scientific reports

OPEN

MDF-Net for abnormality detection by fusing X-rays with clinical data

Chihcheng Hsieh[1], Isabel Blanco Nobre[2,7], Sandra Costa Sousa[2,7], Chun Ouyang[1], Margot Brereton[1], Jacinto C. Nascimento[3], Joaquim Jorge[4] & Catarina Moreira[1,4,5,6]✉

This study investigates the effects of including patients' clinical information on the performance of deep learning (DL) classifiers for disease location in chest X-ray images. Although current classifiers achieve high performance using chest X-ray images alone, consultations with practicing radiologists indicate that clinical data is highly informative and essential for interpreting medical images and making proper diagnoses. In this work, we propose a novel architecture consisting of two fusion methods that enable the model to simultaneously process patients' clinical data (structured data) and chest X-rays (image data). Since these data modalities are in different dimensional spaces, we propose a spatial arrangement strategy, *spatialization*, to facilitate the multimodal learning process in a Mask R-CNN model. We performed an extensive experimental evaluation using MIMIC-Eye, a dataset comprising different modalities: MIMIC-CXR (chest X-ray images), MIMIC IV-ED (patients' clinical data), and REFLACX (annotations of disease locations in chest X-rays). Results show that incorporating patients' clinical data in a DL model together with the proposed fusion methods improves the disease localization in chest X-rays by 12% in terms of Average Precision compared to a standard Mask R-CNN using chest X-rays alone. Further ablation studies also emphasize the importance of multimodal DL architectures and the incorporation of patients' clinical data in disease localization. In the interest of fostering scientific reproducibility, the architecture proposed within this investigation has been made publicly accessible(https://github.com/ChihchengHsieh/multimodal-abnormalities-detection).

According to the Lancet[1], 2019 witnessed a shortage of 6.4 million physicians, 30.6 million nurses, and 2.9 million pharmaceutics personnel across 132 countries worldwide, especially in Low and Medium Income Countries. The situation has worsened after the pandemic since medical staff were disproportionately affected.

Deep Learning (DL) technologies promise to benefit health systems, professionals, and the public, making existing clinical and administrative processes more effective, efficient, and equitable. These technologies have become highly popular in the medical imaging field, where a plethora of applications have been successfully addressed, e.g., breast imaging[2,3], left ventricular assessment[4,5], dermoscopy analysis[6,7] and chest X-rays[8–12], which have gained further attention due to the recent pandemic. Despite their advantages, these systems have notable shortcomings: they require extensive amounts of labelled data to operate correctly. Explicitly labelling anomalies in large amounts of medical images requires the availability of medical experts, who are scarce and expensive. The process is time-consuming and costly, resulting in bottlenecks in research advancements[13]. Additionally, the complex interconnected architectures make DL predictions opaque and resistant to scrutiny, hindering the adoption of AI systems in public health (known as the "black box" problem[14–19]). A system that could automatically annotate or highlight relevant regions in medical images in a similar way to humans would be extremely useful and could save millions of dollars creating breakthroughs in research in AI adoption in Healthcare[13].

Several works in the literature attempt to automatically learn regions of interest indicating the patient's clinical abnormalities. These works mainly use DL approaches which have been found to be efficient and effective on a variety of computer vision tasks[20–23]. Mask R-CNN is one of the most widely used DL architectures to predict regions with abnormalities in images[24]. However, their predictive performance is still low, and these architectures do not take into consideration the process of how expert radiologists assess and diagnose these images. The human component is completely disregarded in most DL studies to predict abnormalities in chest X-rays. This is relevant because when radiologists look at an X-ray image, they experience it in a multimodal world: they

[1]Queensland University of Technology, Brisbane, Australia. [2]Lusíadas Knowledge Center, Imageology Department, Lisbon, Portugal. [3]Institute for Systems and Robotics, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal. [4]Instituto Superior Técnico, University of Lisbon, Portugal, Lisbon, Portugal. [5]Human Technology Institute, University of Technology Sydney, Ultimo, Australia. [6]INESC-ID, Lisbon, Portugal. [7]These authors contributed equally: Isabel Blanco Nobre and Sandra Costa Sousa. ✉email: catarina.pintomoreira@uts.edu.au

see objects, textures, shapes, etc. It is the combination of these modalities that make humans capable of making mental models that generalize well with less data when compared to DL approaches.

