European Journal of Radiology 172 (2024) 111341

Contents lists available at ScienceDirect

European Journal of Radiology

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

Review

Shedding light on ai in radiology: A systematic review and taxonomy of eye gaze-driven interpretability in deep learning

José Neves [a,*], Chihcheng Hsieh [b], Isabel Blanco Nobre [c], Sandra Costa Sousa [c], Chun Ouyang [b], Anderson Maciel [a], Andrew Duchowski [d], Joaquim Jorge [a], Catarina Moreira [e]

[a] Instituto Superior Técnico / INESC-ID, University of Lisbon, Portugal
[b] School of Information Systems, Queensland University of Technology, Australia
[c] Department of Imageology, Lusíadas Knowledge Center, Portugal
[d] Clemson University, South Carolina, USA
[e] Human Technology Institute, University of Technology Sydney, Australia

ARTICLE INFO

Keywords:
Deep learning
Eye tracking
Multimodal fusion
Explainable AI

ABSTRACT

X-ray imaging plays a crucial role in diagnostic medicine. Yet, a significant portion of the global population lacks access to this essential technology due to a shortage of trained radiologists. Eye-tracking data and deep learning models can enhance X-ray analysis by mapping expert focus areas, guiding automated anomaly detection, optimizing workflow efficiency, and bolstering training methods for novice radiologists. However, the literature shows contradictory results regarding the usefulness of eye-tracking data in deep-learning architectures for abnormality detection. We argue that these discrepancies between studies in the literature are due to (a) the way eye-tracking data is (or is not) processed, (b) the types of deep learning architectures chosen, and (c) the type of application that these architectures will have. We conducted a systematic literature review using PRISMA to address these contradicting results. We analyzed 60 studies that incorporated eye-tracking data in a deep-learning approach for different application goals in radiology. We performed a comparative analysis to understand if eye gaze data contains feature maps that can be useful under a deep learning approach and whether they can promote more interpretable predictions. To the best of our knowledge, this is the first survey in the area that performs a thorough investigation of eye gaze data processing techniques and their impacts in different deep learning architectures for applications such as error detection, classification, object detection, expertise level analysis, fatigue estimation and human attention prediction in medical imaging data. Our analysis resulted in two main contributions: (1) taxonomy that first divides the literature by task, enabling us to analyze the value eye movement can bring for each case and build guidelines regarding architectures and gaze processing techniques adequate for each application, and (2) an overall analysis of how eye gaze data can promote explainability in radiology.

1. Introduction

I have always been convinced that the only way to get artificial intelligence to work is to do the computation in a way similar to the human brain. Geoff Hinton

The rapid advancements in medical imaging technology have revolutionized healthcare over the past decades, significantly aiding in disease diagnosis and treatment planning. According to the Radiology Society of North America (RSNA), there exists a shortage of radiologists that is accelerated by the strain of an ageing population, the attrition associated with COVID-19, and a disproportionate growth rate of radiology trainees, which presents a substantial challenge to the sustainability of global healthcare infrastructures [1,2]. Projections from the World Health Organization (WHO) indicate that by 2050, the proportion of individuals over 60 will reach 22% of the world population, almost doubling the necessity for imaging studies relative to 2015 [3]. In the United States alone, between 2012 and 2019, the expansion of the

* Corresponding author at: Instituto Superior Técnico / INESC-ID, University of Lisbon, Portugal.
E-mail addresses: jose.s.neves@tecnico.ulisboa.pt (J. Neves), chihcheng.hsieh@hdr.qut.edu.au (C. Hsieh), isabel.blanco.nobre@lusiadas.pt (I.B. Nobre), sandra.costa.sousa@lusiadas.pt (S.C. Sousa), couyang@qut.edu.au (C. Ouyang), anderson.maciel@tecnico.ulisboa.pt (A. Maciel), aduchow@clemson.edu (A. Duchowski), jorgej@tecnico.ulisboa.pt (J. Jorge), Catarina.pintomoreira@uts.edu.au (C. Moreira).

https://doi.org/10.1016/j.ejrad.2024.111341
Received 31 October 2023; Received in revised form 4 January 2024; Accepted 25 January 2024
Available online 1 February 2024

Medicare beneficiary population surpassed the growth of the diagnostic radiology workforce by approximately 5% [4].

1.1. Deep learning and the black-box problem

Artificial intelligence (AI), especially deep learning (DL) approaches, have demonstrated significant potential in tackling various challenges within radiology [5,6]. According to LeCun [7], there are six fundamental aspects of clinical healthcare that AI has the potential to enhance significantly. These aspects include workflow efficiency, reduced reading time, lowered dose and contrast agents, early disease detection, enhanced diagnostic accuracy, and personalized diagnostics. While DL models have demonstrated remarkable proficiency in various tasks [5], their internal decision-making process remains largely opaque, presenting a so-called 'black-box' problem where the processes leading to a model's decisions remain inaccessible for human scrutiny [8,9]. This opacity often leads to reluctance among practitioners to unconditionally accept outputs from these models without understanding their underpinning logic [10]. This creates a significant hurdle for the broader acceptance and deployment of AI-based technologies, as they are prone to well-established biases, including automation bias, that can result in errors, eroding trust in model predictions and hampering their uptake [11,12]. Hence, there is a clear demand for more human-centric technologies to bridge the gap between DL systems and human understanding, allowing for integrating human contextual knowledge [13,14] [15].

1.2. Towards eye gaze-driven interpretability

A promising avenue for enhancing the performance and explainability of these AI models lies within radiologists' gaze data. These data potentially provide insights into the cognitive processes of radiologists, shedding light on their mental models built in a multimodal environment consisting of patients' medical information, imaging data, and global and local image features. These insights could guide the development of human-centred AI [16], more specifically, multimodal DL architectures that are not purely data-driven but reflect human cognitive processes.

According to Watanabe et al. [17], eye-gaze data from radiologists has the potential to offer critical insights into human decision-making, emphasizing areas of focus, the sequence of observations, the duration and frequency of gaze on critical regions, and implicit expert knowledge—all contributing to more comprehensible AI decisions. By integrating eye-gaze data into a DL model, one can build models that mirror human diagnostic processes, potentially making them more understandable. However, the role of gaze data in deep learning for lesion detection in X-rays remains unexplored. Most works in the literature explore the use of eye gaze together with chest X-ray images for general classification tasks, where the model aims to predict a global label stating whether the patient has a certain lesion or not. Some works try to use explainable AI techniques, such as the application of saliency map methods, to help identify what features the model focuses on when making a decision. However, the recent work of Saporta et al. [18] found that when benchmarking human experts against saliency map methods in radiology, they discovered that saliency maps usually performed worse than the human experts in identifying the relevant regions containing lesions in an X-ray, and this performance gap was bigger when the lesions were small and with complex shapes. Saliency map methods were always very greedy in terms of the localized regions. They could not accommodate the irregularities of the shapes of the lesions, which is a crucial factor in the identification of tumours that usually have spiked and non-round shapes. It remains an open research question whether eye gaze data can improve these saliency maps in global classificattasks or can help guide the learning process in object detection better DL models to identify the regions with lesions. To address this research gap, we conducted a systematic literature review focusing on the impact that eye-tracking data processing techniques and the inclusion of eye-tracking data on multimodal deep learning models can have on the performance and explainability of AI in radiology solutions.

1.3. Contributions

Our contributions to advancing the integration of eye-tracking data in automated radiology through multimodal deep learning are delineated as follows:

- We conducted a comprehensive literature review in line with PRISMA guidelines, centring on the advancements in processing eye-tracking data. This study sheds light on the enhancements in performance and the growing prominence of interpretability within multimodal deep-learning frameworks that incorporate gaze data.
- We introduced a detailed taxonomy to classify various dimensions of gaze data utilization in AI-driven radiology. This structure aims to substantially elevate the understanding of the potential regions where gaze can amplify AI models in terms of their efficiency and interpretability.
- By performing a comparative analysis of diverse gaze-driven methodologies across various tasks, we have distilled best practices tied to specific clinical scenarios. This comparison also offers insights into promising research avenues, emphasizing the crafting of gaze-integrated radiology deep learning models optimally tailored for clinical implementation.

2. Background concepts and related works

In this section, we provide a detailed overview of existing surveys related to utilizing eye-gaze data in deep learning models within the context of radiology. Before proceeding, we must present important concepts of saccades, fixations, fixation/attention maps, and saliency maps.

2.1. Background concepts

Fig. 1 presents a visual overview of the differences between saccades, fixations, and attention heatmaps.

Saccades are rapid eye movements that occur when we shift our gaze from one point of interest to another. This common ocular motion happens many times a day and is primarily involved when reading, scanning a scene, or tracking an object. They are the fastest movements produced by the human body and serve to direct the fovea - the part of the eye with the highest visual acuity - to objects of interest [19].

Fixations refer to periods when our eyes remain relatively still, and the gaze is focused on a particular point in our visual field. These are moments when the eye pauses during saccades to acquire information and are essential for tasks that require detailed visual information. The duration of fixations can provide insights into cognitive processing. For instance, longer fixations could indicate areas of interest or complexity that require more cognitive processing [19].

Fixation or attention maps are visual representations of where viewers focus their gaze or attention when looking at a stimulus, such as an image or a scene. These maps are usually generated by aggregating the fixations of multiple viewers, producing a heatmap that indicates the areas where most viewers concentrated their gaze. They provide a measure of visual attention distribution over a given stimulus. This paper will refer to heatmaps generated using radiologists' gaze data as **attention or fixation maps**.

Saliency maps. In deep learning, heatmaps can be used to interpret what parts of an image a model is "paying attention to" or finding significant for making predictions. These maps are valuable for understanding model behavior, and in the field of radiology, they can also serve as a useful tool for training and assessment. This study will refer to heatmaps generated by deep learning models as **saliency maps.**

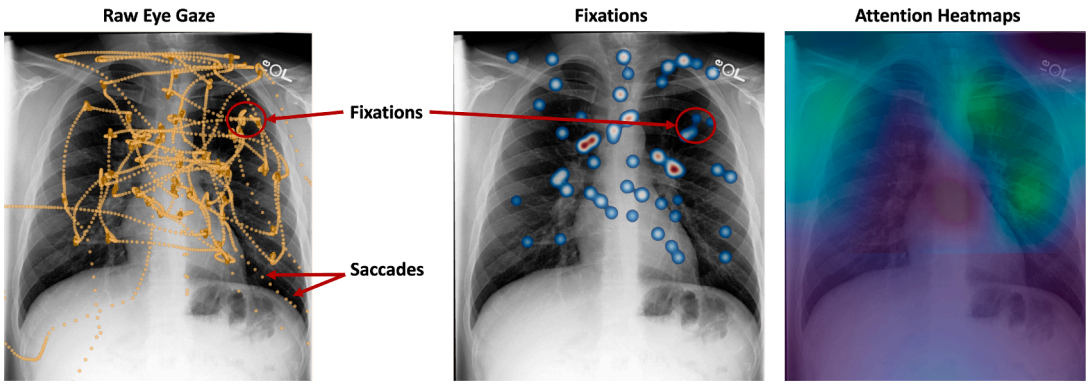

Fig. 1. Difference between saccades, fixations, and attention heatmaps.

2.2. Related work

In Gijp et al. [20], the authors identify the visual search patterns associated with high performance in terms of the time needed first to fixate abnormalities. They found that expert search is characterized by a global-focal search pattern with specific search patterns associated with analyzing chest X-rays and Computed Tomography(CT) scans. However, their investigation did not yield evidence to support that teaching typical visual search strategies could positively affect student performance.

In a comprehensive review, Brunyé et al. [21] examine the use of eye-tracking technology in medical image interpretation. They detail the perceptual and cognitive challenges physicians encounter during diagnosis and underscore the role of eye-tracking in understanding the complex interplay of visual perception, memory retrieval, problem-solving, and decision-making involved in diagnosis. Their work distinguishes between novices and experts in visual interpretation, highlighting the influence of heuristics and biases. Although the visual interpretations could provide insights into explainable mechanisms, the authors did not explore the importance of integrating gaze data into DL models for radiology.

In an in-depth study on visual search in breast radiology, Gandomkar and Mello-Thoms [22] investigated the interpretation of breast images, emphasizing the limitations imposed by human factors and the resultant diagnostic errors, including satisfaction of search, incorrect background sampling, and initial impressions. However, while they discuss the interaction between radiologists and computerized diagnostic tools, they do not investigate how eye gaze data can be integrated with DL. In contrast, our proposed study aims to bridge these gaps, exploring integrating eye-tracking data with AI tools to improve diagnostic accuracy.

Research by Lévêque et al. [23] provides a comprehensive review of eye-tracking studies within medical imaging, covering a broad spectrum of applications, including expertise identification, training development, and understanding visual search patterns. They also conducted their eye-tracking study on screening mammograms interpreted by experienced radiologists, investigating the feasibility of predicting visual attention using computational models. However, their findings suggest that current computer-generated saliency maps are inadequate for accurately predicting human gaze behavior, highlighting the need for significant improvement before they can be employed for gaze-based training.

The study by Arthur and Sun [24] undertakes an in-depth review of eye-tracking use in evaluating radiological image interpretation by medical professionals. It explores four themes: competency assessment, educational tools, visual search behavior, and assistive aid evaluations. The paper highlights that while most studies emphasize competency assessment, the results regarding interpretation speed and eye-metrics are conflicting and require more investigation. Both Lévêque et al. [23,24] studies lack a focus on gaze data in DL techniques or eye-tracking's role in model explainability.

To our understanding, our survey is the first to conduct a comprehensive literature review and a comparative analysis of features present in gaze data that have the potential to lead to deep learning models better suited for aiding clinical practice. We delve into optimal methods for extracting and integrating these data into deep learning models. Furthermore, we explore the potential of eye tracking as a tool to enhance explainability within radiology. In the next section, we explore our implementation of PRISMA, the technique we used to obtain a relevant set of works representative of the existing literature about eye-tracking data integration in AI-powered radiology solutions.

3. Systematic literature review

Our systematic literature review adheres to the guidelines set forth by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [25] framework, providing a standard approach to gathering and distilling information from existing research relevant to our research questions.

In alignment with PRISMA checklist, our review process comprises several steps: (1) formulation of research questions; (2) outlining the literature search strategy and process; (3) extracting, organizing, and analyzing pertinent literature data such as title, abstract, author keywords, and year, specifically related to eye gaze data in multimodal deep learning architectures in radiology; (4) and finally, we identified potential biases and limitations in our review methodology. Fig. 2 illustrates the PRISMA checklist and the results of the four-stage process applied to each database queried: Scopus, IEEE eXplore, PubMed, and Web of Science. The initial search yielded 180 research papers. After removing duplicates and applying inclusion and exclusion criteria, the final selection included 60 papers for detailed analysis. Sections 3.1–3.4 provide details for each step.

In the realm of healthcare, the standard protocol is to register systematic literature reviews in PROSPERO, an international database designed to foster transparency and avoid redundant research in health and social sciences. Our study, employing the widely-cited PRISMA methodology, focuses on the role of eye tracking in enhancing deep-learning models for chest X-ray analysis.

3.1. Research questions

The contemporary literature lacks a concrete answer regarding eye movement data's role and how to best integrate it into AI-powered radiology solutions. Therefore, we are interested in exploring the existing techniques regarding the pre-processing of gaze data, its integration into deep learning architectures, and the contribution to the emergence of interpretability in the model it originates. Thus, we aim to answer the following research questions:

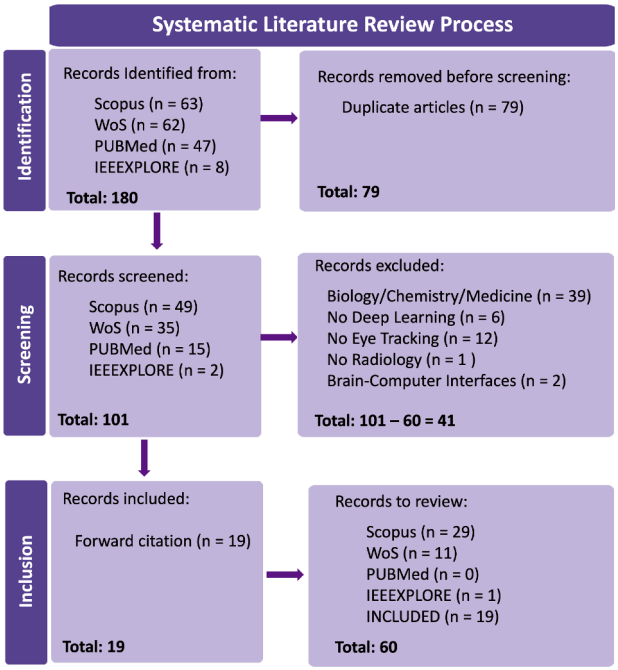

Fig. 2. This figure illustrates the distinct phases of the PRISMA framework followed in our systematic literature review, detailing the process of surveying 60 scholarly articles.

- RQ1 [Structure]: What architectures and fusion techniques are available to integrate eye-tracking data into deep learning solutions to localize and predict lesions?
- RQ2 [Pre-Processing]: How are eye-tracking data pre-processed before being incorporated into multimodal deep learning architectures?
- RQ3 [Interpretability]: How can eye gaze data promote explainability in multimodal deep learning architectures?

3.2. Search process

The literature search was conducted systematically and comprehensively to identify all relevant articles on deep learning architectures that use eye-tracking data. Three people conducted this search process to secure an unbiased selection of works. The databases used for the search included *PubMed*, *Scopus*, *IEEE Xplore Digital Library* and *Web of Science*.We used the following search query in our literature review process:
The terms ”interpretability” or ”explainability” were deliberately excluded from our keyword list to avoid an overly narrow result set. Our search strategy was designed to encompass all pertinent studies that delve into multimodal fusion DL architectures that combine eye-tracking data and X-ray images. Following this collection, our review explored how such a multimodal setup could foster the generation of human-like saliency maps, thereby enhancing the level of explainability. Fig. 2 summarizes our search process.

During our initial screening process, we collated 180 articles from multiple databases, specifically 63 from Scopus, 62 from Web of Science, 47 from PubMed, and 8 from IEEEXplore. After carefully reviewing these gathered articles and removing duplicates, our screening phase concluded with 101 articles.

3.3. Inclusion and exclusion criteria

Our research strategies targeted articles published between 2012 to early 2023. The selection of 2012 as the commencement point stems from its significance in the field of deep learning; it marks the year when AlexNet triumphed in the ImageNet competition [26], thereby catalyzing the subsequent proliferation of research on deep convolutional neural networks. We also considered papers published either in journals or conferences, and we excluded all conference editorial papers and books since they are not subjected to peer review. Table 1 presents our inclusion and exclusion criteria.

Our keywords pulled 39 papers from biology literature proposing various deep-learning approaches to Glaucoma detection. These, however, were excluded due to their lack of alignment with our focus on multimodal Deep Learning, merging eye tracking and radiology. In addition, we removed 12 papers that lacked eye-tracking data and two papers from the Brain-Computer Interfaces area.

