Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Doctoral Consortium Track

# Human-Centred Multimodal Deep Learning Models for Chest X-Ray Diagnosis

**Chihcheng Hsieh**

School of Information Systems, Queensland University of Technology, Brisbane, Australia
Centre for Data Science, Queensland University of Technology, Brisbane, Australia
chihcheng.hsieh@hdr.qut.edu.au

## Abstract

My thesis consists of investigating how chest X-ray images, radiologists' eye movements and patients' clinical data can be used to teach a machine how radiologists read and classify images with the goal of creating human-centric AI architectures that can (1) capture radiologists' search behavioural patterns using their eye-movements in order to improve classification in DL systems, and (2) automatically detect lesions in medical images using clinical data and eye tracking data. Heterogeneous data sources such as chest X-rays, radiologists' eye movements, and patients' clinical data can contribute to novel multimodal DL architectures that, instead of learning directly from images' pixels, will learn human classification patterns encoded in both the eye movements of the images' regions and patients' medical history. In addition to a quantitative evaluation, I plan to conduct questionnaires with expert radiologists to understand the effectiveness of the proposed multimodal DL architecture.

## 1 Research Problem

Deep Learning (DL) technologies are widely used in medical imaging to automatically diagnose and detect subtle differences that may be difficult for human radiologists to detect. However, most DL applications in medical imaging only use image data due to limited literature on medical datasets combining different data modalities. While DL models can outperform humans in diagnosing chest X-rays (CXRs), many times these models get the *predictions correct, but not for the right clinical reasons* [Saporta *et al.*, 2022]. Current state-of-the-art DL models often neglect the multimodal nature of radiologists' perception when examining medical images. This is a crucial limitation since the combination of visual cues, such as shapes, textures, and objects, allows radiologists to form mental models that generalize better with less data. Hence, there is a need to integrate the human component in DL models to improve the accuracy and robustness of medical image diagnosis.

To better understand how humans operate in the radiology reading room, I interviewed a team of expert radiologists about their approach to assessing X-ray images. They emphasized the importance of having access to clinical information about the patient to make an accurate diagnosis. Without this information, they noted that their diagnoses can be more prone to biases and errors. These findings highlight the importance of integrating clinical data in medical image analysis. However, there is a lack of consideration in existing DL architectures in using clinical data for prediction in CXRs. Some recent studies also confirm using accurate patients' *clinical data* together with medical images can improve the performance of human radiologists in terms of accuracy [Castillo *et al.*, 2021].

Recently, two novel datasets (REFLACX [Lanfredi *et al.*, 2022] and EyeGaze [Karargyris *et al.*, 2021]) have been introduced that combine CXRs with radiologists' eye-gaze movements. While previous studies suggest that integrating radiologists' eye-gaze data can aid in detecting potential lesions in CXRs [Ganesan *et al.*, 2018], there remains a lack of research exploring the incorporation of this modality in DL models for medical image analysis. Hence, further research is needed to investigate and leverage the potential of radiologists' eye-gaze data in DL to enhance diagnostic accuracy and mitigate errors.

The aim of this thesis is to address the aforementioned research gaps by investigating various types of multimodal DL architectures and assessing their effectiveness in conjunction with expert radiologists. Specifically, this research endeavors to explore the integration of different data modalities such as patients' clinical data and eye tracking data, to develop robust DL models for medical image analysis.

## 2 Current Research Contributions

So far, I have interviewed expert radiologists to understand how they operate in their radiology reading room, and how they assess CXRs. I also contributed to a public dataset that integrates several different datasets of the literature into a single multimodal dataset that contains clinical data, CXRs, radiologists' audio transcripts, and their eye movements. Using this dataset and information provided by radiologists, I proposed a multimodal DL model that combines clinical data with CXRs.

### 2.1 The MIMIC-EYE Multimodal Medical Dataset

In this thesis, I created MIMIC-EYE [Hsieh *et al.*, 2023], a comprehensive dataset that integrates several MIMIC datasets, including medical images, reports, clinical data, and eye tracking data with gaze and pupil dilation information. The integration of eye tracking data with MIMIC modalities may enhance

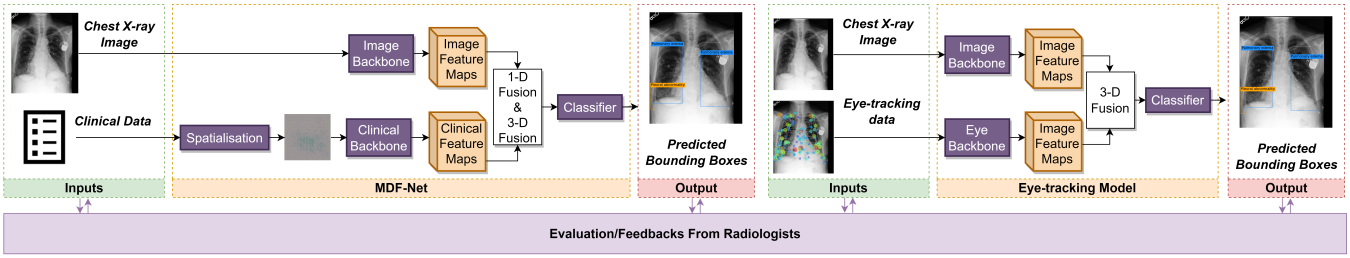


Figure 1: Model on the left is MDF-Net [Hsieh *et al.*, 2023], which involves clinical data into the lesion detection task. The clinical data is passed through the spatialisation module, which increases the dimensionality of the clinical data from 1 to 3. Then, the image and clinical data can be fused in the same dimension. The model on the right is the proposed architecture for incorporating eye-tracking data [Luís *et al.*, 2023].

the understanding of radiologists' visual search patterns and improve DL model development. MIMIC-Eye contains 3,689 tuples of chest X-ray images, eye-gaze data, and radiologists' audio transcripts, of which 1,683 tuples include clinical data.

