data directly to DL models without proper preprocessing. We present EyeXNet, a multimodal DL architecture that combines images and radiologists' fixation masks to predict abnormality locations in chest X-rays. We focus on fixation maps during reporting moments as radiologists are more likely to focus on regions with abnormalities and provide more targeted regions to the predictive models. Our analysis compares radiologist fixations in both silent and reporting moments, revealing that more targeted and focused fixations occur during reporting. Our results show that integrating the fixation masks in a multimodal DL architecture outperformed the baseline model in five out of eight experiments regarding average Recall and six out of eight regarding average Precision. Incorporating fixation masks representing radiologists' classification patterns in a multimodal DL architecture benefits lesion detection in chest X-ray (CXR) images, particularly when there is a strong correlation between fixation masks and generated proposal regions. This highlights the potential of leveraging fixation masks to enhance multimodal DL architectures for CXR image analysis. This work represents a first step towards human-centered DL, moving away from traditional data-driven and human-agnostic approaches.

Keywords: multimodal deep learning; eye tracking; object detection; X-rays; fixation maps

Citation: Hsieh, C.; Luís, A.; Neves, J.; Nobre, I.B.; Sousa, S.C.; Ouyang, C.; Jorge, J.; Moreira, C. EyeXNet: Enhancing Abnormality Detection and Diagnosis via Eye-Tracking and X-ray Fusion. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1055–1071. https://doi.org/10.3390/make6020048

Academic Editor: Andreas Holzinger

Received: 24 March 2024
Revised: 19 April 2024
Accepted: 6 May 2024
Published: 9 May 2024

1. Introduction

Chest X-ray (CXR) imaging is paramount in diagnosing and monitoring various thoracic diseases [1]. Using deep learning models to automatically detect abnormalities in CXRs has demonstrated promising results, augmenting the efficiency of medical diagnosis and alleviating the workload of radiologists [2]. Recently, the release of two major datasets, REFLACX [3] and EyeGaze [4], containing radiologists' eye-tracking data has triggered interest in the development of innovative deep learning (DL) architectures that integrate both radiologists' fixation masks and CXR images. The objective is to harness radiologists' visual patterns to not only enhance the performance of these predictive systems but also to shift towards more human-centric architectures, which learn from human visual patterns [5–7] rather than relying solely on image-level information that is susceptible to

bias [8]. Nonetheless, the efficacy of this approach and its impact on various model architectures remain ambiguous [9]. The literature presents conflicting perspectives on these datasets, with some studies asserting that incorporating eye-tracking data in DL can result in performance enhancements [10–12]. In contrast, others either argue the contrary [13–15] or are inconclusive [16].

In a recent study by Moreira et al. [17], an analysis of the EyeGaze dataset revealed that radiologists tended to focus more on regions of interest (which could potentially contain abnormalities) when verbally reporting their findings as opposed to silently examining the image. To provide further context, during the data collection process, the radiologists initially spent a certain amount of time quietly assessing the image, and only afterwards did they begin to report their observations. This finding suggests that the radiologists' attention was drawn more to the salient areas of the image during the reporting phase, providing valuable insights into the regions that potentially contain abnormalities. However, the authors did not validate their analysis using any deep learning approach to determine whether exclusively considering fixation masks corresponding to reporting moments could lead to a performance advantage.

In this paper, we extend the ideas in Moreira et al. [17] by investigating the impact of incorporating fixation masks, specifically those corresponding to reporting moments, into deep learning models to automatically detect abnormalities in chest X-ray images. Building upon these findings, we explore the potential performance advantages of using these refined fixation masks in conjunction with various DL architectures. By comparing the outcomes of models that incorporate these fixation masks with those that rely solely on image-level information, we seek to provide a clearer understanding of the value of integrating radiologists' visual patterns in the development of more accurate and human-centric systems [18,19]. Recently, the European fundamental rights mandated the use of explainable and interpretable AI in medicine [20], ensuring that patients are informed about the essential functions of AI in an intelligible form before its use. This requirement aligns with the broader European legal framework, which prioritizes human oversight, privacy by design, and non-discrimination in AI applications, including those in the medical domain.

To validate this hypothesis, we propose EyeXNet, a deep learning framework that combines chest X-ray images with fixation masks corresponding to the reporting moments of radiologists. EyeXNet is designed to leverage the valuable insights gained from radiologists' visual patterns during the reporting phase, aiming to improve the performance of automatic abnormality detection in chest X-ray images.

The main contributions of this paper are the following:

- EyeXNet, a DL framework that combines chest X-ray images with fixation maps corresponding to the reporting moments of radiologists, aiming to improve the performance of automatic abnormality detection in chest X-ray images.
- To the extent of our knowledge, our approach is the first in the literature to use the filtering of gaze related to non-reporting moments to obtain more meaningful eye-tracking data as a proxy of the radiologists' attention in the abnormality detection network's training process.
- A comprehensive evaluation of EyeXNet using various DL architectures, including ConvNext, DenseNet, VGG, MobileNet, RegNet, and EfficientNet, and a comparison of their performance with baseline models that relies solely on image-level information.
- An analysis involving a think-aloud experiment with two experienced radiologists sheds light on the challenges faced by radiologists during their assessments and how these challenges may influence the performance of DL models for chest X-ray abnormality detection.

2. Key Eye-Tracking Concepts in Radiology

Eye-tracking data, specifically fixations and saccades, provides a valuable understanding of cognitive activities. Such data are visually manifested through fixation masks, a critical component of EyeXNet.

Fixations denote the instances where the gaze remains stationary over a specific area about the size of the fovea, signifying intensive information processing [21]. These instances suggest that the eye momentarily suspends to examine regions of potential significance closely. This is particularly relevant in radiology as empirical evidence suggests that the initial fixations made by radiologists frequently align with regions containing lesions [3,22]. Thus, fixation data are essential in highlighting areas of diagnostic interest within the image.

Saccades represent swift ocular movements between fixation points. These movements essentially link different fixations, forming a pathway that connects separate regions of in-depth examination within an image [23].

Fixation Masks serve as graphical interpretations of collected fixation data, commonly formed by establishing Gaussians at the centre of each fixation, where the intensity is directly proportional to the duration of the fixation [3]. These masks provide a consolidated visual representation of an image's areas where the observer's gaze has concentrated, thereby spotlighting areas considered informative for particular tasks, such as disease classification [24]. In the context of this study, it is important to note that we use the terms "fixation masks", "fixation maps", and "fixation heatmaps" interchangeably and that they refer to the same concept.

The key challenge is to find the most effective strategy for integrating eye gaze data into DL architectures to enhance lesion detection.

3. Related Work

Several studies have integrated eye-tracking data into multimodal DL architectures. These studies aim to enhance lesion detection in CXR images by incorporating radiologists' eye gaze patterns. Determining the most effective approach with which to incorporate this information into the learning process remains an ongoing research challenge.

There are two main approaches for incorporating eye-tracking (ET) data into predicting image-level labels for CXR images from the EyeGaze dataset [4]. The first approach combines CXR images processed through a CNN with temporal fixation heatmaps using a 1-layer bidirectional long short-term memory network with self-attention [25]. This method yielded a 4% AUC improvement by incorporating temporal fixation heatmaps compared to the baseline model, which used only CXR image data as input. The second approach employed static fixation heatmaps, aggregating all temporal fixations into one image. During training, the model jointly learns from the static fixation heatmap and the image-level label. In the testing phase, the model receives a CXR image as input and outputs both the label and a heatmap distribution of the most crucial locations for the condition. This approach demonstrated similar results to the baseline model, with the added benefit of enhanced interpretability provided by the heatmaps.