3.4. Risk of bias

Like any task involving human judgment, the process of identifying pertinent research is subject to cognitive biases. In conducting this systematic review, we recognize that our decision to limit the search to three databases (Scopus, Web of Science, IEEE, and PubMed) may have led to the omission of some articles. Additional databases like Google Scholar and SpringerLink could have broadened our search. We also did not mine the references from the gathered papers to enhance our search, as the dataset of documents we collected was already substantial, and further extraction would have significantly increased the complexity of the analysis that was performed. Lastly, our search was confined to specific keywords to find papers of interest, which could have inadvertently narrowed our scope, possibly overlooking other relevant articles.

3.5. Word co-occurrence analysis

This literature review aims to examine the impact of eye gaze data on the efficacy of multimodal deep learning architectures in various tasks and explore their role in augmenting explainability. For an exhaustive understanding of the content within the retrieved papers, we conducted a word co-occurrence analysis of the titles, abstracts, and keywords from the final article list, resulting in a bibliometric network that presented the interplay between the different keywords. We used the graphical capabilities of *VOS Viewer*[1], a software tool specifically designed for building and visualizing bibliometric networks, to showcase the results.

Fig. 3 presents the resultant network as a density map, with distinct colors indicating varying publication years. It becomes evident that while our search process began with papers from 2012, the final list predominantly included works from 2017 through 2023. This trend is consistent with the rise of interpretability concerns circa 2016 due to the General Data Protection Regulation (GDPR), leading to a marked shift in AI literature from purely data-driven approaches to those considering interpretability and explainability. (see Fig. 4).

Earlier papers incorporating eye tracking within multimodal deep learning architectures were primarily Computer Aided Diagnostic Systems (CAD) applications, focusing mainly on nodule detection in mammograms. These studies employed eye tracking to inform about radiologists' search patterns and strategies. A noticeable emergence of topics of eye-tracking and interpretability in deep learning can be seen from 2020 onwards.

Interpretability in this context is largely connected with saliency maps—heatmaps derived from Deep Learning systems indicating neuron activation. These maps can reveal which system regions are predominantly engaged in prediction formation. Similar techniques have been proposed in the literature for eye tracking data, suggesting that by highlighting radiologists' fixated regions on an X-ray, the Deep Learning system can learn to identify human-like regions of interest. This explains the observed connections between interpretability, eye tracking, and ConvNet.

[1] https://www.vosviewer.com/

$$
\begin{aligned}
&(\text{"eye-tracking" } OR \text{ "eye tracking" } OR \text{ "gaze"}) \ AND \\
&(\text{"deep learning" } OR \text{ "multi-modal" } OR \text{ "multimodal" } OR \text{ "machine learning" } OR \\
&\text{"artificial intelligence" } OR \text{ "AI" } OR \text{ "neural network" } OR \text{ "neural networks"}) \ AND \\
&(radiol* \ OR \ radiog* \ OR \ xray* \ OR \ x\text{-}ray*)
\end{aligned}
$$

Table 1
Inclusion and exclusion criteria applied in our systematic literature review.

Inclusion	Exclusion
Papers that use DL and eye tracking in radiology	Papers about Biology/ Chemistry and Medicine
Papers about multimodal fusion in radiology	Papers that do not use eye-tracking data
Papers focused on X-rays or CT scans	Papers that do not use deep learning
Papers citing eye tracking datasets in radiology	Papers that are not about radiology
Papers about DL and visual search patterns in radiology	Papers about Brain-Computer Interfaces
Papers published after 2012 (inclusive)	Papers about editorial notes in journals/ conferences

These heatmaps can also facilitate the identification of potential biases, promoting trust and consequently leading to topics related to explainability. Notably, more recent and advanced deep learning architectures such as U-Net, EfficientNet, or Visual Transformers have been explored for tasks like disease localization or automatic X-ray image annotation to enhance explainability.

Interestingly, a noticeable shift toward chest X-ray data was observed between 2020 and 2021, primarily attributable to the COVID-19 pandemic, which spurred interest in exploring DL architectures within the scientific community.

4. Taxonomy for eye gaze-drive approaches in deep learning for radiology

This study aims to provide a structured and comprehensive taxonomy that classifies the different DL eye-tracking studies in the literature. This classification comprises their different strategies, highlighting their

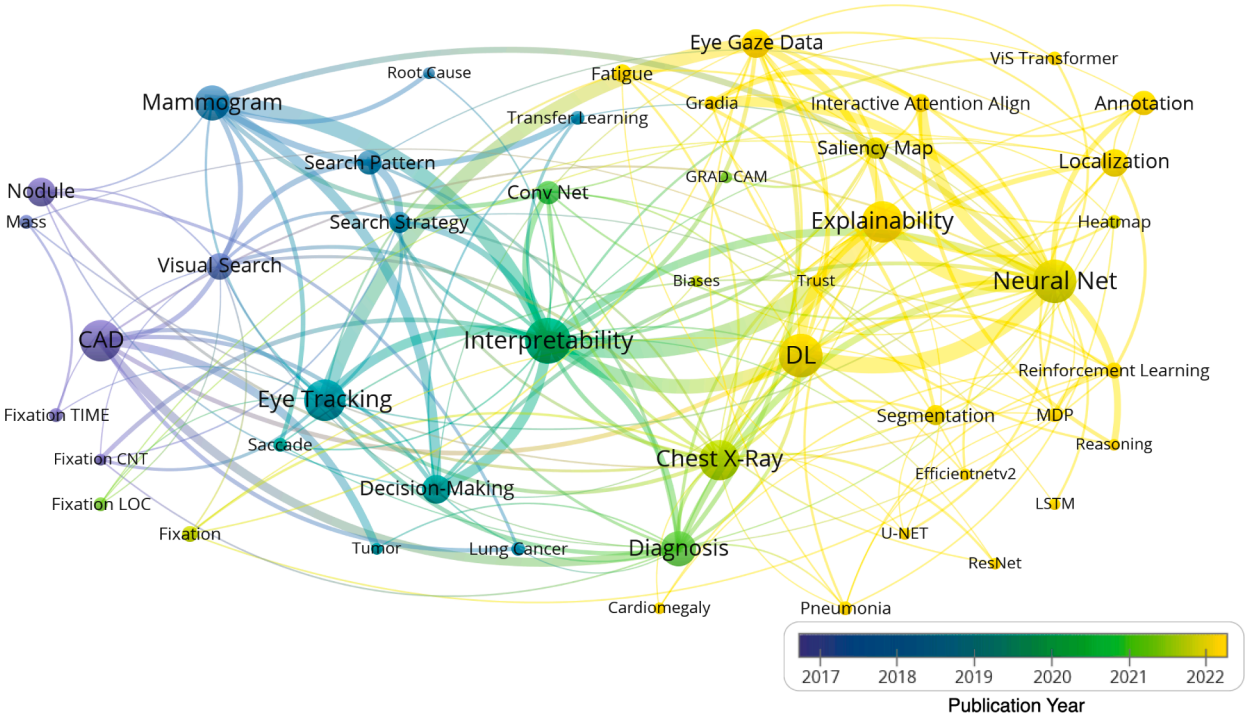

Fig. 3. Network visualization of co-occurrence between topics in the final list of retrieved articles.

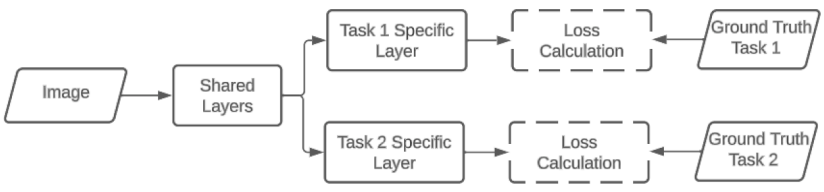

Fig. 4. Hard-Parameter Sharing framework flowchart representation.

unique aspects, commonalities, and divergences while underlining their respective contributions to the field of radiology. We provide detailed information on the taxonomy and approaches of similar works, discussing the datasets, pros and cons of each work, to better contextualize our analysis. The taxonomy starts with dividing the studies by task before assigning labels characterizing their approach. Table 2 presents the different tasks identified in the literature.

Table 2 presents the different tasks identified in the literature.

Our analysis indicates that researchers use eye-tracking data in DL methodologies for radiology on the following tasks: diagnostic classification, detection of objects or Regions of Interest (ROIs), analysis of the expertise levels, evaluation of errors and fatigue, and attention analysis.

4.1. Classification

In machine learning, classification is the process by which a model categorizes an input (e.g., an X-ray or MRI scan) into one of several predefined classes or labels. In AI healthcare applications, this corresponds to performing an automated diagnosis given a certain element about a patient. Deep learning models have shown a special potential for this task and thus aid medical professionals in being more efficient and precise. However, their deployment faces difficulties in interpretability because medical professionals do not have access to the models' internal decision process, ultimately leading to challenges in their adoption in real-world practice. Incorporating eye-tracking data into X-ray analysis promises to enhance diagnostic accuracy, streamline workflows, and foster explainability by capturing expert radiologists' gaze patterns, enabling multi-modal learning towards facilitating real-time, human-in-the-loop decision-making. This section reviews literature exploring various ways this integration can be done. We divided this category into two dimensions: multitasking and single-task classification. Single-task classification consists of predicting a particular diagnostic outcome from an imaging dataset, such as ascertaining the presence of a specific pathology or lesion. On the other hand, multitask classification undertakes multiple analytical operations concurrently on the same image, for instance, a model that predicts the presence of pathologies in an X-ray while simultaneously segmenting regions of interest. The two frameworks' goals are the same but take advantage of different learning approaches. While single-task classification focuses on targeted diagnostic precision, the multitasking approach capitalizes on the synergistic interactions between tasks, facilitating a holistic interpretation of the radiological data.

4.1.1. Architectures and techniques

The most common multitask architecture in our literature review is the encoder-decoder, a hard-parameter sharing framework. Hard parameter sharing involves using the same model weights for multiple tasks. In contrast, soft parameter sharing assigns separate weights for each task but incorporates their differences into the objective function (Crawshaw [98]). Following the structure of the first, Encoder-Decoder architectures have a U-Net-based structure with an added branch responsible for classification. At the same time, the decoder approximates radiologists' fixation maps(Agnihotri et al. [31], Karargyris et al. [32], Huang et al. [29], Watanabe et al. [17]), both branches sharing the weights of their feature extraction layers. They take advantage of the fact that the U-Nets' decoder default output is a mask, and thus, it is possible to learn how to approximate the radiologists' gaze distribution and share that knowledge with the classification branch.

Table 2
Division of the literature by the studied task/output

Task	Studies
Diagnosis Classification	[27–32,17,33–40]
Abnormality Detection	[41–47]
Expertise Level Analysis	[48–51]
Error Detection	[52–57]
Fatigue Analysis	[58–60]
Attention Analysis	[61–64,53,54,65–68]
Others	[69–74]

On the other hand, Bhattacharya et al. [33] described a teacher-student global-focal multitask approach to classification. This framework includes a teacher network that learns how to estimate expert fixation heatmaps and a student network that outputs both a classification and expert heatmap prediction. Inside them, the global component focuses on semantic context, while the focal one emphasizes fine-grained features. During training, the loss term associated with fixation heatmap prediction computes the difference between the ones the teacher and student networks predicted. This results in the need for fewer eye gaze annotated data during training since the teacher network can predict images without associated fixation heatmaps. In addition, its processing of gaze data enables the model to most naturally approximate the radiologists' by having a global-focal component analogous to the one observed in radiologists' X-ray analysis process(Sheridan and Reingold [99]).

The architecture class "Non-ED hard parameter sharing" refers to models that follow the hard parameter sharing framework without using an encoder-decoder architecture([28,27]). Jesse Kim [28] implemented their model with task-specific layers that take the form of two sets of linear layers, one responsible for classification and another for gaze distribution prediction. Saab et al. [27] uses the same network to predict gaze quantitative features and make class predictions. This architecture, while following the hard parameter-sharing framework, offers flexibility by not using an encoder-decoder structure. It is important in cases where we may need a more specific model design and gaze feature extraction process for the radiological task.

The approaches we include in the single-task definition are more diverse than the ones from multitasking. The few shot learning approaches in Ma et al. [35] and Ma et al. [36] tackle shortcut learning, a problem typically reported in ViTs' learning processes. Shortcut learning happens when the model exploits correlation and biases in training data that are not generalizable to unseen data. To tackle this problem, the authors apply an eye gaze-guided mask to screen out image patches that were not significant for the radiologist's decision. The researchers implement a residual connection between the unmasked input and the last ViT encoder layer to retain the relationship between patches. ViTs are emergent models in medical imaging classification. The integration of gaze distribution information in the form of this solution aids this technology in achieving its full potential in the field of automated radiology through a less biased and more generalizable training process.

Rong et al. [37] focuses on performing fine-grained classification. The authors implement a model of gaze augmentation training and knowledge fusion network segments. The gaze augmentation training aims to force the model to focus on regions of human gaze selected as relevant by implementing a sliding window algorithm to find areas that contain human attention and then ranking them according to the windows' averaged pixel values. After that, the model chooses k windows with the highest scores to include in the augmented version of the input image. In the knowledge fusion network segment, both the augmented version and the original image enter different feature extraction layers before an element-wise fusion happens, and the resultant output is moved into a single CNN network.

The experiments with temporal heatmaps of the paper Agnihotri et al. [31] and Karargyris et al. [32] are equivalent except in their evaluation methods, which is the reason why they obtain different results. Temporal heatmaps differ from standard fixation heatmaps by separating the progression of the radiologists' gaze by temporally organized frames instead of plotting the whole gaze distribution over the same image. The model that processed them consists of an encoder-decoder model with an extra input LSTM branch to process the set of temporal heatmaps relating to the input image. Lanfredi et al. [42] directly uses gaze data to diagnose spatial neglect.

These last three approaches have in common the concatenation of a branch processing the original X-ray and another branch processing the fixation heatmaps to achieve feature fusion between the two information sources. This framework allows the model to adapt how features are directly extracted from both sources before completing the image-related feature maps with gaze-related features.

Another group of architectures consists of the ones that, instead of having a decoder responsible for approximating radiologists' gaze distribution, directly add to the classification loss a term describing the proximity between that distribution and the model's saliency maps (Wang et al. [38], Zhu et al. [39], van Sonsbeek et al. [40]). We refer to them as SM to HA (Saliency Maps to Human Attention). These architectures implement the most direct connection between the models' saliency maps and the radiologists' gaze distribution, directly forcing the feature extraction layers' to be influenced by the fixation heatmaps, enabling a more visible transformation in the models' saliency maps, and thus of their explainability.

As for loss functions, cross-entropy is the most common choice for classification.

However, lesion detection requires additional information extraction from gaze data, particularly regarding lesion location.

4.1.2. Datasets and data processing

Except for Saab et al. [27], multitask framework approaches rely on fixation heatmaps as a proxy for radiologists' attention, approximating the models' saliency maps or output mask of the decoder. The fixation heatmaps distinguish themselves for their inherent interpretability for being a spatial distribution that visually salients the regions radiologists' found to be more important. By shifting attention toward the distribution they illustrate, models will tend to become more interpretable. In addition, fixation heatmaps reduce data dimensionality, providing a more direct flow of information regarding the location of important features for classification. On the contrary, in Saab et al. [27] the model predicts gaze quantitative features relevant for classification. This paper is the only one to focus beyond chest X-rays (CXRs) as image inputs in Table 3, focusing also on Magnetic Resonance Imaging (MRIs). The most commonly used dataset is EYEGAZE, which includes chest X-rays and fixation maps for the classes "Normal," "CHF"(congestive heart failure), and "Pneumonia." The remaining datasets are more or less of the same importance, with VinBigData distinguishing itself for its 14 classes.

Fixation heatmaps are also the most common technique to process gaze information in Table 4, with every work surveyed, except Franceschiello et al. [34], using it for expertise attention proxy. Agnihotri et al. [31], Zhu et al. [39], Karargyris et al. [32], Rong et al. [37] used the EYEGAZE dataset while Ma et al. [35,36] used SIIM-ACR, INbreast, FIGRIM and CAT2000, that are two/three labeled datasets. van Sonsbeek et al. [40] used EYEGAZE, REFLACX and Chest X-ray14. Zhu et al. [39], van Sonsbeek et al. [40], Karargyris et al. [32], Agnihotri et al. [31], Rong et al. [37], Ma et al. [36,35] focused on chest X-rays (with Ma et al. [36,35] also using mammography imaging) and Wang et al. [38] on knee X-rays. (see Table 5 and 6).

4.1.3. Results and Interpretation

Despite having found gaze to impact classification negatively, both Karargyris et al. [32] and Agnihotri et al. [31] found gaze to shift the output mask of the decoder towards the expert's gaze distribution. Agnihotri et al. [31] and Watanabe et al. [17] re-reproduced the second approach Karargyris et al. [32] described. The negative results that the cross-validation evaluation technique outputted lead the authors of Agnihotri et al. [31] to state that the collection of gaze data is not worth the effort. However, in a successful attempt to fix the architectures from Agnihotri et al. [31], Karargyris et al. [32], Watanabe et al. [17] introduced a loss term between the saliency map of the encoder and the fixation heatmap into the model's framework, in addition to the segmentation loss and classification loss sum. As a result, the inclusion of gaze increased classification performance, and correct predictions led to higher proximity between the network's saliency maps and radiologists' gaze distribution. This suggests that the type of learning loss is fundamental in the successful fusion of eye gaze data with radiology images, and future studies should reconsider refined loss functions in their models.

One possible explanation for the results this trio of studies obtained is that the U-Net's decoder feature layers can change to better approximate fixation heatmaps without having to affect the encoder feature layer's weights, rendering the extra task effect of little significance. In that case, the models need a direct loss term, forcing the encoder-selected features to approximate the ones salient in the heatmaps.

The correlation between heatmap proximity and classification performance also appears in Jesse Kim [28], although not reporting increased performance with the introduction of gaze data. With the addition of Bhattacharya et al. [33], these were the only studies of our survey to perform classification with more than three labels. The latter also experimented with several datasets with a binary class setting, with two reporting 14 different classes, and the EYEGAZE dataset with three classes, observing improved classification performance in every case. The authors of this study associated datasets with a higher number of labels with increased model dependence on focal features. In comparison, a lower number of labels led to having the global component play a

Table 3
Literature on performing classification with multitask learning. Consult A.

Secondary task	Architecture	Dataset	Papers	Fixation/Image Processing	Image Input	Model	Loss	Metrics
Normalized Gaze Features Prediction	Non-ED Hard Parameter Sharing	EYEGAZE [32], SIIM-ACR [75], Private (MRI), (*)	[27]	Gaze Quantitative Features	CXR, MRI	ResNet-50 [76]	Soft CE	AUC, F1, Precision
Expert Attention Prediction		JPG [77], EYEGAZE [32]	[28]	Fixation Heatmap	CXR	DenseNet-121 [78]	BCE, MSE	AUC, SSIM
	Encoder-Decoder	EYEGAZE [32]	[29]			U-Net [79], EfficientNet [80]	BCE, Dice	AUC, Dice
		SIIM-ACR [75], RSNA [2], CheXpert,[81] *	[30]			ResNet-34 [76]	BCE, Custom IoU	AUC, ACC
		EYEGAZE [32]	[31] [32] [17]			U-Net [79]	BCE, MSE	AUC, AO AUC
						U-Net [79], Grad-CAM [82], DeconvNet [83], Guided Backpropagation [84]		
	Teacher-Student Global-Focal	EYEGAZE [32], Cell [85], NIH [86], VBD [87], RSNA [2], SIIM-ACR [75]	[33]	Fixation Heatmap, Image Augmentation		Swin Transformer [88]	CE, GIoU, MSE	AUC, F1, ACC, Precision, Recall

Table 4
Literature on performing classification with single task learning. Consult A.