### 2.2 MDF-Net: Multimodal Dual-Fusion Network

After creating the MIMIC-Eye dataset, we developed a new architecture called multimodal Dual-Fusion Network (MDF-Net), which uses both CXR images and clinical data for lesion detection. The challenge in using these modalities is that they have different dimensions, with images typically represented as 3-dimensional tensors and clinical data as 1-dimensional tensors. To address this, we introduced a strategy called spatialisation, which elevates 1-D clinical data into 3-D space (Figure 1). Our spatialisation module contains multiple deconvolutional layers that upsample the pixel to the desired size for fusion, allowing us to fuse images and clinical data in 3-D space. Through MDF-Net, we demonstrated that incorporating clinical data improved the generalization of the model and lesion detection performance by +12% on Average Precision (AP). Our novel architecture and spatialisation strategy provide a promising approach for using multimodal data in deep learning for medical image analysis (Table 1).

| | Mask R-CNN (Baseline) | | MDF-Net (3D Fusion only) | | MDF-Net | |
|---|---|---|---|---|---|---|
| **Lesion** | **AP** | **AR** | **AP** | **AR** | **AP** | **AR** |
| **Enl. Card. Sil.** | 54.83 | **100.00** | 69.12 | 94.44 | **70.36** | **100.00** |
| **Atalectasis** | 14.58 | 40.00 | 16.33 | 42.86 | **24.43** | 48.57 |
| **Pleural Abn.** | **16.29** | 42.86 | 13.20 | **52.38** | 16.09 | 28.57 |
| **Consolidation** | 3.12 | 20.00 | **16.93** | **40.00** | 14.29 | 30.00 |
| **Pulm. Edema** | 9.22 | 66.67 | 22.26 | **72.22** | **33.25** | 66.67 |
| **Overall** | 19.61 | 53.90 | 27.57 | **60.38** | **31.69** | 54.76 |

Table 1: Evaluation results of the proposed MDF-Net. Results show that the dual fusion mechanism outperforms the baseline Mask-RCNN by $\approx 12\%$ AP, and the MDF-Net(3D) by 4.12% for a score threshold = 0.05 / IoBB threshold = 0.5. In the table, Enl. Card. Sil. corresponds to Enlarged Cardiac Silhouette, Pleural Abn to Pleural Abnormality, and Pulm. Edema to Pulmonary Edema.

## 3 Future Research Directions

My current research investigates the integration of eye tracking data into DL architectures for lesion detection in CXRs. Initial results showed that directly providing fixation masks of radiologists' gaze patterns as input did not improve the baseline Mask RCNN. Future work will address two research challenges: (1) DL approaches that use raw eye gaze data without considering the intricacies of its human generation process, leading to noisy data that does not effectively contribute to the supervised learning process, and (2) the human-centric nature of eye tracking data, requiring the consideration of human factors in its generation, including the investigation of how radiologists examine CXRs and their visual search patterns.

## Acknowledgements

My PhD research is sponsored by a full scholarship offered by the Centre for Data Science (CDS) at Queensland University of Technology. The material is based on the work supported by the UNESCO Chair on AI&XR; and the Portuguese *Fundação para a Ciência e a Tecnologia (FCT)* under grants no. 2022.09212.PTDC (XAVIER) and no. UIDB/50021/2020.

## References

[Castillo *et al.*, 2021] C. Chelsea *et al.* The effect of clinical information on radiology reporting: A systematic review. *J. of Medical Radiation Sciences*, 68:60–74, 2021.

[Ganesan *et al.*, 2018] A. Ganesan *et al.* A review of factors influencing radiologists' visual search behaviour. *J. of Medical Imaging & Radiation Oncology*, 62:747–757, 2018.

[Hsieh *et al.*, 2023] C. Hsieh *et al.* Mdf-net: Multimodal dual-fusion network for abnormality detection using cxr images and clinical data. *arXiv:2302.13390*, 2023.

[Hsieh *et al.*, 2023] C. Hsieh *et al.* MIMIC-Eye: Integrating MIMIC Datasets with REFLACX and Eye Gaze for Multimodal Deep Learning Applications. *PhysioNet*, 2023.

[Karargyris *et al.*, 2021] A. Karargyris *et al.* Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for AI development. *Scientific Data*, 8, 2021.

[Lanfredi *et al.*, 2022] R. Lanfredi *et al.* REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific Data*, 9, 2022.

[Luís *et al.*, 2023] A. Luís *et al.* Integrating eye-gaze data into CXR DL approaches: A preliminary study, IEEE Conf. on VR and 3D User Interfaces Abstracts and Workshops, 2023

[Saporta *et al.*, 2022] A. Saporta *et al.* Benchmarking saliency methods for chest X-ray interpretation. *Nature Machine Intelligence*, 4:867—878, 2022.