Wang et al. [10] explored the utility of gaze data collected from radiologists by developing a gaze-guided attention network that focuses on disease regions similar to radiologists and outputs the disease label along with an attention map based on annotated bounding boxes and radiologists' gaze information. The results revealed that the radiologist's gaze-guided GA-net architecture outperformed state-of-the-art methods using only images, such as ResNet [26] and the Vision Transformer [27]. Moreover, collecting gaze data was faster than acquiring manually annotated bounding boxes from radiologists while achieving comparable classification accuracy. Other studies have suggested that eye-tracking data can be valuable as the initial fixations made by radiologists often coincide with lesion-containing regions [22].

Nevertheless, some research argues against using saliency maps from human fixations in deep learning models as these systems may rely on background context for object classification, introducing biases [13,15]. Early studies on ET data in CXR analysis [28] identified three types of diagnostic errors associated with eye-tracking data: (1) search errors, where the target is missed; (2) recognition errors, occurring when the eyes fixate on the target but the target remains undetected; and (3) decision errors, which stem from the radiologist's inability to report the findings. These errors could contaminate ET data

collection and adversely impact the performance of deep learning models. In similar studies, the proportion of these errors was found to be approximately 30% search errors, 25% recognition errors, and 45% decision errors [29,30].

4. EyeXNet Architecture

In this section, we introduce EyeXNet, a novel DL framework incorporating CXR images and radiologist fixation masks for improved abnormality detection in CXRs. Figure 1 presents the architecture of EyeXNet, which is a two-stage detector based on Faster R-CNN. We also introduce details on how to reproduce our results (including information about the parameters used and the models trained can be found in our public Github repository: https://github.com/ChihchengHsieh/MIMIC-Eye-applications, (accessed on 8 May 2024)).

The Faster R-CNN framework is a powerful model for object segmentation. The key components are the following.

Backbone Network: This is a deep convolutional network that serves as the feature extractor. In Mask R-CNN, the backbone could be any feature-rich network like ResNet or VGG, designed to process input images and produce a high-dimensional feature map.

Region Proposal Network (RPN): This network proposes candidate object bounding boxes. It slides over the feature map output by the backbone and outputs a set of rectangles (proposals) that are likely to contain objects.

ROI Pooling: This extracts a small feature map for each proposal, aligning the extracted features with the input, which is crucial for accurate mask prediction.

Bounding Box Regression and Classification Head: This part of the network predicts the class of each object and refines the bounding box coordinates proposed by the RPN.

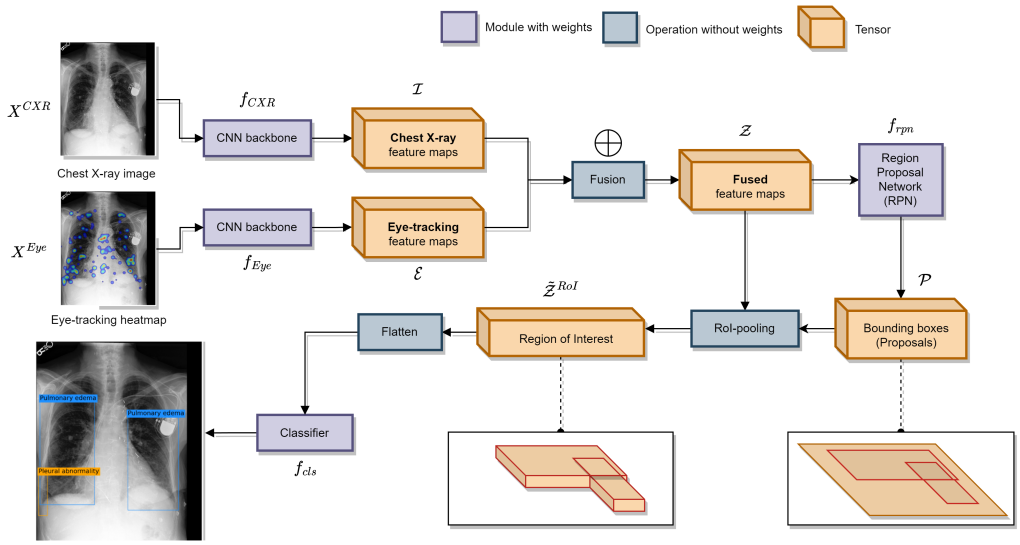

Figure 1. EyeXNet proposed architecture. The model fuses X-ray images with eye gaze fixation masks to yield regions of interest and abnormality labels.

EyeXNet leverages the valuable insights gained from radiologists' visual patterns during the reporting phase to enhance the performance of automatic abnormality detection. The architecture comprises the following key components:

a. Input Layer: The proposed network receives two modalities as inputs: a front (AP or anterior-posterior view) view of CXR images and the respective clinical data. They are defined as:

- A set of CXR images: $X^{\text{CXR}} \in \mathbb{R}^{W \times H \times C}$;
- A set of eye-tracking heatmap: $X^{\text{Eye}} \in \mathbb{R}^{W \times H \times C}$.

In the proposed architecture, we set the dimensions of our image input space to $W = H = 512, C = 1$.

b. Feature Engineering: $X^{\text{Eye}} \rightarrow \mathcal{E} \in \mathbb{R}^{W' \times H' \times D'}$
$X^{\text{CXR}} \rightarrow \mathcal{I} \in \mathbb{R}^{W' \times H' \times D'}$

The goal of this step is to extract the feature maps for both input sources:

b.1. Chest X-ray Transform: we extracted feature maps ($\mathcal{I}$) from the input chest X-ray image (X^{CXR}) using a CNN backbone (f_{CXR}):

$$\mathcal{I} = f_{\text{CXR}}(X^{\text{CXR}}),$$

where the resulting dimensional space is $\mathcal{I} \in \mathbb{R}^{W' \times H' \times D'}$. The values of W', H', and D are decided by the architecture of the backbone. In this work, a variety of the backbone is used to form a robust evaluation.

b.2. Eye-Tracking Heatmap Transform: To fuse eye-tracking heatmaps (X^{Eye}) with chest X-ray images, the same CNN backbone architecture (f_{Eye}) is used to generate the feature maps ($\mathcal{E}$):

$$\mathcal{E} = f_{\text{Eye}}(X^{\text{Eye}}),$$

where $\mathcal{E} \in \mathbb{R}^{W' \times H' \times D'}$, which is the same shape as chest X-ray feature maps ($\mathcal{I}$). Although the architecture is the same, the weights are not shared since eye-tracking heatmaps and chest X-ray images are very different image data types. Once we have feature maps of the same size for both modalities, we can proceed to the fusion phase.

c. Feature Maps Fusion: $\{\mathcal{I}, \mathcal{E}\} \rightarrow \mathcal{Z} \in \mathbb{R}^{W' \times H' \times D'}$

The final feature map ($\mathcal{Z}$) representing the element-wise sum fusion of both modalities is obtained by

$$\mathcal{Z} = \mathcal{C} \oplus \mathcal{E}, \mathcal{Z} \in \mathbb{R}^{H' \times W' \times D'}, \tag{1}$$

where $\oplus$ corresponds to the element-wise sum operation used for fusion. The final $\mathcal{Z}$ corresponds to the 3D feature map representation of the fused information. Next, we use this data representation as input to the Faster-RCNN architecture to perform lesion detection.

d. Region Proposal Network: $\{\mathcal{Z}\} \rightarrow \tilde{\mathcal{Z}}^{\text{RoI}} \in \mathbb{R}^{W_r \times H_r}$

To perform localized abnormality detection, we use the Region Proposed Network (RPN), f_{rpn}, of the Faster-RCNN architecture to generate candidate object bounding boxes, also known as proposals $\mathcal{P}$, given by

$$\mathcal{P} = f_{\text{rpn}}(\mathcal{Z}), \forall p_i \in \mathcal{P} : p_i = (x_i, y_i, w_i, h_i, c_i^{\text{obj}}). \tag{2}$$

RPN learns the coordinates of the generated bounding boxes (x_i, y_i, w_i, h_i), and the corresponding confidence score, c_{obj}, of having an abnormality (object) in the localization of the bounding boxes. This confidence score is used to sort the generated proposals by their predictive relevance.