Architecture	Dataset	Papers	Gaze/Image Processing	Input	Model	Loss	Metrics
Few-Shot Learning	SIIM-ACR [75], INbreast [89], (Gaze data from [27])	[35]	Fixation Heatmap Attention Zones Cropping	CXR, Mammography	ViT [90]	CE	ACC, F1, AUC
	SIIM-ACR [75],INbreast [89], FIGRIM [91], CAT2000 [92]	[36]			EML-NET [93], ViT [90]	CE, KLD, PCC, NSS	
Multimodal DL	EYEGAZE [32]	[37]		CXR	ResNet-50 [76]	CE	ACC
		[32]	Fixation Heatmaps		U-Net [79]	BCE, MSE	AUC
		[31]					
Approximating AM to HA	OAI [94]	[38]	Gaze Probabilistic Distribution Fixation Heatmap	Knee XR	ResNet [76], CAM [95]	BCE, Custom MSE	ACC, MAE
	EYEGAZE [45]	[39]	Fixation Heatmap	CXR	ResNet [76], EfficientNet [80], CAM [95]	CE, Selective MSE	Precision, Recall
	EYEGAZE [32], REFLACX [67], Chest X-ray14 [96]	[40]			VAE [97]	Custom VAE Loss	AUC, F1, Dice, MSE

Table 5
Literature on performing object detection. Consult A.

Secondary Task	Architecture	Dataset	Papers	Gaze/Image Processing	Input	Model	Loss	Metrics
Expert Attention Prediction	Teacher-Student Global-Focal	EYEGAZE [32], REFLACX [67], RSNA [2], SIIM-FISABIO-RSNA [102], NIH [86], VBD [87]	[41]	Fixation Heatmap Radiomics	CXR	Swin Transformers [88]	CE, GIoU, MSE, VAL [41]	AUC, MSE
	Encoder-Decoder	REFLACX [67], JPG	[42]	Extracting Lesion Location from gaze		ResNet [76]	MIL loss	AUC, IoU
FP removal		Private	[43]	Graph representation of gaze, Clustering of nodes, Gaze graph sparsification	Lung CT scans	CNN [103]	CE	ACC, Dice
	General DL	LIDC-IDRI [104]	[68]	3D Attention Estimation		YOLOv3 [105]	MSE, MIL loss	Recall
NA	Probabilistic Graphical Modal	Private	[44]	Gaze Probabilistic Distribution	CXR	PIM [44]	NA	Recall, Spec
	Multimodal DL	REFLACX [67]	[45]	Fixation Heatmap		Custom Mask-RCNN [106]	BCE	Average Precision, Average Recall
	Labeling Algorithm	Images from PubMed and Google images	[46]	Fixation points convex hull	MRI	U-Net [79]	Dice	AUC
		BraTS	[47]	Matching of gaze points with spoken keywords		CNN [103]	NA	ACC

Table 6
Literature on attention analysis. Consult A.

Dataset	Papers	Fixation/Image Processing	Input	Model	Loss	Metrics
Private *	[61]	2D Lesion Segmentation	Panchreas CT, Demographic Tabular Data	RF[110], CNN [103] (Bayesian Combination)	NA	ACC, EMD
REFLACX [67]	[62]	Fixation Heatmap Image Augmentation Center Bias Estimation	CXR	CNN [103], Grad-CAM [82]	BCE	AUC, NCC, PCC
EYEGAZE [32]	[63]	NA		U-Net [79], CNN [103], Grad-CAM [82]	CE	ACC, KLD, NSS
Private *	[64]	Attention Based Clustering Background Supression Cluster Features' Extraction	Mammograms		NA	
	[53]	Attention Based Clustering Histogram Normalization Color Conversion		Inception ResNet V2 [111]	CE	Precision, Recall, ACC, P-value
	[54]			ResNet [76], Inception ResNet V2 [111]		Precision, Recall, ACC, F1, Recall, Spec, NPV, PPV,
SALICON [112], Mammogram eye-tracking dataset [113]	[65]	NA		HRNet [114]	CC, KLD, SIM, NSS	CC, SIM, KLD, NSS
EYEGAZE [32]	[66]	Fixation Heatmaps	CXR	Markov Chains		NA
REFLACX [67]	[67]			NA		NCC

*Authors collected their eye gaze data.

bigger part.

These results may suggest that the task of using gaze to improve classification on a dataset of 14 classes is fairly different from the analogous with 3 classes because it requires the ability to extract fundamentally different features, therefore either requiring different model architectures or multi-resolution frameworks like the global-focal attention modules in Bhattacharya et al. [33], capable of leveraging fixation heatmap and X-ray features' levels' as needed.

Another example of information leveraging is present in Huang et al. [29], where the researchers train a model in a normal/abnormal classification setting, although the dataset they used includes five different labels. Their framework distinguishes itself for its multi-head auxiliary attention block (AAB) that selects important information from the fixation heatmap prediction branch. The authors of this paper compare their approach directly with a hard parameter-sharing implementation analogous to the ones Agnihotri et al. [31], Karargyris et al. [32], Kholiavchenko et al. [30], Watanabe et al. [17], Jesse Kim [28], and Saab et al. [27] described, upon which their approach improves. They conclude that not only is their gaze responsible for raising the class prediction score, but so are the components they added to the general hard parameter-sharing framework.

One thing in common between this approach and the one described in [33] is the separate training of their fixation heatmap estimation components before being used to train the classification one. Another is using attention mechanisms to shift the model's focus into gaze-related features. One or both of these characteristics could contribute to the comparative success of these frameworks. Frameworks' inclusion of attention mechanisms may be an advantage for including gaze into the learning process. Although CNN-based architectures can incorporate gaze to improve classification, they lack a direct way since their attention is implicit. Attention-based architectures have an explicit attention mechanism that can be directly influenced by the radiologist's attention, which may ease the extraction and adaptation of global and focal features.

Another transformer-based approach is the one the researchers in Ma et al. [35] and Ma et al. [36] implemented. Ma et al. [35] compares its approach to the ones from Karargyris et al. [32], Agnihotri et al. [31], and Wang et al. [38], achieving superior performance, and both Ma et al. [35] and Ma et al. [36] observe the increase in performance and reduced shortcut learning through the usage of gaze attention. Although obtaining positive results, the authors recommend caution with gaze incorporation into the learning process, for it is not always reliable, and there should be a balance between features the model learns by itself and those it obtains through the fixation heatmaps.

In fact, expert gaze can potentially introduce harmful biases ([100,101]), as models can end up adopting misdirected attention behavior. However, the attention zone cropping technique from Ma et al. [35] and Ma et al. [36] uses fixation heatmaps to select relevant regions for classification to decrease shortcut learning. Their positive results support the hypothesis that fixation heatmaps select features relevant for classification.

The results from Wang et al. [38] and van Sonsbeek et al. [40] further support this hypothesis. Both these papers compared adding bounding boxes and fixations heatmaps to improve classification, with Wang et al. [38] observing similar results with their addition and van Sonsbeek et al. [40] favoring fixation heatmaps. The authors of Wang et al. [38] also obtained better results from their attention consistency approach than gaze-free implementations of ViTs and ResNets. The model achieved the worst results for the REFLACX. However, it has three times the data of EYEGAZE, leading the authors to question the data quality of the first and the effects of dataset inconsistency. Zhu et al. [39] also obtains superior results in relationship to gaze-free baselines, although the proximity of its model's CAMs to fixation heatmaps only increased for the ResNet backbone.

The results from Wang et al. [38] and van Sonsbeek et al. [40] suggest that fixation heatmaps may be at least as informative as location labels like bounding boxes for models performing classification.

The multimodal approach from Agnihotri et al. [31] and Karargyris et al. [32] obtained different results in both studies. Although unchanged, the authors used the same dataset. The difference resides in using cross-validation to evaluate the model in Agnihotri et al. [31], invalidating the results from Karargyris et al. [32].

4.1.4. Eye gaze's role in classification

The integration of experts' gaze data into an automated deep-learning classification of medical imaging has the potential to augment the diagnostic process significantly. By mirroring the radiologists' attention, it systematically improves models in terms of interpretability and performance. One of the main roles we predicted gaze data to have was to turn the attention of the models into a readable form without necessarily impacting performance. However, from the literature analyzed, we have observed that performance and explainability are inextricably linked. Informing the models of radiologists' fixation data helps them arrive faster and with higher precision to successful feature extraction, for even when models arrive at the relevant data representations by themselves, these can be noisy, with useless information that causes their saliency maps to be very difficult to interpret. This is why eye-tracking data helps models surpass the "black-box problem": it helps them arrive at cleaner representations of the features they need to perform successful classification. However, this data also contains information on the radiologists' biases and misconceptions, leading models to adopt misdirected attention behavior potentially. The processing of gaze data and architecture design must contain leverage and filtering mechanisms for models to arrive at an equilibrium between features they learn by themselves and those they obtain through the fixation heatmaps.

4.2. Object detection

In the intricate field of radiology, mere class prediction of an image often falls short of clinical needs. Pinpointing the exact location of a lesion within the image requires a deeper comprehension of radiological data. Given the often subtle nature of these lesions and the profound implications of oversight, there's a pressing demand for precision. Deep learning models emerge as promising tools to assist radiologists in this nuanced task of lesion detection. Yet, to fully harness their potential, it is crucial to develop mechanisms that elucidate the rationale behind their decisions. Incorporating gaze data into the training of lesion detection models can potentially illuminate the model's focus areas, offering radiologists a transparent and verifiable insight. In this section, we delve into literature that probes the integration of eye-tracking data into medical imaging models for lesion detection.

4.2.1. Architectures and techniques

The architectures we have seen applied in classification can potentially still be useful in lesion detection since, in addition to providing features useful to perform global diagnosis, eye gaze contains information about possible regions containing lesions.

Like the model Bhattacharya et al. [33] used to perform classification, Bhattacharya et al. [41] implemented a teacher-student global-focal architecture for object detection. The researchers of this study observed that, while radiologists tend to focus on fine-grained features, it is possible that global textural features also contain relevant information not present in those features. To include information about these features, the authors implemented a model that computes a joint data representation of visual and radiomics attention and a loss function that includes the distance between this probabilistic distribution and the student's attention. Lanfredi et al. [42] and Khosravan et al. [43] also implemented multitask solutions, both encoder-decoder based, although different in their secondary task (fixation heatmap prediction and false positive removal, respectively). In addition, Khosravan et al. [43] implements a sparsification technique to reduce the dimension of the input

data representation to a fraction of its initial size without changing its topological properties. This technique consists of obtaining clusters of fixations and using them as edges to a graph representation of the distribution of the gaze. By computing this representation, applying a spectral graph sparsification method that preserves the Laplacian, and thus the structural similarity, of the original graph(Spielman and Srivastava [107]) is possible.

This technique is especially relevant for its ability to salient specific locations with a higher presence of fixations while also containing information about the global distribution of the gaze. This ability is especially useful in lesion detection since representations like fixation heatmaps have poorer location-specific representation.

On the other hand, Lanfredi et al. [42] and Stember et al. [47] used gaze data for lesion labeling, by observing the fixations associated with the dictation of certain words by radiologists. However, while the approach from Lanfredi et al. [42] used the fixations from the phrase before and in which the keyword was mentioned (and the ones between), the Stember et al. [47] used the position of the gaze at the beginning of the keyword. The position at the end is to obtain a middle point which is their lesion label. The model of Lanfredi et al. [42] uses multiple instance learning(a weakly supervised learning technique where labels are given to sets instead of individual elements) and outputs a grid of cells within which it performs classification. Stember et al. [46] also attempted to validate the use of gaze data as an object detection label by having the radiologist move his foveal area around meningiomas in MRIs instead of hand-annotated labels.

Automatic labeling is crucial in developing automated radiology approaches since data scarcity is one of the main challenges this field faces. In the approaches we just described, eye gaze data enables the annotation of a lesion's location solely through the focus of the radiologist and thus contributes to a more efficient collection of training datasets. Instead of using expert gaze attention, Wedel et al. [44] modeled the eye movement information from laypeople through a Markov model with partially invisible states. The authors associated fixations with pre-existing ROIs (regions of interest) and assume the transitions between them follow a Markov process. Following a Markov Chain Monte Carlo algorithm, the model gives birth and kills ROIs. Also, in a single-task setting, the goal of Luís et al. [45] was to evaluate the impact of introducing eye gaze data in the learning training process of a Mask R-CNN(He et al. [106]) model by adding an extra input processing arm to its default architecture. The model fuses the feature maps from the layers with a chest X-ray image as input and those with a fixation heatmap, with the result entering the region proposal network segment of the Mask R-CNN(He et al. [106]).

4.2.2. Datasets and data processing

Except for Aresta et al. [68], the works performing object detection focused on chest X-rays, using the same datasets studies utilized for classification in the previous section. These include EYEGAZE, VinBigData, REFLACX, SIIM-FISABIO-RSNA, and NIH, and the most common gaze processing technique also consisting of obtaining fixation heatmaps.

In addition to classification-related features, fixation heatmaps contain information about regions of interest that can be useful for object detection. However, the authors of Wedel et al. [44] used a private dataset that did not include fixation heatmaps. They directly used fixation data as input for their Markov Chain Monte Carlo algorithm. They modeled the fixation density with a Poisson distribution to estimate the presence of nodules in the image. [42] estimated fixations heatmaps and used them as lesion detection labels. To accomplish that, the authors obtained the fixations from 1.5s before the phrase on which the radiologist mentions the label until the last time the radiologist dictates it. They used a mixture of data with labels with and without localization labels to train their model. Similarly, Stember et al. [47] extracted lesion location labels in MRIs by averaging the gaze points of a radiologist from the beginning and ending of lesion referencing keywords. Stember et al. [46] also sought to obtain lesion location labels by having a radiologist move their foveal area around meningioma lesions in MRIs. After that, the authors performed a convex hull to delimit the lesion region. Khosravan et al. [43] processed their fixations by building a graph representation of fixations and applied the BIRCH algorithm(Zhang et al. [108]) to conduct hierarchical clustering followed by a sparsification algorithm to reduce the complexity of the graph without losing the structural similarity between the sparsified and original graphs.

4.2.3. Results and Interpretation

Both Lanfredi et al. [42] and Luís et al. [45] did not observe improvements in object detection with the introduction of gaze data into the learning process. In the first, the model could localize lesions better than a baseline model that did not have access to localization labels and worse than one with hand-annotated ellipses. In addition, the authors observed that neither the segmentation loss from the extra model branch nor having a loss version that included localization error contributed to the model's classification performance. However, the output saliency maps increased similarly to the radiologists' fixation heatmaps. Similarly, Luís et al. [45] did not obtain any improvements related to introducing fixation maps in the learning process. The authors hypothesized that it may be because of the dataset's data collection protocol.

Although Lanfredi et al. [42] and Luís et al. [45] introduced gaze into their learning processes with very different techniques, since they both use the dataset Lanfredi et al. [67] created, the high degree of noise of the data may have contributed to the negative results observed in both cases. The REFLACX dataset has the problem of containing fixations related to the interaction of the participants with the interface, which most certainly would increase the difficulty of having a successful training process while depending solely on it, as both these studies did. The approximation of the output saliency maps to radiologist's attention is similar to what we observed in Agnihotri et al. [31] and Karargyris et al. [32], possibly also for the lack of an adaption like the authors of Watanabe et al. [17] describe. In addition, it is also possible that, although adequate for classification, their framework is not applicable to object detection.

On the other hand, Bhattacharya et al. [41] associated positive results with using both gaze data and radiomics features. Stember et al. [46] observed that their technique and manual labels achieved a high average Dice score, and U-net-based architecture models trained on them achieved similar results. Likewise, Stember et al. [47] used gaze-based location labels to train a deep learning model in object detection and achieved positive results.

The texture radiomics features the model extracted are analogous to the ones Mall et al. [55] and Tourassi et al. [56] showed to be the most relevant for error prediction. The results Bhattacharya et al. [41] obtained showed that the combination of radiomics and fixation features is associated with an increase in performance, showing the complementary nature of both sources of information in a teacher-student setting. The results from Stember et al. [46] and Stember et al. [47] further validate gaze masks as a lesion annotation method. An additional perspective on the role of radiomics features is their representation of the benefits derived from incorporating supplementary, pertinent patient information (such as omics data and clinical history) into the analysis. This approach provides the diagnostic model with a broader context, a strategy consistently observed in the literature to correlate with enhancements in model performance. The inclusion of such multifaceted patient data often leads to more informed and accurate predictions, as supported by various studies [109].

The positive results Khosravan et al. [43] obtained hints at the efficacy of their sparsification technique as a valid gaze processing algorithm and a potential candidate for the missing piece between classification and object detection fixation gaze processing techniques.

4.2.4. Eye gaze's role in object detection

Lesion detection demands a deeper understanding of the image to

pinpoint the precise location of abnormalities than classification, creating a new dimension in which gaze data can be useful.

In this task, we can process gaze similarly to classification and thus provide information about feature selection and relevant regions to the models, but its application has many ramifications here. One of the main challenges in automated diagnosis is the lack of available training data; gaze data can play an important role in more efficient automated labeling solutions.

Another application lies in detecting regions of interest with the potential of containing lesions, upon which radiologists and models can perform a more thorough search. This case is particularly useful for CAD systems operating simultaneously with radiologists, helping them conduct a more effective data analysis.

4.3. Attention Analysis

Enabling models to understand how radiologists' attend to medical images is the cornerstone of improving automated diagnosis. Their fixation patterns contain key information about the relationship they establish between image features and the presence of anomalies. Deep learning models are a very promising emerging technology in healthcare. However, for them to be useful, their attention has to be understandable to radiologists.

Eye tracking data emerges as a potential bridge between the radiologists' and models' attention. This section aims to analyze the information expert gaze contains and compare the attention it represents to naturally occurring attention in trained automated diagnosis models.

4.3.1. Architectures and techniques

We find in the literature on attention analysis to encompass two main themes: the comparison between the attention gaze-free machine learning models acquired during medical image classification training and the attention exhibited on expert's fixation heatmaps(Dmitriev et al. [61], Lanfredi et al. [62], Watson et al. [63]), and the analysis of the relationship between image features and expert's attention(Mall et al. [64], Mall et al. [53], Mall et al. [54], Lou et al. [65]).