Advancements in computer-aided diagnostics (CAD) employing radiomics and deep learning have also been explored in the literature. A compelling example can be seen in a study that presents an attention-augmented Wasserstein generative adversarial network (AA-WGAN) for fundus retinal vessel segmentation. The application of attention-augmented convolution and squeeze-excitation modules highlights regions of interest and suppresses extraneous information in the images, proving effective in segmenting intricate vascular structures[25]. Further, an attention-based glioma grading network (AGGN) for MRI data shows superior performance, highlighting the key modalities and locations in the feature maps even without manually labelled tumour masks[26]. Lastly, a CAD model, Cov-Net, exhibits robust feature learning for accurate COVID-19 detection from chest X-ray images, outperforming other computer vision algorithms[27]. Together, these studies underscore the potential of radiomics and deep learning in improved health anomaly detection, segmentation, and grading.

Multimodal DL consists of architectures that can learn, process, and link information from different data modalities (such as text, images, structured data, etc.)[28–34]. Deep learning can benefit from multimodal data in terms of generalization and performance compared to the unimodal paradigm (see for instance Azam et al.[35] for a comprehensive review in medical multimodal images). In terms of multimodal DL approaches for chest X-ray images, most works in the literature focus on combining image data with text to generate reports, predict diseases or even for lesion detection by training BERT-like models[36–39]. However, regarding disease classification, the medical reports associated with the images used for training already contain some information about the patient's diseases, which may generate biased results.

A recent literature review[40] indicated that clinical data is highly informative and essential for radiologists to interpret and make proper diagnoses. To the best of our knowledge, there are no multimodal DL approaches that combine patients' clinical information with X-ray images to predict the location of abnormalities in chest X-rays. This justifies and motivates our research path in the present paper. Concretely, chest X-ray images and clinical data are aimed at addressing a critical gap in current object detection deep learning approaches in terms of fusing tabular data and image data, which are very scarce. Our interviews with radiologists reinforced the importance of clinical data in making accurate diagnoses from chest X-ray images because radiologists cannot accurately assess an X-ray image without knowing the patient's clinical data. By integrating these two types of data into our model, we aim to capture the nuanced decision-making process of radiologists more effectively. This approach allows the model to leverage the patterns visible in the images and the rich contextual information available in the clinical data, leading to a more holistic and accurate disease localization.

In this paper, we propose the *Multimodal Dual-Fusion Network (MDF-Net)*, which is a novel architecture inspired and extended from Mask R-CNN[24]. MDF-Net can fuse chest X-ray images and clinical features simultaneously to detect regions in chest X-rays with abnormalities more accurately. The proposed architecture uses a two-stage detector comprising a Region Proposal Network (RPN) of Mask R-CNN followed by an attention mechanism to extract information only from Regions of Interest (RoIs). Figure 1 presents a general description of the proposed framework. Figure 1a shows the overall architecture of the proposed model. The prediction process can be divided into three phases.

The initial phase is dedicated to extracting semantic representations, known as feature maps, from the input data. This involves the distinct processing of the two data modalities. One model subsystem calculates a feature map utilizing the input images, whereas a different subsystem computes a feature map based on clinical data. In the second phase, we conduct a fusion operation to fuse the two feature maps above to obtain a joint representation. Then, the third phase deploys the final classifier, the purpose of which is to predict the bounding boxes corresponding to any detected abnormalities. In Fig. 1b, we integrated the *triage* data from MIMIC-IV ED with REFLACX, which contains bounding box annotations, to get corresponding clinical data for each CXR image. The integrated (multimodal) dataset, MIMIC-EYE[43], allows us to perform multimodal learning with the model shown in Fig. 1a. In the end, we performed two ablation studies: one that investigated the impact of the different fusion methods in our architecture; and another that investigated the impact of different sets of clinical features for abnormality detection. The performance of the models was measured using Average Precision (AP) and Intersection Over predicted Bounding Boxes (IoBB) (Fig. 1c).