Using the coordinates of the computed bounding boxes, a RoIPool operation is performed to extract the corresponding Regions of Interest (RoIs), $\tilde{\mathcal{Z}}^{\text{RoI}} = \text{RoIPool}(\mathcal{P}, \mathcal{Z})$. The RoIs result in a data structure with dimensions $\tilde{\mathcal{Z}}^{\text{RoI}} \in \mathbb{R}^{W_r \times H_r}$, where W_r and H_r are hyperparameters. We set W_r and H_r to 7 in our experiments.

e. Output: $\{Z^C, \tilde{\mathcal{Z}}^{\text{RoI}}\} \rightarrow \hat{y}$:

After learning the candidate RoIs, we flatten these data to serve as input to a normal dense neural network, which will perform the final classification. To emphasize the role of the clinical data in this classification process, we concatenate the clinical data representation,

Z^C, with the flattened candidate RoIs, $\tilde{\mathcal{Z}}^{\text{RoI}}$, before classification takes place. The role of the 1-Difusion in the MDF-Net is to provide residual information to further pass the clinical data to deeper layers in our architecture. The final prediction $\hat{y}$ is then obtained by:

$$\hat{y} = f_{\text{cls}}(\text{Flatten}(\tilde{\mathcal{Z}}^{\text{RoI}}) \cup Z^{\text{C}}), \tag{3}$$

where $\cup$ represents the vector concatenation operation, $\tilde{\mathcal{Z}}_{\text{RoI}} = \text{RoIPool}(\mathcal{P}, \mathcal{Z})$, and $\hat{y}$ contains predicted classes $\hat{y}^{\text{cls}}$, bounding boxes $\hat{y}^{\text{bb}}$, and binary masks $\hat{y}^{\text{mask}}$, and where f_{cls} is the final classification layer.

f. Training: In the training phase, five loss terms are considered: four original from the Mask R-CNN framework plus one related to the inclusion of eye tracking data into the model's learning process.
L_{cls}: Cross-entropy between groundtruth abnormality y^{cls} and predicted abnormalities $\hat{y}^{\text{cls}}$. This loss term requires the model to predict the class of abnormalities correctly in the output layer:

$$L_{\text{cls}} = -\sum_i y_i^{\text{cls}} \log(\hat{y}_i^{\text{cls}}) \tag{4}$$

L_{bb}: Bounding box regression loss between ground-truth bounding boxes y^{bb} and predicted bounding boxes $\hat{y}^{\text{cls}}$ is calculated using the smooth-L_1 norm using hyperparameter β. To minimize this loss, the model has to locate abnormalities in the correct areas in the output layer. In our implementation, we set $\beta = \frac{1}{9}$.

$$\begin{gathered} L_{bb} = \sum_{i=1}^{n} l_i, \text{ where} \\ l_i = \begin{cases} 0.5(\hat{y}_i^{\text{bb}} - y_i^{\text{bb}})^2/\beta & \text{, if } \hat{y}_i^{\text{bb}} - y_i^{\text{bb}} < \beta \\ |\hat{y}_i^{\text{bb}} - y_i^{\text{bb}}| - 0.5 * \beta & \text{, otherwise} \end{cases} \end{gathered} \tag{5}$$

$L_{\text{obj}_{\text{rpn}}}$: Binary cross-entropy loss between ground-truth objects y^{obj} and predicted objects c^{obj} (confidence score), which requires RPN to correctly classify whether the proposals (candidate bounding boxes) contain any abnormality.

$$L_{\text{obj}_{\text{rpn}}} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_{\text{obj}}^{(i)} \cdot \log(c_{\text{obj}}^{(i)}) + (1 - y_{\text{obj}}^{(i)}) \cdot \log(1 - c_{\text{obj}}^{(i)})\right) \tag{6}$$

$L_{\text{bb}_{\text{rpn}}}$: Proposal regression loss between proposals (candidate bounding boxes) $p : p \in \mathcal{P}$ and ground-truth bounding boxes y^{bb}, which is also calculated using the same smooth-L_1 norm function for L_{bb}. This loss term aims to improve RPN on localising abnormalities.

We used homoscedastic (task) uncertainty [31] to train the proposed model using these five loss terms by dynamically weighting them for better convergence. Let $\mathcal{L} = \{L_{\text{cls}}, L_{\text{bb}}, L_{\text{mask}}, L_{\text{obj}_{\text{rpn}}}, L_{\text{bb}_{\text{rpn}}}\}$; we used the SGD (stochastic gradient descent) to optimize the overall loss function

$$\arg\min_{\theta,\alpha_l} \sum_{l\in\mathcal{L}} \frac{1}{2\alpha_l^2} l(\theta) + \log \alpha_l^2, \tag{7}$$

where θ is the weights of the model, and α_l is a trainable parameter to weight each task/loss.

5. EyeXNet Complexity Analysis

The overall complexity of EyeXNet largely mirrors that of the original Mask R-CNN model, considering the sum of the complexities of its individual components, primarily structured as follows: $O(NCHW)$, where N is the number of region proposals, C is the number of classes (in our case, five), H is the height of the feature map, and W is the width of the feature map.

The EyeXNet architecture includes five primary components: (1) a deep fully convolutional network, (2) a region proposal network, (3) the ROI pooling, (4) a bounding box regressor, and (5) a classifier, described below.

Deep Fully Convolutional Network: This component uses a structure based on Zeiler and Fergus's [32] smaller, faster model. It extracts $256 \times N \times N$ feature maps from the input image I, which is input for the RPN and ROI pooling layers.

Region Proposal Network: For each pixel in the feature map, the RPN generates K anchor boxes (or candidate windows), typically totalling 2000, considering different scales and ratios. After applying non-maximum suppression, we retain approximately 2000 candidate windows, resulting in a complexity of $O(N^2)$ given the quadratic relationship with the number of proposals per feature map location.

ROI Pooling: This component takes the variable-sized candidate windows and divides each into an HxW grid of sub-windows. It then performs max pooling across these sub-windows to produce a fixed-size output feature map for each region, which is efficient with a complexity of $O(1)$ per region.

We estimate that the computational cost of EyeXNet approximates that of the foundational Mask R-CNN, albeit adapted for including eye-tracking masks. This adaptation introduces additional steps in data preprocessing and integration but does not significantly alter the primary computational complexity, which remains dominated by the convolutional operations and region proposal computations.

6. Experimental Setup

This study aims to evaluate the effectiveness of incorporating radiologist fixation masks obtained from eye-tracking data into EyeXNet to predict the location of various abnormalities in chest X-ray images. We hypothesise that radiologists should focus more on the regions containing abnormalities when they report what they see in the image [17], leading to improved classifier performance. All the experiments were on a single PC with an i9-13900K CPU and an NVIDIA GeForce RTX 4090 GPU with 24 GB RAM both located in Santa Clara, California, USA.

6.1. Models

We evaluated EyeXNet with eight backbones: MobileNet, ResNet18, DenseNet161, EfficientNetB0, EfficientNetB5, ConvNextNet, VGG16, and RegNet. As a baseline, we used EyeXNet, using only the images. To train the models, we used the top five occurring lesions in the MIMIC-EYE dataset: Pulmonary Edema, Enlarged Cardiac Silhouette, Consolidation, Atelectasis, and Pleural Abnormality, as described below:

MobileNet [33] is characterized by its use of depth-wise separable convolutions, a design that reduces the model size and computational cost. This architecture is particularly advantageous for deployment on mobile and edge devices due to its efficiency. However, the simplicity that affords MobileNet its speed can also lead to lower accuracy compared to more complex models, a crucial trade-off in resource-constrained environments.