The studies focusing on the first theme created models for classification without eye movement information. Dmitriev et al. [61] implemented a random forest segment that processed tabular data, a CNN network to process the medical image, and a Bayesian combination to combine both segments. Lanfredi et al. [62] appended a Grad-CAM (Selvaraju et al. [82]) to a CNN with and without attention gates. Watson et al. [63] implemented an ensemble model with a discriminator that attempts to distinguish the sub-models based on their saliency maps.

They used these models to analyze the relationship between attention and interpretability. Dmitriev et al. [61] had a radiologist evaluating the utility of seeing the importance of the random forest attributed to each tabular feature and the Two-dimensional t-distributed Stochastic Neighbor Embedding (t-SNE) visualization of the last hidden layer representations from the CNN component. Lanfredi et al. [62], Watson et al. [63] compared their models' output saliency maps to the radiologists' fixation maps. In addition, to make a valid comparison, the authors of Lanfredi et al. [62] refer to the need to deal with the center bias characteristic of fixation heatmaps in chest X-rays. To address this problem, they compute shuffled versions of their metrics, shuffled AUC (sAUC), and shuffled NCC(sNCC), to penalize saliency in the center of the image and reward focus on regions other than the center.

The works from the second theme(Mall et al. [64,53,54]) have their gaze processing technique in common. They start by identifying and categorizing regions of the image into fixation clusters (FCs), marked peripherally fixated clusters (MPFCs), and never-fixated clusters (NFCs). FC clusters have at least three sequential fixations within 2.5° of the visual angle from each other. MPFCs are areas where a radiologist marked a lesion but has less than three sequential fixations within 2.5°. We call this strategy attention-based clustering. In opposition to fixation heatmaps, this technique provides a discriminated spatial distribution of the radiologists' attention, prioritizing location-specific information instead of the overall morphology of the gaze distribution. Subsequently, the researchers of Mall et al. [53] and Mall et al. [54] implemented deep learning models to predict what kind of attention cluster will a given image patch gather, while the authors of Mall et al. [64] extracted local features (spectral features from Gábor filters and gaze features like dwell time) and global features(spatial-dependency features) to establish differences in these features for the different cluster types through One-way and N-way ANOVA analysis.

These works provide two views on feature extraction: the automatic selection by models to gather an unbiased collection of features that best fits the given task and the manual selection of features that deep learning models may have difficulty obtaining for being abstract and specific. On the other hand, Lou et al. [65] focuses on predicting the radiologists' gaze distribution on mammograms without losing detail. The point of their architecture is to be able to modulate in a parallel manner multiple levels of features by having two or more streams attending to different scales that share information.

The authors of Lanfredi et al. [67] and Moreira et al. [66] also analyzed expert gaze informative power. While the researchers of Lanfredi et al. [67] drew ellipses around each lesion and compared the number of fixations to fall within these regions from the average gaze distribution of the radiologists among all chest X-rays examined and the specific gaze distribution of the radiologists, the authors of Moreira et al. [66] investigated if it is possible to infer Markov chains describing the gaze patterns of the radiologists for both silent and reporting periods and for each specific lesion.

4.3.2. Datasets and data processing

Relating to datasets, studies focused on chest X-rays used either REFLACX(Lanfredi et al. [62], Lanfredi et al. [67]) or EYEGAZE(Watson et al. [63], Moreira et al. [66]), while studies focused on mammograms mainly collected their gaze data and took their images from private sources(Mall et al. [64], Mall et al. [53], Mall et al. [54]) with one exception that used SALICON and ImageNet datasets for model pretraining before fine-tuning with Mammogram eye-tracking dataset ([65]). As for processing techniques, the works Lanfredi et al. [62], Lanfredi et al. [67], Moreira et al. [66], and Lou et al. [65] used fixation heatmaps, while Mall et al. [64], Mall et al. [53], and Mall et al. [54] opted for attention based clustering to organize the data from the gaze. These last three works also used image augmentation techniques, including image rotation and distortion, as well as transformations on cluster image patches like histogram normalization and color conversion to have an adequate input for a CNN(Mall et al. [53], Mall et al. [54]) and background suppression to exclude useless zones from the image. Lanfredi et al. [62] also applies center cropping and horizontal flipping. Mall et al. [53], Mall et al. [54], and Lou et al. [65] also describe and evaluate the use of transfer learning. Both data augmentation and transfer learning represent important strategies for dealing with the problem of training data scarcity in the field of automated radiology.

4.3.3. Results and interpretation

The results from the first theme studies Lanfredi et al. [62], Watson et al. [63] support the hypothesis that, with training, gaze-free classification models will have their attention become closer to radiologists' attention. Jesse Kim [28] and Watanabe et al. [17] had already addressed this hypothesis and associated positive results with higher proximity between saliency maps and expert fixation heatmaps. The results Watson et al. [63] obtained strongly supported this hypothesis by showing that, by forcing consistency between models, their saliency maps become closer to the radiologist's attention than in the experiences Karargyris et al. [32] described, which directly used fixation heatmaps to shift its model's attention. In Lanfredi et al. [62], when using metrics that penalized proximity and rewarded distance from the center of the image, the authors were able to surpass the inter-observer difference

baseline while using a model consisting of CNN layers and a Grad-CAM module with attention gates, showing the model's ability to excel in highlighting specific areas where radiologists fixate more than average. In addition, Lanfredi et al. [67] observed that fixation heatmaps have a stronger relationship to the lesion included in the image than the average gaze distribution, supporting the hypothesis that gaze patterns have a specific component to the analyzed image. On the other hand, Moreira et al. [66] observed that, although gaze patterns recorded during search periods tended to be universal, eye movements from the reporting phase exhibited an arbitrary nature. These results support the hypothesis of an inherent difference between the gaze from the two phases. However, the eye movements while reporting were expected to focus on the found lesions, thus exhibiting more circular, concise, and predictable patterns. From the second theme, Mall et al. [64] observed that regions that attracted different kinds of attention (foveal, peripheral, or none) exhibited different global and local properties. In contrast, Mall et al. [53,54], Lou et al. [65] could predict the attention each region would gather by automatically selecting features with deep learning models. The authors of Lou et al. [65] concluded that both the deep encoder (that focuses on high-degree local features) and shallow encoder (that focuses on more global low-level features) add their specific contributions to the quality of the prediction.

GIoU considers the aspect ratio and size differences between bounding boxes, making it more robust in various scenarios.

4.3.4. Eye gaze's role in attention analysis

The literature we reviewed suggests that well-performing and consistent model saliency maps naturally approximate experts' attention. Although inter-observability and noise exist, the features radiologists use to perform classification are consistent, so models can predict the type of attention each image region will gather. This enables models to use gaze data to fast-track their training and converge into representations of data well-performing models would converge at. Radiologists' attention informs models of their internal decision processes for them to incorporate and mimic them. It has a global-focal framework of analysis that the literature also shows to be advantageous in deep learning models.

4.4. Error detection

Radiologists bear a significant responsibility and are held to high standards. However, the sheer volume of X-rays they examine daily can increase the likelihood of errors. Therefore, a support system that alerts radiologists to potential errors is essential. Eye gaze monitoring, due to its non-intrusive nature, offers a promising avenue for integration into radiologists' workflows. This section reviews literature exploring the relationship between radiologists' eye gaze and their errors. extends the concepts of ROC analysis to scenarios where radiologists provide multiple observations (marks) for each patient. It evaluates radiologists' ability to localize lesions by assessing their performance in correctly marking the lesions and estimating their accuracy.

4.4.1. Architectures and techniques

[52] identified statistical links between gaze characteristics and mistakes made by radiologists. They found that merely looking at a lesion does not guarantee its correct marking and that marked lesions are glanced at more frequently than unmarked ones. This finding underscores the potential of gaze-based techniques for detecting false negative errors. A group of studies [53–57] focused on predicting errors and types of errors using deep learning models and eye gaze data. These studies highlighted the importance of selecting appropriate features for these models. In contrast, some studies [53,54] allowed deep learning models to identify relevant features autonomously; others [55–57] manually extracted features using techniques such as the Grey Level Co-occurrence Matrix and Gabor filters.

In their respective studies, Mall et al. [53] and Mall et al. [54] utilized deep learning models to predict the type of error associated with each cluster. Mall et al. [53] developed a model solely for predicting errors and error types, while Mall et al. [54] employed a hierarchical prediction framework. This framework utilized information from attention prediction in error type estimation and a secondary model to predict the type of false negatives in missed cancer regions, specifically search, perception, and decision errors. Both studies highlighted the advantage of using deep learning models due to their unbiased feature selection. Conversely, Mall et al. [55] focused on analyzing the relationship between image and gaze features with the missed cancer error (false negative). Tourassi et al. [56] aimed to test three hypotheses: whether local lesion region content can predict dwell times, whether local lesion region content and dwell time can predict a decision, and whether local lesion region content, dwell time, and the reader's decision can predict errors. Pietrzyk et al. [57] implemented an SVM-based algorithm designed to provide feedback to radiologists by signaling potential false positives and false negatives. The study also evaluated the impact of this feedback on the radiologists' performance. Commonalities among these three papers include using various machine learning models and manually extracting spectral, texture, and gaze features. Techniques such as the Gabor filter, Grey Level Co-occurrence Matrix, and wavelet packet decomposition were employed for feature extraction. Notably, both Pietrzyk et al. [57] and Mall et al. [55] utilized ANOVA analysis to identify statistically relevant features for use after extraction. Researchers in these works have in common the usage of eye-tracking as a container of features related to errors or as a selector of regions containing important information about them, signaling the potential this data can have in error detection.

4.4.2. Datasets and data processing

Castner et al. [52] concentrated on OPTs and data sourced directly from the eye tracker's processing. In contrast, Mall et al. [53] and Mall et al. [54] extracted image patches from the regions of clusters identified through attention-based clustering. They then performed histogram normalization and color conversion on these patches. Additionally, they utilized image augmentation techniques, such as image rotation and distortion, and employed transfer learning. On the other hand, Mall et al. [55], Tourassi et al. [56], and Pietrzyk et al. [57] extracted texture and spectral analysis features using the Grey Level Co-occurrence Matrix, Gábor filters, and wavelet packet decomposition. Both Mall et al. [55] and Pietrzyk et al. [57] used ANOVA analysis to filter out redundancy from the previously selected features, focusing on those that were statistically different among classes and contained unique information. Furthermore, Mall et al. [55] conducted a Principal Component Analysis (PCA) analysis to determine the extent of original information retained by the features selected through the ANOVA analysis. While the automatic selection of features can be the most adequate approach for a certain error detection-related task, we may require manual selection when the task requires a more specific and abstract set of features. We consider both crucial to radiologist error detection. Eye gaze is the selector of important regions for error prediction. However, given that it reflects the decision processes of the radiologists, it contains important error-related features by itself.

4.4.3. Results and interpretation

In their research, Castner et al. [52] investigated the relationship between eye movements and errors committed by radiologists. Their findings underscored that observing a lesion does not ensure accurate identification. When a lesion area is observed, it is correctly marked merely by chance, thereby emphasizing the importance of addressing false negative errors. Additionally, they observed that marked lesions received more glances on average than those that were not marked, suggesting that techniques leveraging gaze data could detect false negative errors. A methodological divergence emerges within the analysis group of Table 7, revolving around the decision to either manually select features for conventional machine learning models or to allow a

Table 7
Literature on error detection. Consult A.

Papers	Gaze/Image Processing	Feature Extraction Method	Input	Model	Analysis	Metrics
[52]	NA	NA	OPT	NA	χ^2 test to study relationship between gaze recall and marking recall, Comparison of glance frequency in marked and unmarked targets	NA
[53]	Attention Based Clustering Histogram Normalization Color Conversion		Mammograms	Inception ResNet V2 [111]	NA	ACC, Kappa-statistics, Recall, Spec, Precision, NPV, Micro and Macro Precision, F1
[54]				ResNet [76], Inception ResNet V2 [111], Inception V4 [111], NASNet [115], VGGNet-19 [116]		
[55]	Cluster Feature Extraction	Manual extraction of features related to texture and spectral analysis	Mammograms, Cluster Features	SVM [117], Gradient Boosting [118], SGD, ResNet-152 [76], Inception ResNet V2 [111]	N-way ANOVA to find features different in search, decision and perceptual errors, PCA	
[56]			Mammograms, Cluster features, Diagnosis Confidence	RF [110], MLP, Adaboost [119], Logistic regression, Bagging [120], Naive Bayes, BayesNet[121], DMNBtext [122]	Leave-one-reader- -out and Leave- -one-case-out cross-validation sampling analysis	AUC
[57]			CXR, Mouse Clicks,	SVM [117]	JAFROC and ROC analysis	Recall, Localized Recall, Spec, AUC

deep learning model to identify its features autonomously. An interpretation of the current literature suggests that the optimal approach may be context-dependent. Mall et al. [53] and Mall et al. [54] achieved notable results using deep learning models. However, the deep learning model employed by Mall et al. [54] could not distinguish between search, perception, and decision errors. This finding led the authors to hypothesize that Convolutional Neural Networks (ConvNets) may lack the capacity to modulate this level of error granularity and are missing a temporal dimension. Mall et al. [53], Mall et al. [54], and Mall et al. [55] all utilized the same type of data clustering, making them apt for comparison. Mall et al. [55] aimed to achieve what Mall et al. [54] could not: distinguishing between types of false negatives. Mall et al. [55] employed traditional machine learning models and manually selected image and gaze features. They also implemented convolution-based networks as baselines, which they successfully outperformed. They hypothesized that this success was due to the convolution networks' inability to identify higher-order features crucial for differentiating false negative errors. Notably, while the convolution baselines only used image patches from the clusters, the machine learning models incorporated gaze features into their input, along with contrast and texture features from the image patches.

Through an N-Way ANOVA feature analysis, the authors observed that gaze features, such as the average distance between fixations in clusters and dwell time (which was not used due to its error-defining nature), were statistically significant in predicting search, perception, and decision errors. The authors employed the same feature processing pipeline as Mall et al. [64]. Mall et al. [64] demonstrated that clusters with different attention types (obtained using the same process as Mall et al. [53], Mall et al. [54], and Mall et al. [55]) could be distinguished through energy profiles and dwell time. Mall et al. [54] noted that a hierarchical approach combining cluster attention and error prediction tasks leads to improved model performance. Mall et al. [55] showed that dwell time and energy-based features were the most statistically significant predictors of error types. One possible interpretation of these results is the existence of an inherent connection between the features used in attention and error prediction. Tourassi et al. [56] tested the hypothesis that error predictive power exists in the data from gaze, local image, and the radiologist's decision. They extracted features similar to those used by Mall et al. [55] and Mall et al. [64] and concluded that error prediction has both a global and personal component, leading the best feature-model pair to depend on the radiologist, with a global emphasis on energy profiles. Pietrzyk et al. [57] developed a feedback algorithm that adjusted the model to each radiologist to make false positive and false negative predictions. This technique was achieved by extracting spatial frequency features in regions with prolonged dwells. However, this did not result in significant improvements in sensitivity and recall. This outcome introduces another layer that should be considered in future reviews: assuming we can predict the error type in real-time with acceptable performance, what is the impact on the overall performance of the radiologist?.

4.4.4. *Eye gaze role in error detection*

Eye movements contain information about radiologists' attention, and the attention they give to each image patch is directly related to the type of error or detection they perform in that region. Deep learning models can analyze the image patches selected by radiologists' attention and perceive the errors they are more likely to induce by their spectral and texture features. In addition, gaze can be useful to perceive regions that should receive a more thorough search, aiding radiologists to conduct a more complete scan of X-ray data in real-time. Gaze numerical features like dwell time and distance between fixations also showed information on the presence and nature of errors.

4.5. *Fatigue estimation*

Radiology's demanding and strenuous nature often leads to fatigue, contributing to costly errors. As such, the studies summarized in Table 8 aim to quantify this fatigue by analyzing radiologists' eye gaze. These works propose methods to predict the quality of diagnosis through fatigue estimation, utilizing the analysis of eye movements. This approach is advantageous as it allows for data collection without disrupting the radiologists' natural workflow or negatively influencing their diagnostic process. (see Table 9).

4.5.1. *Architectures and techniques*

The three studies included in Table 8 ([58], [59,60]) all focus on chest X-rays and utilize lung segmentation to estimate lung coverage (the area of the lungs the radiologists' gaze covers upon examining the x-

Table 8
Literature on fatigue estimation. Consult A.

Dataset	Papers	Fixation/Image Processing	Input	Model	Analysis	Fatigue Predictors	Criteria
CheXpert [81], RSNA [2], SIIM-ACR [75]	[58]	Gaze and diagnosis feature extraction Lung Segmentation	CXR	U-Net [79]	Fatigue quantification with gaze	Lung coverage, Heatmap distribution, Gaze numerical features	Correlation predictors/ fatigue
VBD [87]	[59]					Lung coverage, Information gain, Numerical features	
	[60]					Lung coverage	Lung coverage/N° XRs read linear regression

Table 9
Literature on expertise level analysis. Consult A.

Dataset	Papers	Gaze/Image Processing	Input	Model	Analysis	Loss	Metrics
Private	[48]	Extraction of participants Scanpaths during image analysis	OPT, Student/ Expert Fixations	VGG [116]		NA	ACC
	NA	OPT, Student/Intermediate/ Expert Saccades	LSTM [126]	Expertise prediction through gaze features	CE	ACC, Recall, Precision	
	[50]		CXR		Study of statistical differences in gaze between expertise levels	NA	
	[51]		CXR,MSK				

rays). The primary objective of Pershin et al. [58] and Pershin et al. [59] is to identify the most effective measure for estimating fatigue. These studies achieve this by comparing the evolution and correlations of various potential predictors with the number of chest X-rays examined by the radiologist, which is considered the ground truth for fatigue progression. Pershin et al. [59] also introduced information gain as a measure of the evolution of lung coverage throughout the diagnosis. Meanwhile, Pershin et al. [60] directly investigates the effectiveness of lung coverage as a fatigue predictor by studying its evolution concerning the number of chest X-rays viewed by radiologists.

4.5.2. Datasets and data processing

The authors of the three studies sourced their eye gaze data from publicly available chest X-ray datasets, such as RSNA, Society for Imaging Informatics in Medicine-American College of Radiology (SIIM-ACR), and CheXpert (all used by [58]), as well as VinDr-CXR (used by [59,60]). They employed a U-Net model for lung segmentation. In addition to lung coverage, [58,59] also derived numerical gaze measurements such as blink rate, average gaze, heatmap area, and gaze travel distance. Furthermore, Pershin et al. [59] utilized fatigue measurement questionnaires like the Stanford Sleepiness Scale (SSQ) [123], the Digit Symbol Substitution Test (DSST) [124], and the Reaction Time Test (RTT) [125] as independent approximations of fatigue for comparison purposes, separate from eye gaze data.