The key contributions of this work are as follows:

(1) We propose a strategy to extract corresponding clinical data for CXR images from the MIMIC[42,44,45] dataset. This strategy is then used to construct our multimodal dataset for the abnormality detection task[43];
(2) We propose a multimodal learning architecture and two fusion methods to fuse tabular and image data;
(3) We demonstrate the effectiveness and importance of clinical data in abnormality detection through ablation studies.

Results

Overall, our experiments show that clinical data plays an important role in abnormality detection. Using clinical data together with chest X-ray images, the proposed MDF-Net architecture achieved a significant improvement in terms of both average precision (AP) (31.69%) and average recall (AR) (54.76%) when compared to the baseline Mask R-CNN model that relies solely on X-ray imaging, which achieved 19.61% AP and 54% AR. This translates into an improvement of 12.08% AP and 0.86% AR. Figure 2(a) shows an example where the proposed MDF-Net was able to correctly predict bilateral pulmonary edema while the Mask R-CNN (baseline) did not identify any abnormality. In terms of the number of false negatives/false positives, the proposed MDF-Net with both fusion methods always generated fewer false positives (FP) and false negatives (FN) for the majority of the

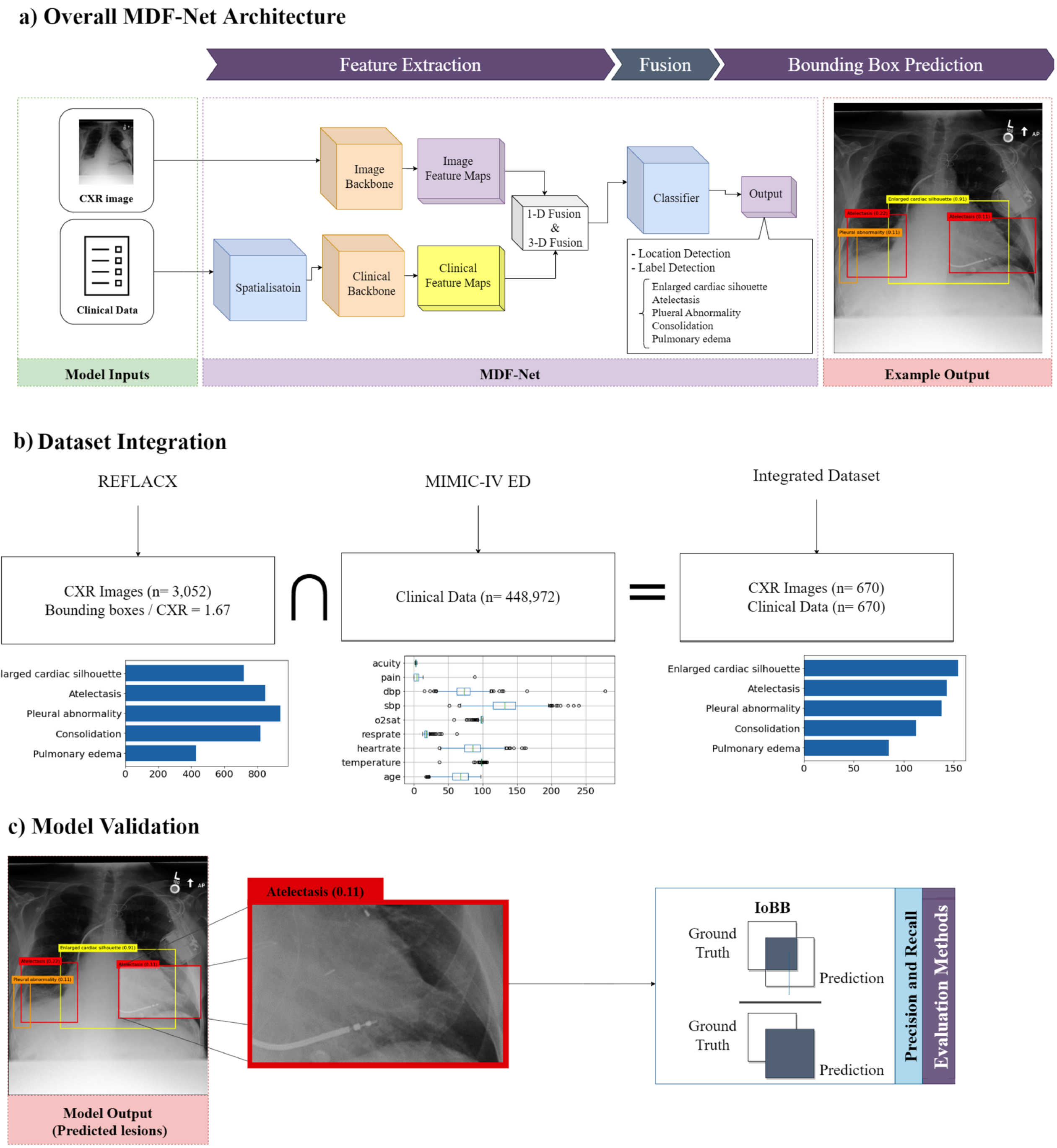
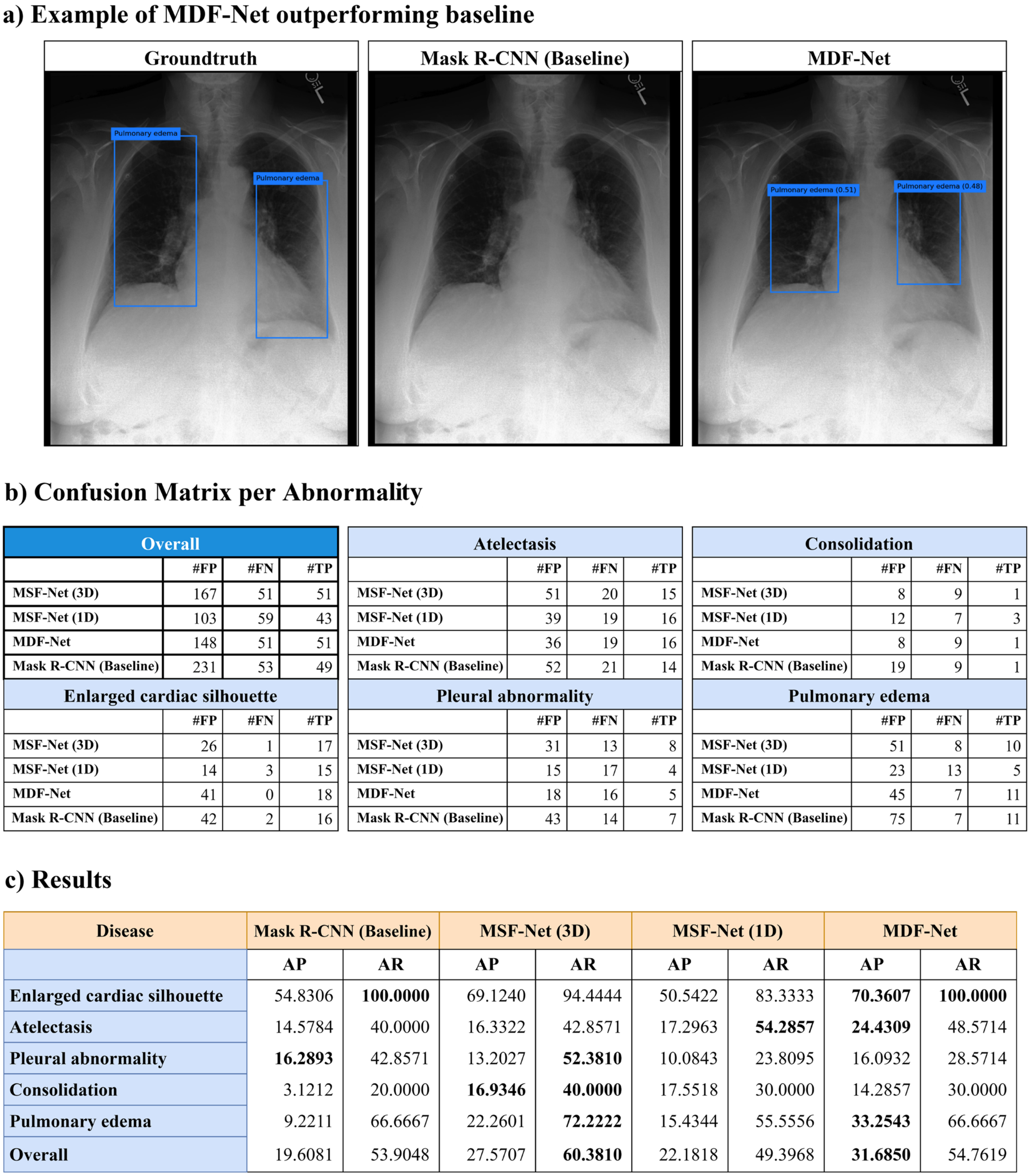