ResNet18 [34] introduces the concept of residual blocks with skip connections. This significant innovation supports the training of deeper network architectures by mitigating the vanishing gradient problem. These properties make ResNet18 robust for various tasks, although it may still be cumbersome for real-time applications due to its relative size compared to more streamlined models.

DenseNet161 [35] stands out due to its dense connectivity pattern, where each layer is connected to every other layer in a feed-forward fashion. This dense connectivity ensures maximal information flow between layers, enhancing feature propagation and reducing overfitting. The downside is the increased computational demand, leading to higher memory consumption and slower inference times.

EfficientNet [36] represents a balanced approach, scaling convolutional neural networks through a compound coefficient that uniformly scales depth, width, and resolution. This methodical scaling has been optimized through neural architecture search, offering a commendable balance between accuracy and efficiency. Nevertheless, the complexity of EfficientNet might pose challenges in implementation, especially in environments where computational resources are limited.

ConvNextNet [37] reflects the latest advancements in integrating transformer-like elements within a conventional CNN framework, pushing the boundaries of what can be achieved with convolutional architectures. Although it offers state-of-the-art performance, ConvNextNet's relatively recent development means it might lack extensive community support and a wealth of deployment experiences, which are invaluable for real-world applications.

VGG16 [38], with its straightforward design of deep convolutional layers, has been a reliable workhorse in the deep learning community. Its simplicity and depth have made it a standard for many benchmark tasks. However, its large size and computational demands make it less ideal for deployment where efficiency is a priority.

RegNet [39] employs a systematic approach to network design, which yields models that scale predictably in performance while being computationally efficient. This predictability is beneficial, yet the rigidity of its design rules may not capture the specific nuances required for all types of diagnostic tasks, potentially limiting its effectiveness in more specialized applications.

6.2. Dataset

In our experiments, we used the MIMIC-EYE dataset [40] to validate a multimodal deep learning architecture for improved chest X-ray abnormality detection. MIMIC-EYE combines multiple MIMIC data sources, including medical images and reports (MIMIC-CXR [41,42] and MIMIC-JPG [43]); clinical data (MIMIC-IV ED [44]); exhaustive patient hospitalization records (MIMIC-IV [45,46]); and eye-tracking data including gaze patterns, pupil dilations, and image annotations (REFLACX [3] and EyeGaze [4]). The dataset comprises 3192 patients, 1644 stays, and 3689 CXR images, offering a robust foundation for evaluating the proposed architecture. Our study used 2122 pairs of CXR images and fixation masks for training, 455 for validation, and 455 for testing. For this study, we focused on the REFLACX subset of this integrated dataset because it contained annotations of CXR abnormalities.

6.3. Data Preprocessing

In the MIMIC-EYE dataset, radiologists were instructed to evaluate chest X-ray images using a think-aloud protocol while their eye movements were recorded. Initially, radiologists silently observed each image before verbally reporting their findings. Moreira et al. [17] observed significant differences in radiologists' eye movements during the silent examination phase compared to when they began speaking. The authors posited that radiologists focused more on regions containing abnormalities during the reporting phase, making these moments valuable for informing deep learning models. In this study, following Moreira et al. [17], we opted to generate fixation masks solely for the moments when

radiologists were speaking as these instances exhibited a stronger correlation between fixations and the ground truth locations of abnormalities in the X-ray images (Figure 2).

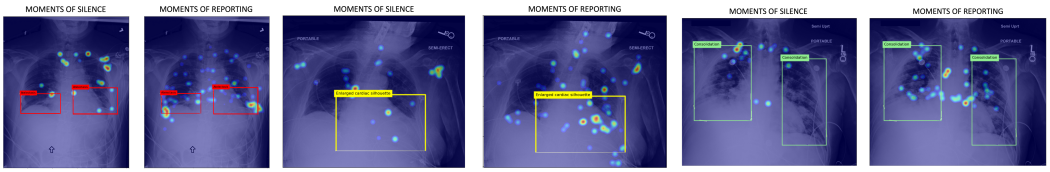

Figure 2. Examples of chest X-rays featuring radiologist fixations during both silent observation and reporting phases, alongside their ground truths. The fixations made during the reporting phase demonstrate a stronger correlation with the ground truth, suggesting that these moments may provide informative cues for DL models.

6.4. Evaluation

We employed the Free-Response Receiver Operating Characteristic (FROC) analysis to evaluate the proposed EyeXNet's performance on the MIMIC-EYE dataset. FROC analysis is a sophisticated method used to evaluate the performance of diagnostic systems, particularly those involved in detecting and identifying multiple occurrences of objects, such as tumors or lesions, within medical images. Unlike the traditional Receiver Operating Characteristic (ROC) curves, designed for binary classification tasks, FROC curves provide a more nuanced evaluation. They allow for the assessment of systems where true and false positives can occur multiple times within the same image set [47].

We assessed the performance of our model at various false positive rates, specifically 0.5, 1, 2, and 4 false positives per image. The FROC curve is generated by plotting the sensitivity (true positive rate) against the average number of false positives per image (FPI) at these different thresholds,

$$FROC@i = [FPI@(i,t), Sensitivity@(i,t)] \quad \text{where } t \in [0,1] \tag{8}$$

where $FROC@i$ represents the FROC value of a given class i, $Sensitivity@(i,t)$ is the true positive rate at threshold t for class i, $FPI@(i,t)$ is the number of false positives at threshold t per image for lass i, and t is the score threshold (probability confidence). A higher FROC value indicates a better diagnostic system performance, implying that the model can detect more true positives with fewer false positives at a given threshold. A higher FROC value primarily results from: (1) increased detection sensitivity, (2) reduced false positives, and (3) model robustness. A higher FROC value suggests that the system achieves an optimal balance between sensitivity and Precision, which is critical for clinical reliability. It indicates that the system effectively identifies true positives with fewer false positives, enhancing its utility in patient care. The FROC curve plot analysis is crucial in medical imaging because it offers a detailed measure of a system's ability to correctly identify multiple pathological findings while quantifying the rate of false alarms. Such information is vital for determining the practical utility of diagnostic systems in clinical settings, where high sensitivity and controlled false positive rates are essential for effective patient management.

To supplement our quantitative analysis, we enlisted the expertise of two seasoned radiologists, each with over 15 and 25 years of experience, to provide insights into the best- and worst-performing predictions made by our proposed EyeXNet model, using the top-performing backbone. Our goal was to determine whether instances where the model struggled to make accurate predictions were also challenging for human radiologists to evaluate.

6.5. Model Complexity Analysis

The computational complexity of the proposed architecture can be approximated to the complexity of the original Mask R-CNN model. Since both the CXR image and fixation masks are fused, the resulting feature maps approximate the performance of the Mask R-CNN with a single modality. It can be estimated by summing the complexities of the components of the original Mask R-CNN model:

$$O(NHW) + O(NCHW) + O(NCHW) \approx O(NCHW),$$

where N is the number of region proposals, C is the number of classes (five classes, in our case), H is the feature map height, and W is the feature map width.

7. Results

Table 1 displays the sensitivity results at false positive rates of 0.5, 1, 2, and 4, along with the corresponding mean FROC, $FROC@[0.5, 1.0, 2.0, 4.0]$, for various backbones employed in the proposed EyeXNet. The FROC curves for each backbone, depicted in Figure 3, illustrate the trade-off between sensitivity and false positive rates for different predicted classes. These curves aid in identifying an optimal threshold that maximizes sensitivity while minimizing false positive rates, thereby providing valuable insights into each backbone's performance.