4.5.3. Results and interpretation

The convergent methodologies employed in these studies lend credence to the hypothesis that a somewhat standardized technique for estimating fatigue through gaze data is emerging. Pershin et al. [58] discovered a strong negative correlation between lung coverage and the number of chest X-rays viewed by the radiologist, identifying lung coverage as the most effective predictor compared to other metrics. Similarly, Pershin et al. [59] noted a robust relationship but found that information gain, a more comprehensive iteration of the information encapsulated by the lung coverage measure, was superior. Pershin et al. [60] further corroborated the efficacy of lung coverage, and by extension, lung coverage derivative predictors, as reliable approximations of fatigue.

4.5.4. Eye gaze's role in fatigue estimation

Gaze data shows a promising role in radiologists' fatigue estimation. It enables the estimation of the area the radiologist covered during the diagnosis process, which showed a direct correlation with the physicians' mental exhaustion. Subsequently, eye movements enable the creation of a system that analyzes the diagnosis quality of the radiologists' in real-time, which is extremely important in a profession with such a pronounced tendency to extreme workloads and low margin of error.

4.6. Expertise level analysis

In a field where errors have the potential to come at a very high cost, the current techniques to measure the experience of a radiologist are mainly subjective and non-comprehensive. Recording eye gaze data has the advantage of not interfering with the natural workflow of the physicians. It enables us to learn to analyze the level of expertise of a subject, what they lack in their search patterns, and what strategies are associated with higher performance to educate radiologists and build more efficient automated diagnosis systems with expert-like attention.

4.6.1. Architectures and techniques

We find two main themes in the expertise level analysis literature: solutions that use gaze features to predict radiologists' level of expertise (Castner et al. [48,49]) and the study of the statistical differences between expert and novice gaze behavior(Donovan and Litchfield [50], McLaughlin et al. [51]). The two solutions on the first theme use different types of gaze data: one uses fixation scan paths to cluster study subjects into student and expert categories(Castner et al. [48]), and another uses saccadic features as input for an LSTM network that conducts an expertise class prediction between student, intermediate and expert. The authors of Castner et al. [48] also perform class prediction using their scan paths and a KNN algorithm. Both works analyze performance and possible expertise-discriminating features on the second theme, including mean fixation duration, number of fixations, and time-to-first hit the lesion. Donovan and Litchfield [50] used JAFROC to conduct performance analysis, while McLaughlin et al. [51] used One-way ANOVA analysis, Kruskal–Wallis, and Mann–Whitney U tests to find significant differences between expertise levels. In addition, the authors of McLaughlin et al. [51] verified if the accuracy and confidence of each group could be reflected in the variability of their eye tracking heat maps.

4.6.2. Datasets and data processing

The four studies used datasets from private origins and collected their gaze data. In the first group of analysis, data processing consisted of extracting fixation locations in the form of square patches with the

fixation point in their centers(Castner et al. [48]) and extracting saccadic features from gaze data(Castner et al. [49]). These features were: saccade length, saccade amplitude, saccade acceleration average, saccade acceleration peak, saccade deceleration peak, saccade velocity average, saccade velocity peak, and position of saccade peak velocity. Both studies focused on Optical Projection Tomography(OPT) imaging and analyzed only the first 45 s of the participants' gaze. The second group of analysis consisted of statistical studies, where the collected data were diagnostic and gaze numerical features such as accuracy, diagnostic confidence, time to first fixation, dwell time, and other already referred measurements. Numerical features associated with gaze and fixations came directly from the eye tracker. McLaughlin et al. [51] computed fixation heatmaps to analyze statistical differences between expertise levels and included Musculoskeletal(MSK) imaging in their experiments along with chest X-rays, while Donovan and Litchfield [50] focused on chest X-rays only. The expertise levels classes for each case were(in crescent order of experience): dentistry students and professional dentists (novice and expert)(Castner et al. [48]), sixth and tenth-semester dental students and experienced dentists (novice, intermediate, and expert)(Castner et al. [49]), 1st and 3rd-year radiographers and expert radiologists(Donovan and Litchfield [50]), and students and radiographers and reporting radiographers(McLaughlin et al. [51]).

4.6.3. Results and interpretation

There is a disagreement between Donovan and Litchfield [50] and McLaughlin et al. [51] about using gaze information to predict expertise, although both focus on OPT imaging. Castner et al. [49] states that, although the spatial information in scan paths contains important features for expertise analysis, saccade behavior over time contains patterns related to key intervals in expert visual search. The two studies do not establish a strong hypothesis regarding the best approach. Using scan paths, Castner et al. [48] was able to cluster experts with significantly high accuracy. At the same time, there is a clear distinction between expertise levels in the evolution of features like saccade length, average velocity, and peak velocity over time, hinting at the predictive power that both information sources contain. Experts exhibit similar behavior, while students are diverse, and, therefore, easier to classify as non-experts(Castner et al. [48]). In addition, how easily experts and novices can be distinguished depends on the image chosen for examination(Castner et al. [48]). Donovan and Litchfield [50] and McLaughlin et al. [51] focused on statistical differences in gaze characteristics. Contrary to Castner et al. [48] and Castner et al. [49], they worked with chest X-rays and musculoskeletal imaging. The authors observed that the time participants arrived at the lesion's region did not significantly change with experience. However, once seen, participants with more experience tended to fixate more on the lesion's region, given that they more easily recognize it as interesting. This resulted in features such as "time to first hit" not having predictive power, contrary to measurements like mean fixation duration and mean fixation count on lesions. This is also compatible with the error-gaze relationship analysis on Castner et al. [52] that establishes that radiologists tend to fixate more on marked lesions. However, both studies also refer to the difficulties associated with inter-observer variability(Donovan and Litchfield [50]), different learning rhythms between expertise levels(Donovan and Litchfield [50]), and similarities between naive observers and experts when trying to modulate expertise levels([49,50]).

4.6.4. Eye gaze role in expertise level analysis

Both fixation and saccadic eye movement data are shown in the literature to contain important information regarding the radiologists' expertise. Gaze characteristics, number of fixations, saccade length, and dwell time are directly tied to performing strategies demonstrating expertise. Eye gaze also enabled researchers to analyze diagnosis behavior associated with higher performance and experience, permitting radiology students to compare their search patterns with experts and mimic their techniques.

4.7. Other tasks and themes

The studies we discuss in this section approach unique themes and could not join a larger group for comparison. Moreira et al. [73] provided an overview of the potential applications of VR (virtual reality) and eye-tracking technology in radiology. Most eye gaze collection protocols in studies regarding eye tracking for automated solutions in radiology rely on desktop eye trackers, and the authors show that most of their disadvantages, like limited permitted head movement, are overcome with a VR setting. Bhattacharya et al. [69] and Lanzer et al. [72] expand on two applications: training visual search patterns in radiology students and studying the visual search patterns of expert surgeons during procedures. The first included two techniques of EMME (Eye movement modeling example) based training interventions: one showing moving fixation on a blurred and another on a non-blurred background, improving detection sensitivity and specificity in focal lung pathology on chest x-rays. In contrast, the second performed a statistical analysis regarding gaze in catheter-based cardiovascular interventions. On the other hand, Drew et al. [70] studies the downsides of one of the applications of gaze data in automated healthcare. The authors of this work evaluated the prejudicial bias-aiding detection systems that can be introduced in the radiologist's diagnosis. They found that the aiding system influenced the radiologist's eye movements by diminishing their search area. Panetta et al. [74] created a dataset of dental X-rays with expert gaze data in the form of fixation heatmaps.

4.7.1. Results and interpretation

Bhattacharya et al. [69,72] established additional eye gaze radiology-related applications to which our survey did not give enough relevance, namely medical image interpretation training and access to cognitive process information during surgical procedures, obtaining results that elucidate the potential of the technology in these fields. Eye tracking recording technology has limitations, and Moreira et al. [73] provided a possible solution for some of them by having the eye tracker directly on the subject's head in a VR setting. Drew et al. [70] focused on one of the potential dangers of the solutions we described in the previous section by investigating the possible bias these systems can introduce into the human examination of medical imaging.

5. Implementation recommendations

Beyond offering a detailed taxonomy of gaze-integrated solutions across various tasks in the literature, this survey aspires to impart insights into the practical implementation of these solutions. However, our literature selection fell short of providing holistic pipelines emblematic of best practices in the domain of gaze utilization and processing in deep learning for automated radiology. Therefore, our focus is narrowed to presenting recommendations based on techniques that consistently showcased superior results across different tasks.

5.1. Recommendations regarding classification using gaze

Drawing from the literature we analyzed concerning solutions that perform classification using gaze data, we advocate the following strategies:

- Employing transformer models, given their innate attention mechanism that intuitively aligns with the attention patterns of radiologists.
- Using a leveraging model (analogous to Wang et al. [38]) so that it selects only useful features for classification.
- Utilizing fixation heatmaps for processing gaze data, given their dual advantages of enhanced interpretability and dimensionality reduction.

- Incorporating attention zone cropping, as was described in [35–37] since it is an effective way to introduce gaze into the learning process and to reduce the tendency ViTs have towards shortcut learning.
- Introducing a loss term quantifying the difference between the models' saliency maps and fixation heatmaps tends towards better results since the encoder feature extraction layer may not be affected by the secondary task in the case of encoder-decoder hard parameter sharing framework.
- Using cross-validation to evaluate your models to avoid misleading results caused by lack of generalization and consistency.

5.2. Recommendations regarding object detection using gaze

Drawing from the literature we analyzed concerning solutions that perform object detection using gaze data, we recommend the following:

- Using radiomics for global feature extraction and a multi-resolution architecture like Teacher-Student Global-Focal Transformer, since radiomics are informative of semantic context and global-focal mechanism are analogous to how radiologists process medical images.
- Using fixation heatmaps to process gaze information because of their interpretability, dimension reduction, and regions of interest indicative properties.
- Using radiologist's fixations during interpretation as a more efficient method to obtain location labels.
- Considering graph-sparsified to process gaze information because of its dimension reduction and regions of interest indicative properties.

5.3. Recommendations regarding attention analysis using gaze

Drawing from the literature we analyzed concerning solutions that perform attention analysis using gaze data, we recommend:

- Comparing radiologists' fixation heatmaps to saliency maps to verify whether the model selects an efficient representation of the input data, including in gaze-free models.
- Using ensemble frameworks to force model consistency and thus the convergence into selecting an efficient input data representation.
- Using shuffled metrics that penalize proximity to biased image regions to obtain a valid comparison between radiologists' fixation heatmaps and models' saliency maps.
- Using global-focal frameworks to learn radiologists' attention more efficiently since it is a processing mechanism analogous to the one they use.

5.4. Recommendations regarding error detection using gaze

Drawing from the literature we analyzed concerning solutions that perform error detection using gaze data, we recommend the following:

- Performing attention-based clustering as Mall et al. [53], Mall et al. [54], and Mall et al. [55] to obtain image patches containing features relevant to error detection.
- Extracting image features with Gábor filters (for spectral features), Grey level co-occurrence matrix (for texture features), and numerical gaze features like dwell time and the average distance between fixations inside fixation clusters since deep learning models can fail at arriving at an adequate set of features for their abstract and specific nature.
- Performing N-way ANOVA analysis and PCA to distinguish statistically significant features that contain relevant predictive information.
- Finding the best performing features-model pair for each radiologist to achieve model personalization.

5.5. Recommendations regarding expertise level analysis using gaze

Drawing from the literature we analyzed concerning solutions that perform expertise-level analysis using gaze data, we recommend:

- Selecting difficult-to-analyze images capable of distinguishing expertise.
- Using saccadic features such as saccade length, average velocity, and peak velocity to predict expertise since they have been successfully used to predict radiologists' expertise.
- Using fixation scan paths and features like mean fixation duration and mean fixation count to predict expertise since they have been successfully used to predict radiologists' expertise.

5.6. Recommendations regarding fatigue estimation using gaze

Drawing from the literature we analyzed concerning solutions that perform fatigue estimation using gaze data, we recommend:

- Performing Lung segmentation with a U-Net-based model to obtain lung coverage.
- Obtaining lung coverage or lung coverage-related metrics like information gain, since there is a strongly negative correlation between them and the number of X-rays the radiologists' analyzed.

6. Answers to Research Questions

6.1. What architectures and fusion techniques are available to integrate eye-tracking data into deep learning approaches to localize and predict lesions?

Gaze data contain valuable insights regarding the radiologists' mental processes preceding the assigning of a diagnosis. However, efficiently extracting this information requires an adequate architecture. Architectures can adopt a multitask or a single-task fusion framework to achieve this. In this discussion, we will first delve into the multitask architectures we have found, followed by the single-task architectures, and conclude with the differences we have perceived between the problems of image classification and lesion detection solutions in healthcare. The multitask fusion technique creates extra tasks to arrive at the most valuable data representations through additional angles, resulting in a more generalized learning process. In our literature, it is also possible to envision it as a knowledge transfer technique between the radiologist and the model. Extra tasks frequently involve learning to predict radiologists' attention, enabling the model to use the same image characteristics as radiologists to make predictions. We find three architectural solutions that follow this definition in the literature, namely: Encoder-Decoder(5), Teacher-Student(6), and Non-ED hard-parameter sharing(7).

Encoder-decoder and Non-ED hard-parameter sharing employ highly similar fusion techniques. Both architectures consist of a common set of shared layers. These shared layers feed into two distinct components: one is responsible for predicting the distribution of radiologists' gaze, while the other handles making predictions. The purpose of having common feature extraction layers is to have both tasks influence input data representation selection. However, the Encoder-Decoder and Non-ED hard-parameter sharing differ in their components to perform the extra task. While Encoder-Decoder uses a decoder to achieve this goal, Non-ED hard-parameter sharing has a more versatile definition that allows for more specific implementations.

Relative to the Encoder-Decoder architecture, from the trio of studies Agnihotri et al. [31], Karargyris et al. [32], Watanabe et al. [17], we have observed that to guarantee the significant influence of the extra task into the global feature extraction layers, it is essential to include an additional loss term consisting of the distance between the models' saliency maps and the corresponding radiologists' attention map. Relative

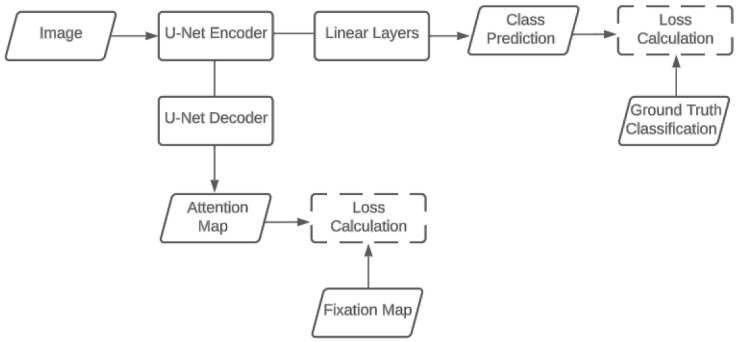

Fig. 5. Encoder-Decoder architecture flowchart representation.

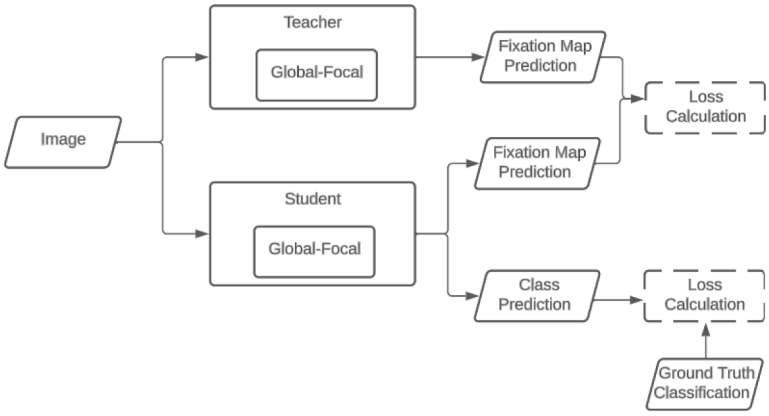

Fig. 6. Teacher-Student architecture flowchart representation.

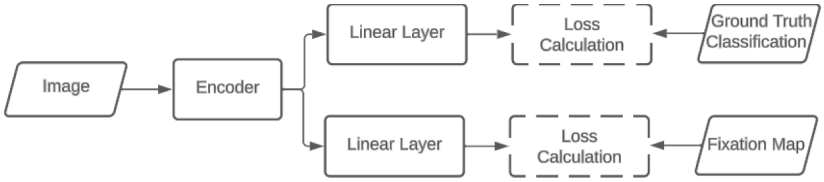

Fig. 7. Non-ED Hard-Parameter Sharing architecture flowchart representation.

to the Teacher-Student architecture, the teacher is trained first to learn to predict the experts' attention. The student learns to perform the main task (classification or lesion detection) and to predict radiologists' attention. The loss term relative to the extra task comes from comparing the student's and the teacher's predicted attention. In addition, these blocks have SEMA connections to allow them to share information.

We also found three single-task architectures in the literature: attention zone cropping(8), multimodal DL(9), and the saliency map to human attention architecture(10). These are more diverse from each other than the multitask architectures, and the last is even very analogous to their framework. Instead of having a task-specific set of layers, SM to HA(Saliency Maps to Human Attention) architectures add a saliency map generator to their feature extraction layers and the difference between its output and the radiologists' attention to the models' loss function. On the other hand, the first two referred approaches use the experts' attention maps right from the input: attention zone cropping by using them to crop out non-relevant regions and multimodal by having an additional arm to directly obtain data representations from it that

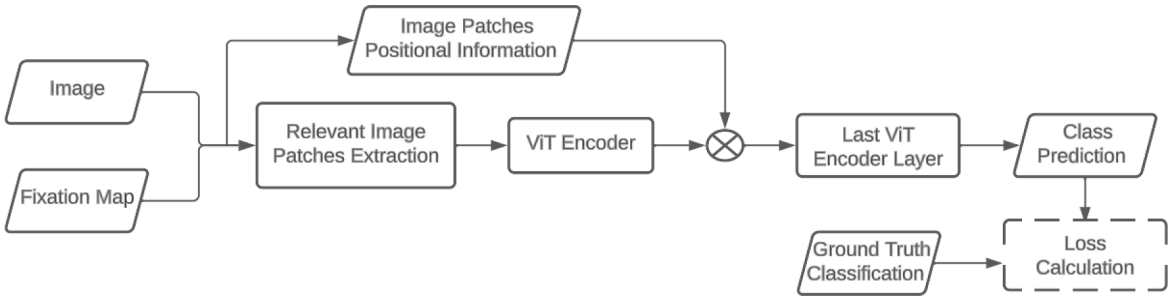

Fig. 8. Attention zone cropping architecture flowchart representation.

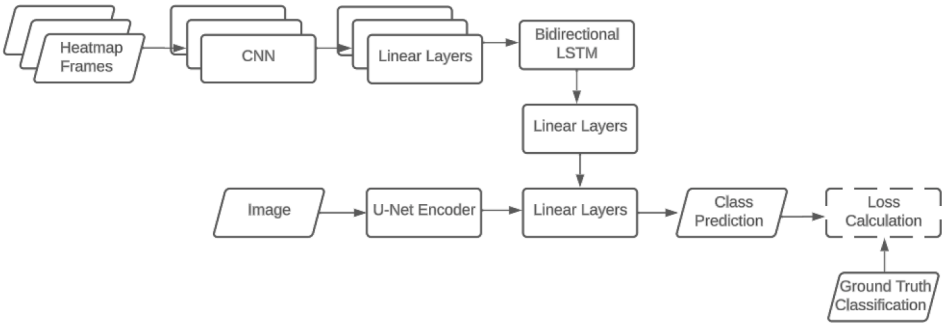

Fig. 9. Multimodal architecture flowchart representation.