Table 1 presents varying performance outcomes between the baseline models (utilizing chest X-ray images solely) for DenseNet, EfficientNet, and EyeXNet; it combines radiologists' fixation maps with ConvNextHowever, these results do not directly suggest the effectiveness of integrating fixation maps into a multimodal architecture.

Table 1. Sensitivity at average false positive rates of 0.5, 1, 2, and 4, along with the corresponding mean FROC scores for different EyeXNet backbones using both chest X-ray (CXR) images and radiologists' fixation masks, as well as the CXR images alone. This comparison highlights the performance of each EyeXNet backbone in detecting abnormalities, considering the influence of incorporating fixation masks.

Backbone	Setting	Sensitivity @[0.5]	Sensitivity @[1.0]	Sensitivity @[2.0]	Sensitivity @[4.0]	mFROC @[0.5, 1, 2, 4]	mRecall
MobileNet [33]	image only	0.489	0.665	0.813	0.916	0.720	0.345
	fixation maps	**0.520**	**0.686**	**0.830**	**0.946**	**0.745**	**0.448**
ResNet18 [34]	image only	0.547	0.700	0.844	0.927	0.755	0.348
	fixation maps	**0.571**	**0.735**	**0.881**	**0.969**	**0.789**	**0.448**
DenseNet161 [35]	image only	**0.597**	**0.796**	**0.922**	0.967	**0.820**	0.396
	fixation maps	0.564	0.721	0.860	0.957	0.775	**0.414**
EfficientNetB5 [36]	image only	0.565	0.744	**0.886**	0.951	0.787	**0.491**
	fixation maps	0.566	**0.752**	0.883	0.961	0.791	0.391
EfficientNetB0 [36]	image only	0.575	0.750	0.877	0.947	0.786	0.325
	fixation maps	**0.581**	**0.764**	**0.895**	0.971	**0.803**	**0.414**
ConvNextNet [37]	image only	0.574	**0.748**	0.902	**0.965**	0.797	0.457
	fixation maps	**0.597**	**0.759**	**0.907**	**0.975**	**0.808**	**0.491**
VGG16 [38]	image only	0.570	0.753	0.897	**0.970**	0.798	0.443
	fixation maps	**0.582**	**0.784**	**0.914**	**0.987**	**0.817**	0.430
RegNet [39]	image only	0.432	0.626	0.786	0.910	0.688	**0.470**
	fixation maps	**0.565**	**0.734**	**0.869**	**0.959**	**0.782**	0.430
	Overall Best Model	ConvNextNet [w/fixations]	DenseNet161 [Image only]	DenseNet161 [Image only]	VGG16 [w/fixations]	DenseNet161 [image only]	ConvNextNet [w/fixations]

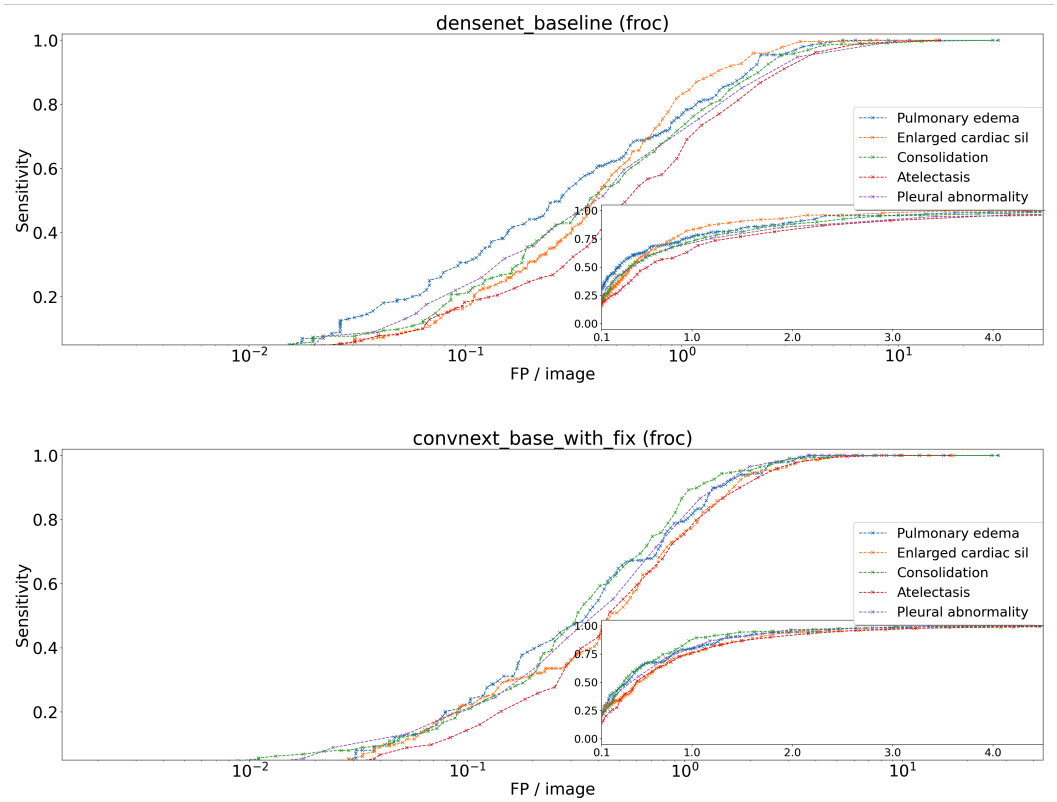

Figure 3. Comparison of FROC curves for DenseNet (image only) and ConvNext (with fixation masks), illustrating their performances in detecting various abnormalities in chest X-ray images. The curves demonstrate the similarities in performance between the two models, highlighting the effectiveness of both approaches for abnormality detection in CXRs.

To enhance the understanding of these results and determine whether fixation masks in reporting moments improve performance, an in-depth analysis of each model's confusion matrix was conducted (Table 2). Upon analyzing the number of True Positives (TPs), False Positives (FPs), and False Negatives (FNs), it becomes evident that ConvNextNet achieved superior performance when utilizing the fixation masks. ConvNextNet detected a higher number of TPs and fewer FNs than all other tested backbones. Furthermore, ConvNextNet registered fewer FPs than its baseline model. These findings are supported by the top-ranking performances exhibited by ConvNextNet in terms of average Recall and average Precision, demonstrating the effectiveness of incorporating fixation masks in the model. However, other backbones were close to ConvNextNet in terms of performance, such as EfficientNet using the images alone.

The researchers in Luís et al. [5] also implemented a Mask-RCNN-based approach to incorporate eye tracking data into an abnormality detection network's learning process in a technique analogous to the one described here, using the same datasets. However, contrary to this work, the authors of Luís et al. [5] obtained better results for instances of the model that used only the original CXR, further hinting at the usefulness of the selection of gaze associated with reporting moments included in models' learning processes,

Table 2. Performance Compare EyeXNet with different backbones using chest X-ray (CXR) images and radiologist fixation masks or just the CXR images. The table presents the number of True Positives (TPs), False Positives (FPs), and False Negatives (FNs); the Average Recall (mRecall); and the Average Precision (mPrecision) for each model. ConvNext, when using fixation masks, demonstrates the best performance in both Average Recall and Average Precision, highlighting its effectiveness in abnormality detection.