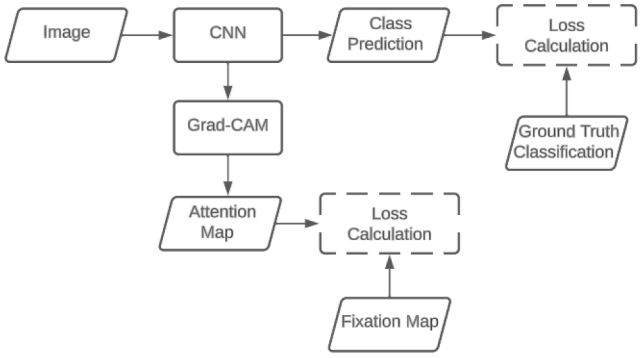

Fig. 10. SM to HA, an architecture that directly approximates Grad-CAM saliency map to radiologists' fixation maps, flowchart representation.

will be concatenated to the segment that processes the original input image.

Almost all the architectures surveyed in the literature use fixation heatmaps as a proxy for the radiologists' attention and have a segment and/or loss term dedicated to learning to approximate this attention as a means to incorporate it into the learning process of the main task. Although most came first or exclusively from the literature on classification (with few, like Teacher-Student, having instances on both tasks), the ideas behind them also apply to lesion detection. However, lesion detection necessitates additional information extraction from gaze data, particularly regarding lesion location. Subsequently, some works attempt to use gaze data directly to obtain regions of interest before thoroughly analyzing those regions for lesion detection in an explicit global-focal strategy. A notable example of this strategy is the work done in Khosravan et al. [43], where the researchers extract regions of interest directly from the graph representation of their gaze data before applying a network performing false positive reduction and lesion segmentation to each.

Contrary to classification, the literature we reviewed on object detection was insufficient to establish patterns and compare approaches comprehensively. We need to analyze more works on the topic to more thoroughly investigate the differences between the two tasks and the limitations classification-original approaches deal with in a lesion detection setting.

6.2. How is eye-tracking data pre-processed before being incorporated into multimodal deep learning architectures?

The first layers of deep learning models typically feature extraction layers that attempt to automatically extrapolate the most meaningful data representations of the input for the task. This mechanism has the advantage of reducing the bias characteristic of manually chosen features and obtaining a personalization of the data representation relative to a task that is difficult to reproduce in a non-automatic mechanism. However, learning to extract meaningful features from gaze data can be daunting, given its complexity, individual variability, and task dependence. Thus, the correct pre-processing of this data is essential to secure a successful introduction into a model's learning process.

The first step in the literature we have reviewed is always fixation filtering. Fixations are frequently stated to be the most representative aspect of radiologists' attention in gaze data. In most cases, filtering takes place in the eye tracker itself. Despite its significance, we considered the algorithms that execute this filtering out of the scope of this paper.

We found two main techniques for processing fixation data: fixation heatmaps and attention-based clustering. Fixation heatmaps are created by superimposing Gaussian distributions, centered at each fixation point, onto the input image. The intensity of these distributions is proportional to the duration of each fixation. The literature divides them into static fixation heatmaps and temporal fixation heatmaps. The difference resides in displaying the full fixation data into one image or separating it temporally according to a certain frequency. Fixation heatmaps typically appear in the literature regarding classification, lesion detection, and attention analysis.

On the other hand, we have attention-based clustering. It divides the image by foveal-sized regions according to the number of fixations they have received. Typically, the categories are foveal clusters (FCs), peripheral clusters (PC), and never-fixated clusters (NFC). FCs contain at least three temporally sequential fixation points. PCs and NFC have less than three (at least one) and no sequential fixations, respectively. This technique is most present in the error detection and attention analysis literature.

However, a relevant example of adapting this technique to lesion detection emerges in Khosravan et al. [43], instead of having several categories, their algorithm uses fixations to select several ROIs as graph vertices. Then, it applies a sparsification technique that reduces the complexity of the data without altering its structure similarity, which means it reduces the amount of redundant data. The authors follow by performing false positive reduction and lesion segmentation to the regions the fixation processing algorithm classified as ROIs. This technique is an example of how the location-specific information attention-based clustering gathers can complement the global context information of fixation heatmaps to provide more adequate fixation processing techniques for lesion detection.

6.3. How can eye gaze data promote explainability in multimodal deep learning architectures?

By incorporating eye gaze data, the features deep learning models select align with the image characteristics radiologists deem relevant for diagnosis. Thus, these models become more interpretable and their decision process more transparent. This allows radiologists to peek inside the models' "black-box", comparing the models' insight with their own, not only enabling the detection of neural networks' malfunctions but also adding to radiologists' decision processes as a verifiable second

opinion that leads experts to reflect upon their own decisions and to examine more thoroughly regions the model deemed to be important. Eye gaze data also promotes explainability for deep learning model architectures that do not include gaze. A good-performing model's saliency maps tend towards the radiologists' gaze distribution on the same inputs. Subsequently, it is possible to use eye-tracking data as a baseline to evaluate models' training process. The main processing technique for processing gaze is the computation of fixation heatmaps, enabling their interpretability, region of interest indicative, and dimensionality reduction properties to arise in the models' saliency maps. Initially, the introduction of gaze data into models' learning processes aimed to increase model explainability while not affecting its performance. However, the research we have reviewed suggests that these two properties go hand in hand. By aiding the model to arrive at a cleaner representation of features needed for the task (that is, the ones radiologists' tend to focus on), eye-tracking data enables models to not only have more medically meaningful saliency maps but to achieve higher scores in performance metrics.

However, gaze data also carries radiologists' biases. Effective processing and architectural design are crucial to allow models to arrive at an adequate equilibrium between the features they learn and those they obtain through the fixation heatmaps.

7. Future research opportunities

While our study has provided an extensive review of the use of eye gaze data and its integration into deep learning architectures for radiology applications, several promising research directions can extend the capacities and aptitudes of this interdisciplinary field.

7.1. Possible alternative approaches to classification and object detection using gaze

Eye gaze data provides the radiologists' expertise as prior knowledge to influence the models' training process direction. However, no radiologist establishes diagnosis on image data alone, nor is expertise transmitted in the vacuum without establishing the biological significance of the regions we refer to. Deep learning models have to consume large amounts of data compared to humans because they lack basic intuition about the world in which they are working. Model frameworks like Bayesian neural networks and Physics-Informed Neural Networks (PINNs) are two solutions that aim to increase the models' prior knowledge, thus conducting a more interpretable model that needs less training to arrive at higher performances. Prior knowledge appears in Bayesian neural network frameworks as conditional distributions over the models' weights. This framework feature enables the model to include insights about the properties of the system it is analyzing. These can be known pathology properties, anatomical structures, or the type of attention certain characteristics usually gather from radiologists.

On the other hand, PINNs incorporate prior knowledge by having a loss term that assures that the solutions satisfy the physical properties we already know a solution should have. In this way, they are more prone to be useful in introducing knowledge about the dynamics and relationships of biological systems. More recently, Large Language Models(LLMs) have been having successfully employed in the medical field to further enhance the integration of prior medical knowledge into the training and decision-making processes of deep learning models ([127,128]). Their language processing and knowledge incorporation ability makes them a must in the future research for explainability and prior-medical-knowledge incorporation into deep learning models. Although they have some difficulties remaining factual, approaches that attempt to reduce their tendency to generate inaccuracies ([129]) are already being developed in order to enable this technology to reach their full potential. These three approaches are examples of techniques with the potential to provide models with an intuitive understanding of the pre-conditions of the biological systems they are analyzing, providing more medically meaningful predictions that are, therefore, more interpretable.

7.2. Explainable user interfaces for medical training

With the growing complexity of multimodal fusion algorithms and the inherent "black-box" nature of some DL models, there is a pressing need to develop interfaces that can provide mechanisms for the user to scrutinize the model's predictions [8,13]. While the domain of explainable AI (XAI) offers methodologies to distil a sub-symbolic linguistic representation (often manifesting as association rules amongst features or attention maps in visual data) [130,131], the challenge resides in approaches that can map this sub-symbolic representation into a semantic construct that is understandable to human cognition in an explainable user interface (XUI), thereby facilitating rigorous scrutiny [132,133]. In medical training, XUIs combine domain-specific nuances inherent to medical data, providing explanations that resonate with expert interpretations [134,135]. The diverse expertise of users, from novices to experienced radiologists, amplifies the challenge, necessitating adaptable interfaces that cater to this broad spectrum. Eye-tracking offers a potential solution by capturing users' gaze patterns, shedding light on their cognitive processes and areas of focus [19]. This technology can guide XUIs to deliver tailored, context-aware explanations, especially by contrasting expert and novice gaze patterns. For instance, expert gaze patterns can serve as benchmarks, helping to identify areas where novices might need more guidance or clarification. Through such synergy, eye-tracking with XUIs can elevate the efficacy of DL models in medical training, paving the way for personalized and contextually rich explanations.

7.3. Towards human-centered AI for medical diagnosis

Human-centered AI refers to AI systems designed and deployed primarily focusing on human values, needs, and capabilities. Rather than viewing AI as an autonomous entity, human-centered AI emphasizes collaboration, where the technology synergizes with human intentions, thus augmenting human capabilities Shneiderman [16].

In medical applications, it is pivotal that the locus of control remains firmly with the human. DL models should be regarded as decision support systems. Their role is to complement the intricate work of radiologists, not to substitute them. By ensuring that DL models are auxiliary tools, we underscore the invaluable expertise and judgment that medical professionals bring to patient care. This collaborative approach enhances diagnostic accuracy and fosters trust between practitioners and their technology.

A salient application of human-centered AI emerges in the realm of medical training [136]. Here, AI systems can be meticulously tailored to individual radiologists, whether they are seasoned experts or budding novices. AI systems have also been trained to generate reports automatically, a tiresome and inefficient process for radiologists[137]. In addition, by leveraging data from eye-tracking technology, which offers insights into a radiologist's cognitive patterns and areas of focus, these AI systems can personalize these reports' generation. Such reports can highlight areas that align or diverge from expert gaze patterns, offering targeted feedback and guidance. Another relevant example are Computer Aided Diagnosis(CAD) systems that assist radiologists in the interpretation and analysis of medical images, acting as a second opinion. This technology can also be integrated with eye-tracking technology, through using the radiologists' patterns to discover zones of the image that require more examination and segment/classify suspicious regions that attracted the practitioners' focus[43]. Nevertheless, it is also important to keep in mind that these systems can also add negative effects amid the positives should they not be implemented correctly [70].

This fusion of AI and eye-tracking, although it is an open research question, promises to provide a more individualized, effective, and

human-centric approach to medical training.

7.4. Virtual reading rooms for medical diagnosis

Integrating eye-tracking with advanced AI and virtual 3D imaging in CXR diagnostics heralds the dawn of innovative virtual reading rooms (VRRs) tailored for medical diagnosis. As we witness an accelerating digital transformation in healthcare, VRRs have become a focal point, promising a leap in diagnostic accuracy, efficiency, and immersive user experience. Herein, we illuminate prospective avenues of exploration and development in this domain.

Personalized User Interfaces (UI). Harnessing eye-tracking data, VRRs could dynamically adjust UI elements based on the diagnostician's viewing patterns. This ensures that pertinent information remains at the forefront, optimizing the visual workspace for enhanced efficiency.

AI-driven Diagnostic Suggestions with 3D Imaging. Monitoring gaze patterns with 3D CXR visualizations allows AI algorithms to pinpoint areas of heightened interest or uncertainty. This integration can then lead to timely diagnostic suggestions, highlighting specific regions of the X-ray that mandate a more in-depth evaluation.

Collaborative Virtual Environments with VR Integration. VRRs could serve as shared virtual spaces, enabling teams of diagnosticians to review and deliberate on cases jointly. Synchronized eye-tracking data can align their perspectives, ensuring synchronized focus on pertinent regions.

Training and Feedback Mechanisms. By juxtaposing the gaze patterns of novices against seasoned experts within a VR environment, VRRs can offer invaluable feedback. This real-time critique can pinpoint overlooked or misinterpreted areas, fostering a richer learning experience.

Ergonomics and Fatigue Monitoring: Vigilance is paramount in diagnostics. With extended reading sessions potentially inducing fatigue, VRRs can use eye-tracking metrics, such as blink frequency or fixation duration, to gauge alertness, proposing timely breaks or alerting the diagnostician to potential lapses in concentration.

User Experience (UX) Optimization with VR Environments: For VRRs to gain traction, they must promise a fluid UX. Endeavors should aim to refine the interface, curtailing latency and facilitating effortless integration with existing hospital IT ecosystems.

7.5. Remote collaboration for medical diagnosis

The rise of digitization, propelled by the recent pandemic, has accelerated the adoption of remote working environments in many fields, including radiology. However, the practical application of remote collaboration in radiology poses unique challenges, particularly when leveraging the collective knowledge of multiple radiologists in different geographical locations. We believe several promising research directions could extend deep learning and eye-tracking capacities and aptitudes in this context.

Enhanced Communication through Shared Gaze Data Currently, communication between radiologists in a remote setting relies heavily on textual or verbal exchanges. Future work could focus on utilizing shared gaze data to facilitate visual communication. By allowing radiologists to see where their colleagues are looking at the medical image, they can better understand their thought processes, leading to more effective collaboration and potentially more accurate diagnoses.

Real-Time Decision Support Systems Building on the real-time diagnostic support systems concept, future studies could focus on developing systems that support real-time collaboration between multiple radiologists. Such systems could leverage deep learning models to analyze the gaze patterns of multiple radiologists in real time and generate consensus predictions or highlight areas of discrepancy for further discussion. This could facilitate effective decision-making, even when the experts are distributed across different locations.

Machine Learning Models for Conflict Resolution Another research direction could be developing machine learning models specifically designed to resolve conflicts in diagnoses. By analyzing the gaze data and the associated diagnoses of multiple radiologists, these models could learn to identify situations where there may be uncertainty or disagreement and suggest the most probable diagnosis based on the collective gaze and decision data.

Privacy and Security in Remote Collaboration While the possibilities of remote collaboration are enticing, they come with their own set of challenges. The privacy and security of patient data in a remote setting is a paramount concern. Future work could explore innovative ways to ensure data security while enabling effective collaboration. Additionally, techniques to anonymize gaze data while maintaining its usefulness for collaboration and machine learning could be an important research direction.

7.6. Ethical considerations

While promising, the intersection of deep learning, eye gaze data, and radiology comes with significant ethical considerations that must be carefully addressed. We outline a few key areas below:

Privacy and Consent The use of eye gaze data implies a certain level of invasion of the privacy of individuals. While this data is collected to improve radiological practices, obtaining explicit consent from the individuals involved is crucial. Precise information regarding the data purpose, storage, sharing, and future use should be communicated effectively. Furthermore, robust measures need to be in place to ensure the de-identification of gaze data and associated radiological images so that the privacy of patients is not compromised.

Security Medical data transfer, storage, and processing, including eye gaze data and radiological images, should comply with data protection regulations such as GDPR and HIPAA. The risk of data breaches and misuse cannot be underestimated. Hence, stringent security measures are essential. These include secure data transfer protocols, encrypted storage, and restricted access.

Fairness and Bias Deep learning models are known to be susceptible to biases in the training data. Therefore, care should be taken to ensure that the eye gaze data collected represents a diverse range of radiologists with varying experience levels, specialties, and backgrounds. A biased dataset could lead to models favoring specific eye gaze patterns, disadvantaging certain radiologists or patients.

Transparency and Accountability The 'black-box' nature of many deep learning models can create accountability issues. It is necessary to ensure that the use of these models in radiology does not undermine the responsibility of healthcare professionals in medical decision-making. Additionally, there should be transparency about when and how these models are used so that patients and practitioners can make informed decisions about their healthcare.

Human Autonomy While the automation of radiology tasks can help increase efficiency and address shortages of trained radiologists, care should be taken to ensure that it does not undermine the autonomy of healthcare professionals. It is essential to strike a balance where these systems aid the radiologists in their decision-making process rather than replace them.

These ethical considerations provide a roadmap for future research and development. We believe that integrating deep learning and eye gaze data in radiology can only reach its full potential if these ethical considerations are given the importance they deserve.

8. Conclusion

In conclusion, deep learning techniques show tremendous potential in revolutionizing various domains, including radiology. The increasing number of imaging examinations globally and the shortage of medical professionals call for innovative solutions to improve healthcare services.

The lack of medical knowledge and base "common sense" in AI

models presents significant challenges in achieving deployment levels in the medical field, where the risk tolerance is very low, training data is scarce, and deep learning networks' "black-box" nature does not allow insight into their decisions. Introducing gaze data into their learning process directly tackles these challenges by helping the models arrive at cleaner versions of the data representations already essential to their successful training. It does not merely contribute to improving model explainability or performance. It bridges the two by guaranteeing the models' ability to converge at medically meaningful features that radiologists can understand and are more useful for the task.

This happens because eye-tracking data contains information about the factors radiologists account to make their decisions, leading the incorporation of this data into deep learning models to be a type of knowledge transfer that significantly diminishes the training data needed for quality convergence. In other applications like error detection, fatigue analysis, and expertise level analysis, it enables models to analyze the quality of the radiologists' decision processes and provide real-time support to mitigate the risk of error and non-comprehensive analysis.

Within this survey, we provide a systematic and comprehensive taxonomy that organizes architectures and gaze-processing techniques according to the associated task. This allowed us to perform a comparative analysis to obtain insights about the best practices for implementing solutions for each case and share them as recommendations.

The literature we reviewed provides foundational insights, yet a significant opportunity exists to establish further comprehensive guidelines on optimal pipelines for each task. The potential of incorporating gaze data into deep learning processes is vast, and many promising applications await deeper exploration. Continued research and collaboration between the medical and AI communities are crucial to unlocking the full potential of gaze data and ushering in a new era of intelligent and efficient healthcare approaches.

Declaration of Competing Interest

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

Acknowledgements

Omitted due to double blind review.

Appendix A. Metrics, Machine Learning and Statistical Analysis

In the realm of machine learning and statistical analysis, the assessment of model performance and the selection of appropriate loss functions are paramount. Their roles are somewhat related, but while metrics evaluate the model's performance after training, loss functions guide the model to increase its predictive power. However, some concepts can perform both roles (as displayed in the corresponding tables). This section will present the mathematical expressions of the metrics and loss functions in the literature. In addition, we will also give a brief introduction to some machine learning and statistics concepts that may need some clarification. In the following expressions:

- n is the number of data points.
- i is a data point element.
- j is a class.
- y_i is the element i of the set of ground truth values.
- y is the element i of the set of predicted values.
- p_j is the probability for the j^{th} class that the model predicted.
- y_i^f is the radiologists fixation map for data element i.
- $\widehat{y}_i^f$ is the fixation map the model predicted for the data element i.
- y_i^s is the model's saliency map for the data element i.
- $\mathrm{cov}(X,Y)$,$\sigma(X)$, and $\mu(X)$ are the covariance between the X and Y variables, the standard deviation of the X variable, and the mean of the X variable, respectively.
- TP is the number of true positives, FP the number of false positives, TN the number of true negatives, and FN the number of false negatives.
- A and B are two objects or regions in an image.
- C_1 and C_2 are constants to stabilize the division when the means and variances are close to zero.
- T_{ij} represents the flow or amount of pixel value transported from pixel value i in image A to pixel value j in image B, $d(i,j)$ is the distance between pixel intensities i and j.
- C is the total number of classes. $\mathrm{Precision}_j$ is the precision for the class j.

Mean Squared Error (MSE):

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{A.1}$$

MSE measures the average squared difference between the actual values (y) and the predicted values ($\widehat{y}$). It quantifies how far off your predictions are from the true values and penalizes larger errors more heavily.

Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right| \tag{A.2}$$

MAE calculates the average absolute difference between the actual values (y) and the predicted values ($\widehat{y}$). It provides a more linear and less sensitive measure of errors compared to MSE.

Cross-Entropy (CE, BCE for binary problems):

$$\mathrm{CE} = -\sum_i y_j \log\left(p_j\right) \tag{A.3}$$

Cross-entropy is used in classification problems to quantify the dissimilarity between the true class labels (y) and the predicted probabilities (p). It punishes significant errors more strongly and is often used as a loss function for training neural networks.

Soft Cross-Entropy:

$$\mathrm{SoftCE} = -\sum_i \sum_j y_{ij} \log\left(p_{ij}\right) \tag{A.4}$$

Soft Cross-Entropy is an extension of Cross-Entropy used for multi-class classification problems. It measures the dissimilarity between true class distributions (y) and predicted class probabilities (p).

Normalized Scanpath Saliency (NSS)[65]:

$$\mathrm{NSS} = \frac{1}{\sum_i y_i^f} \sum_i \frac{\widehat{y}_i^f - \mu\left(\widehat{y}^f\right)}{\sigma\left(\widehat{y}^f\right)} y_i^f \tag{A.5}$$

NSS assesses the quality of eye-tracking data by comparing the fixation locations ($\widehat{y}^f$) to their distribution in an image. It normalizes the saliency values and computes their relationship with the fixation data.

Kullback–Leibler Divergence (KLD)[65]:

$$\mathrm{KLD} = \sum_i y_i^s \log\left(\epsilon + \frac{y_i^s}{\epsilon + \widehat{y}_i^f}\right), \ \epsilon = 2.2204 \times 10^{-16} \tag{A.6}$$

KLD quantifies the difference between two probability distributions, typically used in information theory. In this context, it measures the divergence between the true saliency distribution (y^s) and the predicted distribution ($\widehat{y}^f$).

Linear Correlation Coefficient (CC)[65]:

$$\mathrm{CC} = \frac{\mathrm{cov}(y^s, \widehat{y}^f)}{\sigma(y^s)\sigma(\widehat{y}^f)}, \tag{A.7}$$

CC evaluates the linear relationship between two sets of data, here, between actual saliency values (y^s) and predicted saliency values ($\widehat{y}^f$).

Similarity (SIM)[65]:

$$\mathrm{SIM} = \sum_i \min\left(y_i^s, \widehat{y}_i^f\right) \tag{A.8}$$

SIM computes the similarity between two sets of data, typically between actual saliency values (y^s) and predicted saliency values ($\widehat{y}^f$). It measures how well the two datasets overlap.

Normalized Cross-Correlation (NCC)[62]:

$$\mathrm{NCC} = \frac{1}{P-1} \sum_i \frac{\widehat{y}_i^f - \mu\left(\widehat{y}_i^f\right)}{\sigma\left(\widehat{y}_i^f\right)} \cdot \frac{y_i^f - \mu\left(y_i^f\right)}{\sigma\left(y_i^f\right)} \tag{A.9}$$

NCC assesses the similarity between two sets of data while normalizing for their variances and means. It is commonly used in image processing and computer vision tasks.

Attention Overlap (AO)[31]:

$$\mathrm{AO} = \frac{\sum\limits_{\mathrm{i} \in \mathrm{BB}} \widehat{y}_i^f}{\sum\limits_{\mathrm{i} \in \mathrm{Img}} \widehat{y}_i^f}, \ \text{where } \widehat{y}_i^f = \begin{cases} i & i > 100 \\ 0 & i \leqslant 100 \end{cases}, \tag{A.10}$$

AO measures the overlap between the predicted saliency map ($\widehat{y}^f$) and the ground truth region of interest in an image. It's used to evaluate the performance of models in focusing on specific image regions.

Structural Similarity Index (SSIM):

$$\mathrm{SSIM} = \frac{(2 \cdot \mu(A) \cdot \mu(B) + C_1) \cdot (2 \cdot \mathrm{cov}(\mathrm{A}, \mathrm{B}) + C_2)}{\left(\mu(A)^2 + \mu(B)^2 + C_1\right) \cdot \left(\sigma(A)^2 + \sigma(B)^2 + C_2\right)} \tag{A.11}$$

SSIM is a metric to assess the structural similarity between two images, typically an original and a modified version. It considers luminance, contrast, and structure for comparison.

Earth Mover's Distance (EMD):

$$\mathrm{EMD} = \min \sum_{i_1=1}^{n} \sum_{i_2=1}^{m} T_{ij} \cdot d\left(i,j\right) \tag{A.12}$$

EMD calculates the minimum effort required to transform one distribution into another. It's used in various applications, including image matching and shape analysis.

Pearson's Correlation Coefficient (PCC):

$$\mathrm{PCC} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \cdot \sum (Y_i - \overline{Y})^2}} \tag{A.13}$$

PCC measures the linear correlation between two data sets, typically denoted as X and Y. It quantifies the strength and direction of the relationship between the two datasets.

Recall:

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{A.14}$$

Recall is a metric used in binary classification to assess the ability of a model to identify all positive instances. It's calculated as the ratio of true positives to the sum of true positives and false negatives.

Precision:

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{A.15}$$

Precision is a metric used in binary classification to evaluate the accuracy of positive predictions. It's calculated as the ratio of true positives to the sum of true positives and false positives.

Micro-Precision:

$$\mathrm{Micro-Precision} = \frac{\sum_{j=1}^{C} TP_i}{\sum_{j=1}^{C} \left(TP_i + FP_i\right)} \tag{A.16}$$

Micro-precision calculates the precision for each class and then averages them. It's commonly used in multi-class classification scenarios.

Macro-Precision:

$$\mathrm{Macro-Precision} = \frac{1}{C} \sum_{j=1}^{C} \mathrm{Precision}_j \tag{A.17}$$

Macro-precision calculates the precision for each class separately and then averages these values. It provides equal weight to all classes, regardless of their size.

Accuracy (ACC):

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{A.18}$$

Accuracy is a measure of overall model performance in classification problems. It quantifies the ratio of correct predictions (true positives and true negatives) to the total number of instances.

Specificity (Spec):

$$\mathrm{Spec} = \frac{TN}{TN + FP} \tag{A.19}$$

Specificity measures a model's ability to identify negative instances in binary classification correctly. It's calculated as the ratio of true negatives to the sum of true negatives and false positives.

False Positive Rate (FPR):

$$\mathrm{FPR} = \frac{FP}{FP + TN} \tag{A.20}$$

False Positive Rate calculates the rate of false positive predictions in binary classification. It's the complement of specificity, measuring the proportion of false positives to all actual negatives.

Negative Predictive Value (NPV):

$$\mathrm{NPV} = \frac{TN}{TN + FN} \tag{A.21}$$

Negative Predictive Value evaluates a model's ability to correctly identify negative instances in binary classification. It's calculated as the ratio of true negatives to the sum of true negatives and false negatives.

F1-Score (F1):

$$\mathrm{F1-score} = \frac{2{\cdot}\mathrm{TP}}{2{\cdot}\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{A.22}$$

F1-Score is the harmonic mean of precision and recall, providing a balance between these two metrics in binary classification.

Receiver Operating Characteristic (ROC) Curve:

It is a visual representation of the ability of the model to discriminate between positive and negative classes. It plots the recall metric against FPR.

Area Under the ROC Curve (AUC):

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}\left(t\right) dt \tag{A.23}$$

AUC quantifies the overall performance of a model by calculating the area under the ROC curve. A higher AUC indicates better discrimination between classes.

Intersection over Union (IoU):

$$\mathrm{IoU} = \frac{A \cap B}{A \cup B} \tag{A.24}$$

IoU measures the overlap between two regions (A and B) by dividing the area of their intersection by the area of their union.

Generalized Intersection over Union (GIoU):

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{C\mathrm{n}(A \cup B)}{C} \tag{A.25}$$

GIoU is an extension of IoU that considers the difference between the union and the area outside both regions, providing a more accurate measure of overlap.

Dice Score (Dice):

$$\mathrm{Dice} = \frac{2{\cdot}|A \cap B|}{|A| + |B|} \tag{A.26}$$

The Dice Score quantifies the similarity between two sets (A and B) by calculating the ratio of the intersection to the sum of their sizes. It's commonly used in image segmentation tasks.

Analysis of Variance (ANOVA):

N-way ANOVA is a statistical technique that helps determine whether there are statistically significant differences in the means of groups based on the combinations of multiple categorical factors(also known as "ways"). The formula for N-way ANOVA depends on the number of factors and their levels. For a two-way ANOVA, we would have, for example:

$$Y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij} \tag{A.27}$$

Where:

- Y_{ij} as quantification of the statistical relationship.
- μ is the overall mean.
- α_i represents the effect of i.
- β_j represents the effect of j.
- $(\alpha\beta)_{ij}$ represents the interaction between i and j.

Jackknife Free Response Receiver Operating Characteristic(JAFROC) Analysis:

JAFROC extends the concepts of ROC analysis to better suit radiology, enabling radiologists to provide multiple predictions for each patient to evaluate their ability to recognize lesions and estimate their accuracy.

Principal Component Analysis(PCA):

PCA is a mathematical transformation technique used to reduce the dimensionality of a certain dataset for data simplification and feature extraction. It consists of obtaining the covariance matrix of the dataset in question and computing its eigenvectors as principal components.

Multiple Instance Learning(MIL):

Most supervised learning techniques involve assigning a certain label to a data element. Instead, MIL performs set-level classification by grouping data elements in bags labeled for a single class. In MIL, each bag can contain multiple instances or data elements, and the label assigned to the bag reflects whether the bag belongs to a particular class.

Multi-Layer Perceptron(MLP):

MLP is a type of neural network that uses multiple layers of perceptrons(basic units of neural networks) to approximate a function capable of

corresponding input data to the desired output dataset. Each perceptron in the network processes information and makes decisions. When multiple layers of these interconnected perceptrons are used, the network can learn and represent complex relationships within the data.

Stochastic Gradient Descent(SGD):

SGD is an optimization algorithm that enables neural networks to change their weights based on the gradient of their loss function to reduce the loss function score by approximating a function that better fits the relationship between inputs and outputs. It does this by randomly selecting a small subset (mini-batch) of the training data to compute the gradient, introducing a stochastic (random) element that helps escape local minima and converge to a better solution.

Logistic Regression: Logistic regression is a technique that fits relationships between input variables and discrete outcomes. The logistic regression model is trained using optimization algorithms such as gradient descent to find the best set of coefficients that maximize the likelihood of the observed data. Once trained, the model can make predictions by estimating the probability of an instance belonging to each class and assigning it to the class with the highest probability.

References

[1] M. Henderson, Radiology facing a global shortage specialty affected by covid-19, aging population and demand for imaging, 2022. URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

[2] G. Shih, C. wu, S. Halabi, M. Kohli, L. Prevedello, T. Cook, A. Sharma, J. Amorosa, V. Arteaga, M. Galperin-Aizenberg, R. Gill, M. Godoy, S. Hobbs, J. Jeudy, A. Laroia, P. Shah, D. Vummidi, K. Yaddanapudi, A. Stein, Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia, Radiology: Artificial Intelligence 1 (2019) e180041.

[3] W.H. Organisation, Ageing and health (2022). URL: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

[4] A. Khurana, B. Patel, R. Sharpe, Geographic variations in growth of radiologists and medicare enrollees from 2012 to 2019, J. Am. College Radiol. 19 (2022) 1006–1014.

[5] M.A. Azam, K.B. Khan, S. Salahuddin, E. Rehman, S.A. Khan, M.A. Khan, S. Kadry, A.H. Gandomi, A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics, Comput. Biol. Med. 144 (2022) 105253.

[6] C. Hsieh, I.B. Nobre, S.C. Sousa, C. Ouyang, M. Brereton, J.C. Nascimento, J. Jorge, C. Moreira, Mdf-net for abnormality detection by fusing x-rays with clinical data, Scientific Reports 13 (2023) 15873.

[7] Y. LeCun, A path towards autonomous machine intelligence, Open Review 62 (2022).

[8] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38.

[9] Z.C. Lipton, The mythos of model interpretability, Communications ACM 61 (2018) 36–43.

[10] J. Egger, C. Gsaxner, A. Pepe, K.L. Pomykala, F. Jonske, M. Kurz, J. Li, J. Kleesiek, Medical deep learning—a systematic meta-review, Comput. Methods Programs Biomed. 221 (2022) 106874.

[11] J.W. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, Proc. Nat. Acad. Sci. 116 (2019) 22071–22080.

[12] S. El Kafhali, L. Alzubaidi, A. Al-Sabaawi, J. Bai, A. Dukhan, A.H. Alkenani, A. Al-Asadi, H.A. Alwzwazy, M. Manoufali, M.A. Fadhel, A.S. Albahri, C. Moreira, C. Ouyang, J. Zhang, J. Santamaría, A. Salhi, F. Hollman, A. Gupta, Y. Duan, T. Rabczuk, A. Abbosh, Y. Gu, Towards risk-free trustworthy artificial intelligence: Significance and requirements, Int. J. Intell. Syst. 2023 (2023) 4459198.

[13] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications, Inform. Fusion 81 (2022) 59–83.

[14] C. Hsieh, C. Moreira, C. Ouyang, Dice4el: interpreting process predictions using a milestone-aware counterfactual approach, in: in: 2021 3rd International Conference on Process Mining (ICPM), IEEE, 2021, pp. 88–95.

[15] J. Del Ser, A. Barredo-Arrieta, N. Díaz-Rodríguez, F. Herrera, A. Saranti, A. Holzinger, On generating trustworthy counterfactual explanations, Inf. Sci. 655 (2024) 119898.

[16] B. Shneiderman, Human-Centered AI, Oxford University Press, 2022.

[17] A. Watanabe, S. Ketabi, Khashayar, Namdar, F. Khalvati, Improving disease classification performance and explainability of deep learning models in radiology with heatmap generators, arxiv (2022).

[18] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S.Q. Truong, C.D. Nguyen, V.-D. Ngo, J. Seekins, F.G. Blankenberg, A.Y. Ng, et al., Benchmarking saliency methods for chest x-ray interpretation, Nature, Machine Intelligence 4 (2022) 867–878.

[19] A. Duchowski, Eye Tracking Methodology - Theory and Practice, Springer Link, 2017.

[20] A. van der Gijp, C.J. Ravesloot, H. Jarodzka, M. van der Schaaf, I. van der Schaaf, J.P.J. van Schaik, T.J. ten Cate, How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology, Adv. Health Sci. Educ. 22 (2016) 765–787.

[21] T. Brunyé, T. Drew, D. Weaver, J. Elmore, A review of eye tracking for understanding and improving diagnostic interpretation, Cognitive Research: Principles and Implications 4 (2019).

[22] Z. Gandomkar, C. Mello-Thoms, Visual search in breast imaging: A review, The British Journal of Radiology 92 (2019) 20190057.

[23] L. Lévêque, H. Bosmans, L. Cockmartin, H. Liu, State of the art: Eye-tracking studies in medical imaging, IEEE Access PP (2018) 1–1.

[24] E. Arthur, Z. Sun, The application of eye-tracking technology in the assessment of radiology practices: A systematic review, Applied Sciences 12 (2022) 8267.

[25] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, P. Group, et al., Preferred reporting items for systematic reviews and meta-analyses: the prisma statement, International journal of surgery 8 (2010) 336–341.

[26] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 25, Curran Associates Inc, 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[27] K. Saab, S.M. Hooper, N.S. Sohoni, J. Parmar, B. Pogatchnik, S. Wu, J. A. Dunnmon, H.R. Zhang, D. Rubin, C. Ré, Observational supervision for medical image classification using gaze data, in: M. de Bruijne, P.C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham, 2021, pp. 603–614.

[28] Z.L. Jesse Kim, Helen Zhou, Do you see what i see? a comparison of radiologist eye gaze to computer vision saliency maps for chest x-ray classification, arxiv (2022).

[29] Y. Huang, X. Li, L. Yang, L. Gu, Y. Zhu, H. Seo, Q. Meng, T. Harada, Y. Sato, Leveraging human selective attention for medical image analysis with limited training data, arxiv (2021).

[30] M. Kholiavchenko, I. Pershin, B. Maksudov, T. Mustafaev, Y. Yuan, B. Ibragimov, Gaze-based attention to improve the classification of lung diseases, in: O. Colliot, I. Ivsgum (Eds.), Medical Imaging 2022: Image Processing, volume 12032, International Society for Optics and Photonics, SPIE, 2022, p. 120320C, https://doi.org/10.1117/12.2612767.

[31] P. Agnihotri, S. Ketabi, Khashayar, Namdar, F. Khalvati, Using multi-modal data for improving generalizability and explainability of disease classification in radiology, arxiv (2022).

[32] A. Karargyris, S. Kashyap, I. Lourentzou, J. Wu, A. Sharma, M. Tong, S. Abedin, D. Beymer, V. Mukherjee, E.A. Krupinski, M. Moradi, Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development, arxiv (2020).

[33] M. Bhattacharya, S. Jain, P. Prasanna, Radiotransformer: A cascaded global-focal transformer for visual attention-guided disease classification, arxiv (2022).

[34] B. Franceschiello, T.D. Noto, A. Bourgeois, M.M. Murray, A. Minier, P. Pouget, J. Richiardi, P. Bartolomeo, F. Anselmi, Machine learning algorithms on eye tracking trajectories to classify patients with spatial neglect, Comput. Methods Programs Biomed. 221 (2022) 106929.

[35] C. Ma, L. Zhao, Y. Chen, L. Zhang, Z. Xiao, H. Dai, D. Liu, Z. Wu, Z. Liu, S. Wang, J. Gao, C. Li, X. Jiang, T. Zhang, Q. Wang, D. Shen, D. Zhu, T. Liu, Eye-gaze-guided vision transformer for rectifying shortcut learning, arxiv (2022a).