Backbone	Setting	#TP	#FP	#FN	mRecall	mPrecision
MobileNet	image only	690	**3387**	844	0.345	0.096
	fixation maps	**931**	6485	**603**	**0.448**	**0.111**
ResNet18	image only	685	**3014**	849	0.348	0.110
	fixation maps	**878**	5197	**656**	**0.448**	**0.124**
DenseNet161	image only	**849**	4920	**685**	0.396	0.105
	fixation maps	821	**3983**	713	**0.414**	**0.123**
EfficientNetB5	image only	**915**	8657	**619**	**0.491**	**0.126**
	fixation maps	819	**3710**	715	0.391	0.117
EfficientNetB0	image only	691	**2707**	843	0.325	0.103
	fixation maps	**842**	3755	**692**	**0.414**	**0.123**
ConvNextNet	image only	905	5145	629	0.457	0.138
	fixation maps	**945**	**4758**	**589**	**0.491**	**0.150**
VGG16	image only	**918**	6122	**616**	0.443	0.122
	fixation maps	888	**4228**	646	0.430	0.120
RegNet	image only	**945**	8008	**590**	**0.470**	0.095
	fixation maps	848	**4260**	686	0.430	**0.118**
	Overall Best Model	ConvNextNet [w/fixations]	EfficientNetB0 [Image only]	ConvNextNet [w/fixations]	ConvNextNet [w/fixations]	ConvNextNet [w/fixations]

Human Grounded Results

The qualitative analysis from the think-aloud experiment, involving two experienced radiologists evaluating the four best-performing and worst-performing chest X-ray images processed by EyeXNet using a ConvNext backbone and fixation masks, yielded the following findings:

Radiologist Assessment Variability and Discrepancies from Groundtruth Annotations. The evaluations provided by the two expert radiologists exhibited considerable deviations from the ground truth annotations in the REFLACX dataset. Although it is well documented in the literature that the assessment of X-ray images is subject to variability [48], it is worth noting that in only a few instances did the diagnoses made by our radiologists align with the ground truth.

Challenges in Image Interpretation due to External Devices and Limited Clinical Information. The radiologist faced considerable difficulty interpreting multiple images as the presence of various external devices such as tubes, surgical sutures, bone prostheses, and pacemakers complicated them. These devices suggested a complex medical history for the patients. Lacking access to clinical information and relying solely on the images, the radiologist's task of accurately assessing the images was further complicated. Some recent studies have already been proposed where they try to incorporate patients' clinical data in multimodal DL frameworks to improve diagnostic accuracy [49].

Challenges in Data Quality and Contextual Factors. The chest X-ray images that yielded the lowest performance from the classifier were deemed challenging to assess by our radiologists, primarily due to image quality issues: (1) The images appeared too dark, obscuring specific details; (2) the patients were not in the standard AP position but rather semi-erect or angled, making the detection of certain lesions difficult to evaluate; and (3) the absence of

comprehensive patient clinical history made it problematic for the radiologists to determine whether a lesion in the chest X-ray represented a previous or a current medical issue.

While our results confirm our initial hypothesis, we conducted various analyses to comprehend the diverse performances exhibited by the different backbones when utilizing fixation masks.

8. Discussion

In the existing literature, complete fixation masks (without distinguishing between fixations made during moments of silence and moments of reporting) are employed in multimodal DL architectures for abnormality detection or diagnosis classification in CXRs. However, no study has conclusively demonstrated a clear superiority advantage of using fixation masks.

Contrary to existing literature, we hypothesized that radiologists would likely focus more on the chest X-ray regions with abnormalities while reporting their findings, potentially enhancing system performance [17]. Our findings suggest that incorporating radiologists' fixation maps into EyeXNet with ConvNext may result in more True Positives and fewer False Negatives, ultimately leading to improved Recall and Precision. However, other backbones, such as EfficientNet or DenseNet, achieved comparable performances in terms of sensitivity at different false positive rates by relying solely on chest X-ray images. Indeed, ConvNextNet tends to generate more False Positives than other architectures, which accounts for its performance deficit in this metric. Nevertheless, ConvNextNet appears to be the most suitable backbone for abnormality detection when combined with fixation masks in EyeXNet.

To investigate these results further, we propose various technical factors, such as potential redundancy in fixation masks, which could explain DenseNet's architectural advantages and ConvNextNet's high generation of FPs. We also identify several data quality aspects, including discrepancies between the annotated abnormalities in the original REFLACX dataset and the abnormalities reported in the MIMIC-CXR medical records from the hospital information system.

8.1. Redundancy in Fixation Masks

It is possible that the fixation masks used in the study do not provide significant additional information beyond what is already present in the chest X-ray images, leading to only marginal improvements when combined with the images. DenseNet's inherent ability to extract complex features from the images alone may be sufficient for the task, rendering the fixation masks less impactful. Moreover, fixation masks contain considerable noise, such as multiple revisits to a specific region. Radiologists at different stages of their careers also exhibit varying fixation patterns; less experienced radiologists tend to make more fixations compared to experts [50,51]. Figure 4 provides an example illustrating the noisiness of fixation masks.

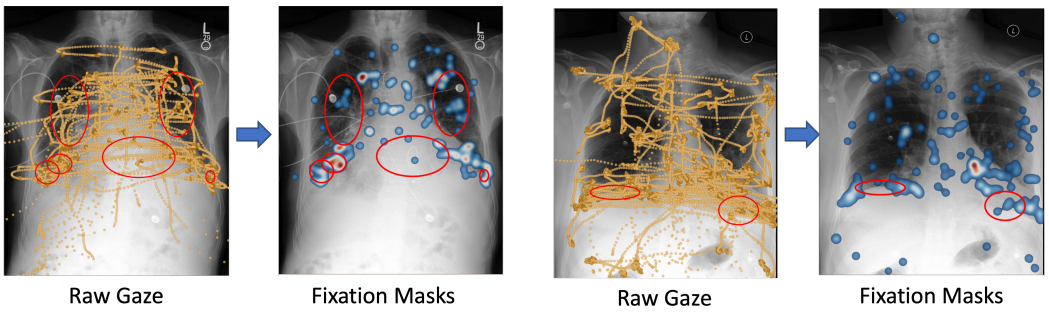

Figure 4. Two distinct chest X-ray images, each annotated with red ellipses by different radiologists. These patterns reveal that the raw gaze data are intricate and noisy, highlighting the complexity of interpreting these visual cues.

It is also plausible that the fixation masks do not provide significant additional information beyond what is already present in the chest X-ray images. However, they contribute to our study's decrease in FNs and the increase in TPs. Moreover, several studies in the literature have reported substantial reductions in FPs when employing fixation masks [52–55].

In the case of DenseNet, this neural network architecture has been specifically designed to learn complex features from image data efficiently. Its ability to adeptly capture these features from chest X-ray images may already be sufficient for identifying abnormalities, reducing the added value of incorporating fixation masks. As a result, the redundancy in the fixation masks may contribute to the observed performance differences between models using images alone and those that also include fixation masks. DenseNet's inherent proficiency in feature extraction might render the additional data from fixation masks less impactful on the model's performance.

8.2. Model Architecture.

DenseNet, ConvNext, and EfficientNet feature unique architectural designs that impact their capacity to learn and extract features from the input data. DenseNet's dense connections excel at reusing features and enhancing gradient flow, which could lead to superior FROC performance without the need for redundant information from fixation maps. DenseNet achieves this by connecting each layer to every other layer in a feed-forward manner, allowing for the more efficient use of parameters and better information flow during training.

Conversely, ConvNext and EfficientNet may possess architectures specifically optimized for attaining higher Recall rates, prioritizing identifying true positive cases. ConvNext utilizes a hierarchical structure composed of multiple convolutional stages, with each stage successively increasing the network's depth. This hierarchical design enables the model to capture features at various scales, which can be beneficial for detecting abnormalities of differing sizes and shapes. Within this hierarchical configuration, feature maps may offer a more comprehensive feature extraction process as our results indicate that ConvNext performs better when using both images and fixation masks compared to relying solely on images.

EfficientNet, on the other hand, employs a compound scaling method that simultaneously adjusts the model's depth, width, and resolution. By scaling all three dimensions together, EfficientNet can achieve a better balance between model complexity and accuracy, which may contribute to its heightened performance in terms of Recall. However, performance improvement is not observed when fusion with fixation masks is attempted, as shown in Tables 1 and 2. This could be due to EfficientNet's already optimized architecture, which efficiently extracts essential features from the images, potentially making the additional information from fixation masks less impactful. Additionally, noise and redundancy in fixation masks may hinder the model's ability to effectively capitalize on the supplementary data, thus not contributing significantly to overall performance.

8.3. Data Quality Challenges.

To gain a deeper understanding of the data quality challenges identified by our experienced radiologists, we examined the original medical reports associated with each chest X-ray. We then compared the labels extracted using NegBio, a natural language processing tool specifically designed for this task, to the annotations present in the REFLACX dataset. Our objective was to determine the number of images with annotations that deviated from the original medical reports.

Considering the five lesion types that EyeXNet predicted, we discovered that in over 30% of cases, the annotations on the chest X-rays diverged from the lesions mentioned in the hospital medical reports. The discrepancy rate for Pulmonary Edema was notably lower, only 16% .

9. Conclusions

This study explored the effectiveness of incorporating radiologists' fixation maps into DL models to detect abnormalities in chest X-ray images. We compared the performance of different backbones, including ConvNext, DenseNet, and EfficientNet, combined with EyeXNet and various fixation maps. Our findings indicate that using fixation maps in EyeXNet with ConvNext can improve Recall and Precision by increasing the number of true positives and reducing the number of false negatives. However, other backbones, such as EfficientNet and DenseNet, achieved similar performance using only CXR images, suggesting that fixation maps might not always contribute significant additional information. Overall, integrating the fixation masks in a multimodal DL architecture outperformed the baseline model in 5 out of 8 experiments regarding average Recall and 6 out of 8 in terms of average Precision.

We conclude that the effectiveness of incorporating fixation masks into a multimodal DL architecture depends on the correlation between the fixation masks and the generated proposal regions. When the fixation masks do not align well with the proposal regions, the information from the image feature maps extracted by the backbone dominates during the fusion process. As a result, the performance is comparable to the baseline Mask R-CNN, which considers only the image input; however, in cases where the fixation masks correlate well with the proposal regions, the fusion process benefits from the additional information provided by the fixation masks. This leads to improved lesion detection for CXR images. Consequently, integrating fixation masks, which capture radiologists' classification patterns, into a multimodal DL architecture proves advantageous for enhancing the overall lesion-detection task. These observations highlight the potential benefits of leveraging fixation masks in multimodal DL architectures and emphasize the importance of considering the alignment between fixation masks and proposal regions when designing and evaluating such architectures for CXR image analysis.

Author Contributions: Conceptualization, C.H., A.L., C.O. and C.M.; Methodology, C.H., A.L. and C.M.; Software, C.H. and A.L.; Validation, C.H., A.L., J.N., I.B.N. and S.C.S.; Formal analysis, C.H.; Investigation, C.H., A.L., J.N., I.B.N., S.C.S., C.O., J.J. and C.M.; Resources, I.B.N. and S.C.S.; Data curation, C.H.; Writing—original draft, C.H., A.L., J.N. and C.M.; Writing—review & editing, C.H., C.O., J.J. and C.M.; Visualization, C.H.; Supervision, C.O., J.J. and C.M.; Project administration, J.J.; Funding acquisition, J.J. and C.M. All authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: The work reported in this article was partially supported under the auspices of the UNESCO Chair on AI & VR and by national funds through Fundação para a Ciência e a Tecnologia with references DOI:10.54499/UIDB/50021/2020, DOI:10.54499/DL57/2016/CP1368/CT0002 and 2022.09212.PTDC (XAVIER project).

Data Availability Statement: All of the analyses, the code, and the EyeXNet framework are publicly available on GitHub: https://github.com/ChihchengHsieh/MIMIC-Eye-applications, accessed on 8 May 2024.

Acknowledgments: The authors thank the reviewers for their constructive and helpful comments.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Parker, M.S.; Chasen, M.H.; Paul, N. Radiologic Signs in Thoracic Imaging: Case-Based Review and Self-Assessment Module. *Am. J. Roentgenol.* **2009**, *192*, S34–S48. [CrossRef] [PubMed]
2. Moses, D.A. Deep learning applied to automatic disease detection using chest X-rays. *J. Med. Imaging Radiat. Oncol.* **2021**, *65*, 498–517. [CrossRef] [PubMed]
3. Bigolin Lanfredi, R.; Zhang, M.; Auffermann, W.F.; Chan, J.; Duong, P.A.T.; Srikumar, V.; Drew, T.; Schroeder, J.D.; Tasdizen, T. REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Sci. Data* **2022**, *9*, 350. [CrossRef] [PubMed]