[36] C. Ma, L. Zhao, Y. Chen, D.W. Liu, X. Jiang, T. Zhang, X. Hu, D. Shen, D. Zhu, T. Liu, Rectify vit shortcut learning by visual saliency, arxiv (2022b).

[37] Y. Rong, W. Xu, Z. Akata, E. Kasneci, Human attention in fine-grained classification, arxiv (2021).

[38] S. Wang, X. Ouyang, T. Liu, Q. Wang, D. Shen, Follow my eye: Using gaze to supervise computer-aided diagnosis, IEEE Trans. Med. Imaging 41 (2022) 1688–1698.

[39] H. Zhu, S. Salcudean, R. Rohling, Gaze-guided class activation mapping: Leveraging human attention for network attention in chest x-rays classification, arxiv (2022).

[40] T. van Sonsbeek, X. Zhen, D. Mahapatra, M. Worring, Probabilistic integration of object level annotations in chest x-ray classification, arxiv (2022).

[41] M. Bhattacharya, S. Jain, P. Prasanna, Gazeradar: A gaze and radiomics-guided disease localization framework, in: L. Wang, Q. Dou, P.T. Fletcher, S. Speidel, S. Li (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2022, Springer Nature Switzerland, Cham, 2022, pp. 686–696.

[42] R.B. Lanfredi, J.D. Schroeder, T. Tasdizen, Localization supervision of chest x-ray classifiers using label-specific eye-tracking annotation, arxiv (2022).

[43] N. Khosravan, H. Celik, B. Turkbey, E.C. Jones, B. Wood, U. Bagci, A collaborative computer aided diagnosis (c-cad) system with eye-tracking, sparse attentional model, and deep learning, Med. Image Anal. 51 (2019) 101–115.

[44] M. Wedel, J. Yan, E. Siegel, A. Li, Nodule detection with eye movements, Journal of Behavioral Decision Making 29 (2016) n/a–n/a.

[45] A. Luís, C. Hsieh, I.B. Nobre, S.C. Sousa, A. Maciel, C. Moreira, J. Jorge, Integrating eye-gaze data into cxr dl approaches: A preliminary study, arxiv (2023).

[46] J. Stember, H. Celik, E. Krupinski, P. Chang, S. Mutasa, B. Wood, A. Lignelli, G. Moonis, L. Schwartz, S. Jambawalikar, Eye tracking for deep learning segmentation using convolutional neural networks, J. Digit. Imaging 32 (2019).

[47] J. Stember, H. Celik, D. Gutman, N. Swinburne, R. Young, S. Eskreis-Winkler, A. Holodny, S. Jambawalikar, B. Wood, P. Chang, E. Krupinski, Integrating eye-tracking and speech recognition accurately annotates mri brain images for deep learning: Proof of principle, Radiology, Artif. Intell. 3 (2020) e200047.

[48] N. Castner, T. Kübler, K. Scheiter, J. Richter, T. Eder, F. Hüttig, C. Keutel, E. Kasneci, Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing, arxiv (2020).

[49] N. Castner, J. Frankemölle, C. Keutel, F. Huettig, E. Kasneci, Lstms can distinguish dental expert saccade behavior with high plaque-urracy, in: in: 2022 Symposium on Eye Tracking Research and Applications, ETRA '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 1–7, https://doi.org/10.1145/3517031.3529631.

[50] T. Donovan, D. Litchfield, Looking for cancer: Expertise related differences in searching and decision making, Applied Cognitive Psychology 27 (2013) 43–49.

[51] L. McLaughlin, R. Bond, C. Hughes, J. McConnell, S. McFadden, Computing eye gaze metrics for the automatic assessment of radiographer performance during x-ray image interpretation, Int. J. Med. Informatics 105 (2017) 11–21.

[52] N. Castner, S. Klepper, L. Kopnarski, F. Huettig, K. Scheiter, J. Richter, E. Kasneci, C. Keutel, Overlooking: The nature of gaze behavior and anomaly detection in expert dentists, in: in: Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, 2018, pp. 1–6, https://doi.org/10.1145/3279810.3279845.

[53] S. Mall, P. Brennan, C. Mello-Thoms, Modeling visual search behavior of breast radiologists using a deep convolution neural network, Journal of Medical Imaging 5 (2018) 1.

[54] S. Mall, P. Brennan, C. Mello-Thoms, Can a machine learn from radiologists' visual search behaviour and their interpretation of mammograms—a deep-learning study, J. Digit. Imaging 32 (2019).

[55] S. Mall, E. Krupinski, C. Mello-Thoms, Missed cancer and visual search of mammograms: what feature-based machine-learning can tell us that deep-convolution learning cannot, in: R.M. Nishikawa, F.W. Samuelson (Eds.), Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment, volume 10952, International Society for Optics and Photonics, SPIE, 2019, p. 1095216, https://doi.org/10.1117/12.2512539.

[56] G. Tourassi, S. Voisin, V. Paquit, E. Krupinski, Investigating the link between radiologists' gaze, diagnostic decision, and image content, J. Am. Med. Inform. Assoc. 20 (2013) 1067–1075.

[57] M.W. Pietrzyk, D. Rannou, P.C. Brennan, Implementation of combined SVM-algorithm and computer-aided perception feedback for pulmonary nodule detection, in: C.K. Abbey, C.R. Mello-Thoms (Eds.), Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment, volume 8318, International Society for Optics and Photonics, SPIE, 2012, p. 831815, https://doi.org/10.1117/12.911577.

[58] I. Pershin, M. Kholiavchenko, B. Maksudov, T. Mustafaev, B. Ibragimov, AI-based analysis of radiologist's eye movements for fatigue estimation: a pilot study on chest X-rays, in: C.R. Mello-Thoms, S. Taylor-Phillips (Eds.), Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment, volume 12035, International Society for Optics and Photonics, SPIE, 2022, p. 120350Y, https://doi.org/10.1117/12.2612760.

[59] I. Pershin, M. Kholiavchenko, B. Maksudov, T. Mustafaev, D. Ibragimova, B. Ibragimov, Artificial intelligence for the analysis of workload-related changes in radiologists' gaze patterns, IEEE Journal of Biomedical and Health Informatics PP (2022b) 1–10.

[60] I. Pershin, T. Mustafaev, D. Ibragimova, B. Ibragimov, Changes in radiologists' gaze patterns against lung x-rays with different abnormalities: a randomized experiment, J. Digit. Imaging 36 (2023).

[61] K. Dmitriev, J. Marino, K. Baker, A.E. Kaufman, Visual analytics of a computer-aided diagnosis system for pancreatic lesions, IEEE Trans. Visual Comput. Graphics 27 (2021) 2174–2185.

[62] R.B. Lanfredi, A. Arora, T. Drew, J.D. Schroeder, T. Tasdizen, Comparing radiologists' gaze and saliency maps generated by interpretability methods for chest x-rays, arxiv (2023).

[63] M. Watson, B.A.S. Hasan, N. Al Moubayed, Learning how to mimic: Using model explanations to guide deep learning training, in: in: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1461–1470, https://doi.org/10.1109/WACV56688.2023.00151.

[64] S. Mall, P. Brennan, C. Mello-Thoms, Fixated and not fixated regions of mammograms: A higher-order statistical analysis of visual search behavior, Academic Radiology 24 (2017) 442–455.

[65] J. Lou, H. Lin, D. Marshall, R. White, Y. Yang, S. Shelmerdine, H. Liu, Predicting radiologist attention during mammogram reading with deep and shallow high-resolution encoding, IEEE International Conference on Image Processing (ICIP) 2022 (2022) 961–965, https://doi.org/10.1109/ICIP46576.2022.9897723.

[66] C. Moreira, D.M. Alvito, S.C. Sousa, I.M.G.B. Nobre, C. Ouyang, R. Kopper, A. Duchowski, J. Jorge, Comparing visual search patterns in chest x-ray diagnostics, in: Proceedings of the 2023 Symposium on Eye Tracking Research and Applications, ETRA '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 1–6. URL: doi: 10.1145/3588015.3588403. doi:10.1145/3588015.3588403.

[67] R.B. Lanfredi, M. Zhang, W.F. Auffermann, J. Chan, P.-A.T. Duong, V. Srikumar, T. Drew, J.D. Schroeder, T. Tasdizen, Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays, Scientific Data 9 (2022).

[68] G. Aresta, C. Ferreira, J. Pedrosa, T. Araujo, J. Rebelo, E. Negrao, M. Morgado, F. Alves, A. Cunha, I. Ramos, A. Campilho, Automatic lung nodule detection combined with gaze information improves radiologists' screening performance, IEEE Journal of Biomedical and Health Informatics (2020).

[69] M. Bhattacharya, S. Jain, P. Prasanna, Training focal lung pathology detection using an eye movement modeling example, arxiv (2021).

[70] T. Drew, C. Cunningham, J. Wolfe, When and why might a computer-aided detection (cad) system interfere with visual search? an eye-tracking study, Academic radiology 19 (2012) 1260–1267.

[71] A. van der Gijp, C.J. Ravesloot, H. Jarodzka, M.F. van der Schaaf, I.C. van der Schaaf, J.P.J. van Schaik, T.J. ten Cate, How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology, Adv. Health Sci. Educ. (2016).

[72] P. Lanzer, M. Al-Naser, S. Bukhari, A. Dengel, E. Krupinski, Eye tracking in catheter-based cardiovascular interventions: Early results, Journal of Medical Imaging 4 (2017) 035502.

[73] C. Moreira, I.B. Nobre, S.C. Sousa, J.M. Pereira, J. Jorge, Improving x-ray diagnostics through eye-tracking and xr, in, in: IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, 2022, pp. 450–453.

[74] K. Panetta, R. Rajendran, A. Ramesh, S. Rao, Tufts dental database: A multimodal panoramic x-ray dataset for benchmarking diagnostic systems, IEEE Journal of Biomedical and Health Informatics PP (2021) 1–1.

[75] A. Zawacki, C. Wu, G. Shih, J. Elliott, M. Fomitchev, M.H. ParasLakhani, P. Culliton, S. Bao, Siim-acr pneumothorax segmentation (2019). URL: https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation.

[76] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. arXiv:1512.03385.

[77] A.E.W. Johnson, T.J. Pollard, N.R. Greenbaum, M.P. Lungren, C. ying Deng, Y. Peng, Z. Lu, R.G. Mark, S.J. Berkowitz, S. Horng, Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019. arXiv:1901.07042.

[78] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks (2018) arXiv:1608.06993.

[79] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, 2015. arXiv:1505.04597.

[80] M. Tan, Q.V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. arXiv:1905.11946.

[81] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. Mong, S. Halabi, J. Sandberg, R. Jones, D. Larson, C. Langlotz, B. Patel, M. Lungren, A. Ng, Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 590–597.

[82] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization, CoRR abs/1610.02391 (2016).

[83] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, 2015. arXiv:1505.04366.

[84] J.T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, 2015. arXiv:1412.6806.

[85] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M.K. Prasadha, J. Pei, M.Y. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V.A. Huu, C. Wen, E.D. Zhang, C.L. Zhang, O. Li, X. Wang, M.A. Singer, X. Sun, J. Xu, A. Tafreshi, M.A. Lewis, H. Xia, K. Zhang, Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (2018) 1122–1131.e9.

[86] G. Shih, C. wu, S. Halabi, M. Kohli, L. Prevedello, T. Cook, A. Sharma, J. Amorosa, V. Arteaga, M. Galperin-Aizenberg, R. Gill, M. Godoy, S. Hobbs, J. Jeudy, A. Laroia, P. Shah, D. Vummidi, K. Yaddanapudi, A. Stein, Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia, Radiology: Artificial Intelligence 1 (2019) e180041.

[87] H.Q. Nguyen, K. Lam, L.T. Le, H.H. Pham, D.Q. Tran, D.B. Nguyen, D.D. Le, C.M. Pham, H.T.T. Tong, D.H. Dinh, C.D. Do, L.T. Doan, C.N. Nguyen, B.T. Nguyen, Q. V. Nguyen, A.D. Hoang, H.N. Phan, A.T. Nguyen, P.H. Ho, D.T. Ngo, N.T. Nguyen, N.T. Nguyen, M. Dao, V. Vu, Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2022. arXiv:2012.15029.

[88] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows (2021) arXiv:2103.14030.

[89] I. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. Cardoso, J. Cardoso, Inbreast: Toward a full-field digital mammographic database, Academic radiology 19 (2011) 236–248.

[90] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021. arXiv:2010.11929.

[91] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, A. Oliva, Intrinsic and extrinsic effects on image memorability, Vision research 116 (2015) 165–178.

[92] A. Borji, L. Itti, Cat2000: A large scale fixation dataset for boosting saliency research, 2015. arXiv:1505.03581.

[93] S. Jia, N.D.B. Bruce, Eml-net:an expandable multi-layer network for saliency prediction, 2019. arXiv:1805.01047.

[94] M. Nevitt, D. Felson, G. Lester, The osteoarthritis initiative, Protocol for the Cohort Study 1 (2006).

[95] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, 2015 arXiv:1512.04150.
[96] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 3462–3471. URL: https://doi.org/10.1109 doi:10.1109/cvpr.2017.369.
[97] D.P. Kingma, M. Welling, Auto-encoding variational bayes (2022) arXiv:1312.6114.
[98] M. Crawshaw, Multi-task learning with deep neural networks, A survey (2020) arXiv:2009.09796.
[99] H. Sheridan, E. Reingold, The holistic processing account of visual expertise in medical image perception: A review, Frontiers in Psychology 8 (2017) 1620.
[100] T. Donovan, D. Litchfield, Looking for cancer: Expertise related differences in searching and decision-making, Applied Cognitive Psychology 27 (2013) 43–49.
[101] Z. Qi, S. Khorram, F. Li, Visualizing deep networks by optimizing with integrated gradients, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019.
[102] A. Kemp, A. Zawacki, C. Carr, G. Shih, J. Mongan, J. Elliott, K. ParasLakhani, P. Culliton, Siim-fisabio-rsna covid-19 detection (2021). URL: https://kaggle.com/competitions/siim-covid19-detection.
[103] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, Curran Associates Inc., Red Hook, NY, USA, 2012, p. 1097–1105.
[104] G. Aresta, C. Ferreira, J. Pedrosa, T. Araújo, J. Rebelo, E. Negrão, M. Morgado, F. Alves, A. Cunha, I. Ramos, A. Campilho, Automatic lung nodule detection combined with gaze information improves radiologists' screening performance, IEEE Journal of Biomedical and Health Informatics 24 (2020) 2894–2901.
[105] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018. arXiv:1804.02767.
[106] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, 2018. arXiv:1703.06870.
[107] D.A. Spielman, N. Srivastava, Graph sparsification by effective resistances, CoRR abs/0803.0929 (2008).
[108] T. Zhang, R. Ramakrishnan, M. Livny, Birch: An efficient data clustering method for very large databases, SIGMOD Rec. 25 (1996) 103–114.
[109] A. Holzinger, B. Haibe-Kains, I. Jurisica, Why imaging data alone is not enough: Ai-based integration of imaging, omics, and clinical data, European Journal of Nuclear Medicine and Molecular Imaging 46 (2019).
[110] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.
[111] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, 2016. arXiv:1602.07261.
[112] M. Jiang, S. Huang, J. Duan, Q. Zhao, Salicon: Saliency in context. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1072–1080, https://doi.org/10.1109/CVPR.2015.7298710.
[113] L. Lévêque, P. Young, T. Wales, H. Liu, Studying the gaze patterns of expert radiologists in screening mammography: A case study with breast test wales, in: in: Proceedings of the 28th European Signal Processing Conference, 2021, pp. 1249–1253, https://doi.org/10.23919/Eusipco47968.2020.9287678.
[114] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, Deep high-resolution representation learning for visual recognition, 2020. arXiv:1908.07919.
[115] B. Zoph, V. Vasudevan, J. Shlens, Q.V. Le, Learning transferable architectures for scalable image recognition, 2018. arXiv:1707.07012.
[116] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015. arXiv:1409.1556.
[117] V.N. Vapnik, The nature of statistical learning theory, Springer-Verlag, New York Inc, 1995.
[118] J. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 29 (2000).
[119] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1997) 119–139.
[120] L. Breiman, Bagging predictors, Machine Learning 24 (2004) 123–140.
[121] S. Andersen, Judea pearl, probabilistic reasoning in intelligent systems: Networks of plausible inference, Artif. Intell. 48 (1991) 117–124.
[122] J. Rennie, L. Shih, J. Teevan, D. Karger, Tackling the poor assumptions of naive bayes text classifiers, in: Proceedings of the Twentieth International Conference on Machine Learning 41, 2003.
[123] A. Shahid, K. Wilkinson, S. Marcu, C.M. Shapiro, Stanford Sleepiness Scale (SSS), Springer, New York, New York, NY, 2012, pp. 369–370, https://doi.org/10.1007/978-1-4419-9893-4_91.
[124] D.R. McLeod, R.R. Griffiths, G.E. Bigelow, J. Yingling, An automated version of the digit symbol substitution test (dsst), Behavior Research Methods & Instrumentation 14 (1982) 463–466.
[125] A.R. Jensen, Clocking the Mind: Mental Chronometry and Individual Differences, Elsevier, 2006.
[126] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[127] Y. Ling, Bio+clinical bert, bert base, and cnn performance comparison for predicting drug-review satisfaction, 2023. arXiv:2308.03782.
[128] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: Training a large language-and-vision assistant for biomedicine in one day, arxiv (2023).
[129] Y. Chen, Q. Fu, Y. Yuan, Z. Wen, G. Fan, D. Liu, D. Zhang, Z. Li, Y. Xiao, Hallucination detection: Robustly discerning reliable answers in large language models, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 245–255, https://doi.org/10.1145/3583780.3614905.
[130] C. Moreira, Y.-L. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, P. Bruza, Linda-bn: An interpretable probabilistic approach for demystifying black-box predictive models, Decis. Support Syst. 150 (2021) 113561.
[131] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (2019) e1312.
[132] M.N. Hoque, K. Mueller, Outcome-explorer: A causality guided interactive visual interface for interpretable algorithmic decision making, IEEE Trans. Visual Comput. Graphics 28 (2021) 4728–4740.
[133] L. Alzubaidi, A. Al-Sabaawi, J. Bai, A. Dukhan, A.H. Alkenani, A. Al-Asadi, H. A. Alwzwazy, M. Manoufali, M.A. Fadhel, A. Albahri, et al., Towards risk-free trustworthy artificial intelligence: Significance and requirements, Int. J. Intell. Syst. 2023 (2023).
[134] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018).
[135] B. Wickramanayake, C. Ouyang, C. Moreira, Y. Xu, Generating purpose-driven explanations: The case of process predictive model inspection, in: in: International Conference on Advanced Information Systems Engineering, Springer, 2022, pp. 120–129.
[136] L. Sun, C. Yin, Q. Xu, W. Zhao, Artificial intelligence for healthcare and medical education: a systematic review, Am. J. Transl. Res. 15 (2023) 4820–4828.
[137] Z. Wang, L. Liu, L. Wang, L. Zhou, R2gengpt: Radiology report generation with frozen llms, 2023. arXiv:2309.09812.