4. Karargyris, A.; Kashyap, S.; Lourentzou, I.; Wu, J.T.; Sharma, A.; Tong, M.; Abedin, S.; Beymer, D.; Mukherjee, V.; Krupinski, E.A.; et al. Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. *Sci. Data* **2021**, *8*, 92. [CrossRef]
5. Luís, A.; Hsieh, C.; Nobre, I.B.; Sousa, S.C.; Maciel, A.; Moreira, C.; Jorge, J. Integrating Eye-Gaze Data into CXR DL Approaches: A Preliminary study. In Proceedings of the 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Shanghai, China, 25–29 March 2023.
6. Pershin, I.; Mustafaev, T.; Ibragimova, D.; Ibragimov, B. Changes in Radiologists' Gaze Patterns Against Lung X-rays with Different Abnormalities: A Randomized Experiment. *J. Digit. Imaging* **2023**, *36*, 767–775. [CrossRef] [PubMed]
7. Castner, N.; Kuebler, T.C.; Scheiter, K.; Richter, J.; Eder, T.; Hüttig, F.; Keutel, C.; Kasneci, E. Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing. In Proceedings of the ACM Symposium on Eye Tracking Research and Applications, Stuttgart, Germany, 2–5 June 2020; pp. 1–10.
8. Saporta, A.; Gui, X.; Agrawal, A.; Pareek, A.; Truong, S.Q.; Nguyen, C.D.; Ngo, V.D.; Seekins, J.; Blankenberg, F.G.; Ng, A.Y.; et al. Benchmarking saliency methods for chest X-ray interpretation. *Nat. Mach. Intell.* **2022**, *4*, 867–878. [CrossRef]
9. Neves, J.; Hsieh, C.; Nobre, I.B.; Sousa, S.C.; Ouyang, C.; Maciel, A.; Duchowski, A.; Jorge, J.; Moreira, C. Shedding light on ai in radiology: A systematic review and taxonomy of eye gaze-driven interpretability in deep learning. *Eur. J. Radiol.* **2024**, *172*, 111341. [CrossRef] [PubMed]
10. Wang, S.; Ouyang, X.; Liu, T.; Wang, Q.; Shen, D. Follow My Eye: Using Gaze to Supervise Computer-Aided Diagnosis. *IEEE Trans. Med. Imaging* **2022**, *41*, 1688–1698. [CrossRef] [PubMed]
11. Saab, K.; Hooper, S.M.; Sohoni, N.S.; Parmar, J.; Pogatchnik, B.; Wu, S.; Dunnmon, J.A.; Zhang, H.R.; Rubin, D.; Ré, C. Observational Supervision for Medical Image Classification Using Gaze Data. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021, Strasbourg, France, 27 September–1 October 2021; de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 603–614.
12. Watanabe, A.; Ketabi, S.; Namdar, K.; Khalvati, F. Improving disease classification performance and explainability of deep learning models in radiology with heatmap generators. *Front. Radiol.* **2022**, *2*, 991683. [CrossRef]
13. Nie, W.; Zhang, Y.; Patel, A. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations. *arXiv* **2018**, arXiv:1805.07039.
14. Agnihotri, P.; Ketabi, S.; Khalvati, F. Using Multi-modal Data for Improving Generalizability and Explainability of Disease Classification in Radiology. *arXiv* **2022**, arXiv:2207.14781.
15. Qi, Z.; Khorram, S.; Li, F. Visualizing Deep Networks by Optimizing with Integrated Gradients. *arXiv* **2019**, arXiv:1905.00954.
16. Lanfredi, R.B.; Arora, A.; Drew, T.; Schroeder, J.D.; Tasdizen, T. Comparing radiologists' gaze and saliency maps generated by interpretability methods for chest X-rays. *arXiv* **2021**, arXiv:2112.11716.
17. Moreira, C.; Alvito, D.; Sousa, S.C.; Nobre, I.B.; Ouyang, C.; Kopper, R.; Duchowski, A.; Jorge, J. Comparing Visual Search Patterns in Chest X-ray Diagnostics. In Proceedings of the ACM on Computer Graphics and Interactive Techniques (ETRA), Tübingen, Germany, 29 May–3 June 2023.
18. Shneiderman, B. *Human-Centered AI*; Oxford University Press: Oxford, UK, 2022.
19. El Kafhali, S.; Alzubaidi, L.; Al-Sabaawi, A.; Bai, J.; Dukhan, A.; Alkenani, A.H.; Al-Asadi, A.; Alwzwazy, H.A.; Manoufali, M.; Fadhel, M.A.; et al. Towards Risk-Free Trustworthy Artificial Intelligence: Significance and Requirements. *Int. J. Intell. Syst.* **2023**, *2023*, 4459198. [CrossRef]
20. Stöger, K.; Schneeberger, D.; Holzinger, A. Medical artificial intelligence: The European legal perspective. *Commun. ACM* **2021**, *64*, 34–36. [CrossRef]
21. Holmqvist, K.; Andersson, R. *Eye-Tracking: A Comprehensive Guide to Methods, Paradigms and Measures*; Oxford University Press: Oxford, UK, 2017.
22. Nodine, C.F.; Kundel, H.L. Using eye movements to study visual search and to improve tumor detection. *RadioGraphics* **1987**, *7*, 1241–1250. [CrossRef] [PubMed]
23. Duchowski, A.T. *Eye Tracking Methodology: Theory and Practice*; Springer: Berlin/Heidelberg, Germany, 2003.
24. Rong, Y.; Xu, W.; Akata, Z.; Kasneci, E. Human Attention in Fine-grained Classification. In Proceedings of the 32nd British Machine Vision Conference (BMVC), Online, 22–25 November 2021.
25. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
26. Woo, S.; Park, J.; Lee, J.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
28. Carmody, D.; Nodine, C.; Kundel, H. Finding lung nodules with and without comparative visual scanning. *Percept. Psychophys.* **1981**, *29*, 594–598. [CrossRef]
29. Krupinski, E.A. Visual scanning patterns of radiologists searching mammograms. *Academic radiology* **1996**, *3 2*, 137–44. [CrossRef]
30. Hu, C.H.; Kundel, H.L.; Nodine, C.F.; Krupinski, E.A.; Toto, L.C. Searching for bone fractures: A comparison with pulmonary nodule search. *Acad. Radiol.* **1994**, *1*, 25–32. . [CrossRef] [PubMed]
31. Kendall, A.; Gal, Y.; Cipolla, R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. *arXiv* **2017**, arXiv:1705.07115.

32. Matthew Zeiler, D.; Rob, F. Visualizing and understanding convolutional neural networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
33. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
35. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
36. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning. PMLR, Long Beach, CA, USA, 10–15 June 2019; pp. 6105–6114.
37. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
38. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 568–576
39. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.; He, K.; Dollár, P. Designing network design spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10428–10436.
40. Hsieh, C.; Ouyang, C.; Nascimento, J.C.; Pereira, J.a.; Jorge, J.; Moreira, C. MIMIC-Eye: Integrating MIMIC Datasets with REFLACX and Eye Gaze for Multimodal Deep Learning Applications (version 1.0.0). *PhysioNet* **2023**. [CrossRef]
41. Johnson, A.E.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.Y.; Mark, R.G.; Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **2019**, *6*, 317. [CrossRef] [PubMed]
42. Johnson, A.E.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.y.; Mark, R.G.; Horng, S. MIMIC-CXR Database (version 2.0.0). *PhysioNet* **2019**. . [CrossRef]
43. Johnson, A.E.; Pollard, T.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.y.; Peng, Y.; Lu, Z.; Mark, R.G.; Berkowitz, S.J.; Horng, S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv* **2019**, arXiv:1901.07042.
44. Johnson, A.; Bulgarelli, L.; Pollard, T.; Celi, L.A.; Mark, R.; Horng IV, S. MIMIC-IV-ED (Version: 2.2). *PhysioNet* **2023**. . [CrossRef]
45. Johnson, A.; Bulgarelli, L.; Pollard, T.; Horng, S.; Celi, L.A.; Mark, R. MIMIC-IV (version 2.2). *Physionet* **2023**. [CrossRef]
46. Goldberger, A.; Amaral, L.; Glass, L.; Hausdorff, J.; Ivanov, P.; Mark, R.; Mietus, J.; Moody, G.; Peng, C.; Stanley, H. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2020**, *101*, e215–e220. [CrossRef] [PubMed]
47. Chakraborty, D.P. A brief history of free-response receiver operating characteristic paradigm data analysis. *Acad. Radiol.* **2013**, *20*, 915–919. [CrossRef]
48. Ganesan, A.; Alakhras, M.; Brennan, P.C.; Mello-Thoms, C. A review of factors influencing radiologists' visual search behaviour. *J. Med. Imaging Radiat. Oncol.* **2018**, *62*, 747–757. [CrossRef] [PubMed]
49. Hsieh, C.; Nobre, I.B.; Sousa, S.C.; Ouyang, C.; Brereton, M.; Nascimento, J.C.; Jorge, J.; Moreira, C. MDF-Net: Multimodal Dual-Fusion Network for Abnormality Detection using CXR Images and Clinical Data. *arXiv* **2023**, arXiv:2302.13390.
50. Borys, M.; Plechawska-Wójcik, M. Eye-tracking metrics in perception and visual attention research. *EJMT* **2017**, *3*, 11–23.
51. Harezlak, K.; Kasprowski, P. Application of eye tracking in medicine: A survey, research issues and challenges. *Comput. Med. Imaging Graph.* **2018**, *65*, 176–190. [CrossRef] [PubMed]
52. Mall, S.; Brennan, P.C.; Mello-Thoms, C.R. Modeling visual search behavior of breast radiologists using a deep convolution neural network. *J. Med. Imaging* **2018**, *5*, 035502. [CrossRef]
53. Mall, S.; Brennan, P.C.; Mello-Thoms, C. Can a Machine Learn from Radiologists' Visual Search Behaviour and Their Interpretation of Mammograms—A Deep-Learning Study. *J. Digit. Imaging* **2019**, *32*, 746–760. [CrossRef] [PubMed]
54. Mall, S.; Brennan, P.; Mello-Thoms, C. Fixated and Not Fixated Regions of Mammograms, A Higher-Order Statistical Analysis of Visual Search Behavior. *Acad. Radiol.* **2017**, *24*, 442–455. [CrossRef]
55. Khosravan, N.; Celik, H.; Turkbey, B.; Jones, E.C.; Wood, B.; Bagci, U. A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning. *Med. Image Anal.* **2019**, *51*, 101–115. [CrossRef] [PubMed]