Figure 1. Panel (**a**) An overview of the proposed MDF-Net architecture. Panel (**b**) Integration of the CXR images from MIMIC-CXR[41] with the clinical data of MIMIC IV[42]. Panel (**c**) IoBB is used to evaluate the model between ground truth and predictions.

chest abnormalities that it was trained on: 148 FP and 51 FN for MDF-Net compared to 231 FP and 53 using Mask R-CNN alone (Fig. 2b).

We conducted a series of ablation studies to further validate the efficacy and robustness of the proposed MDF-Net. These systematic experiments were designed to dissect and independently evaluate different components of the proposed system.

Firstly, the effectiveness of our fusion methods was scrutinized together with the different backbone architectures. Fusion methods within the MDF-Net are pivotal as they integrate the distinct feature maps extracted from the chest X-ray images and clinical data, creating a unified representation for the final prediction. By modifying or omitting these fusion methods, we gauged their impact on the model's overall performance, thereby assessing their contribution to the successful disease localization in chest X-rays (Figs. 3, 4).



a) Example of MDF-Net outperforming baseline

b) Confusion Matrix per Abnormality

Overall	#FP	#FN	#TP
MSF-Net (3D)	167	51	51
MSF-Net (1D)	103	59	43
MDF-Net	148	51	51
Mask R-CNN (Baseline)	231	53	49

Atelectasis	#FP	#FN	#TP
MSF-Net (3D)	51	20	15
MSF-Net (1D)	39	19	16
MDF-Net	36	19	16
Mask R-CNN (Baseline)	52	21	14

Consolidation	#FP	#FN	#TP
MSF-Net (3D)	8	9	1
MSF-Net (1D)	12	7	3
MDF-Net	8	9	1
Mask R-CNN (Baseline)	19	9	1

Enlarged cardiac silhouette	#FP	#FN	#TP
MSF-Net (3D)	26	1	17
MSF-Net (1D)	14	3	15
MDF-Net	41	0	18
Mask R-CNN (Baseline)	42	2	16

Pleural abnormality	#FP	#FN	#TP
MSF-Net (3D)	31	13	8
MSF-Net (1D)	15	17	4
MDF-Net	18	16	5
Mask R-CNN (Baseline)	43	14	7

Pulmonary edema	#FP	#FN	#TP
MSF-Net (3D)	51	8	10
MSF-Net (1D)	23	13	5
MDF-Net	45	7	11
Mask R-CNN (Baseline)	75	7	11

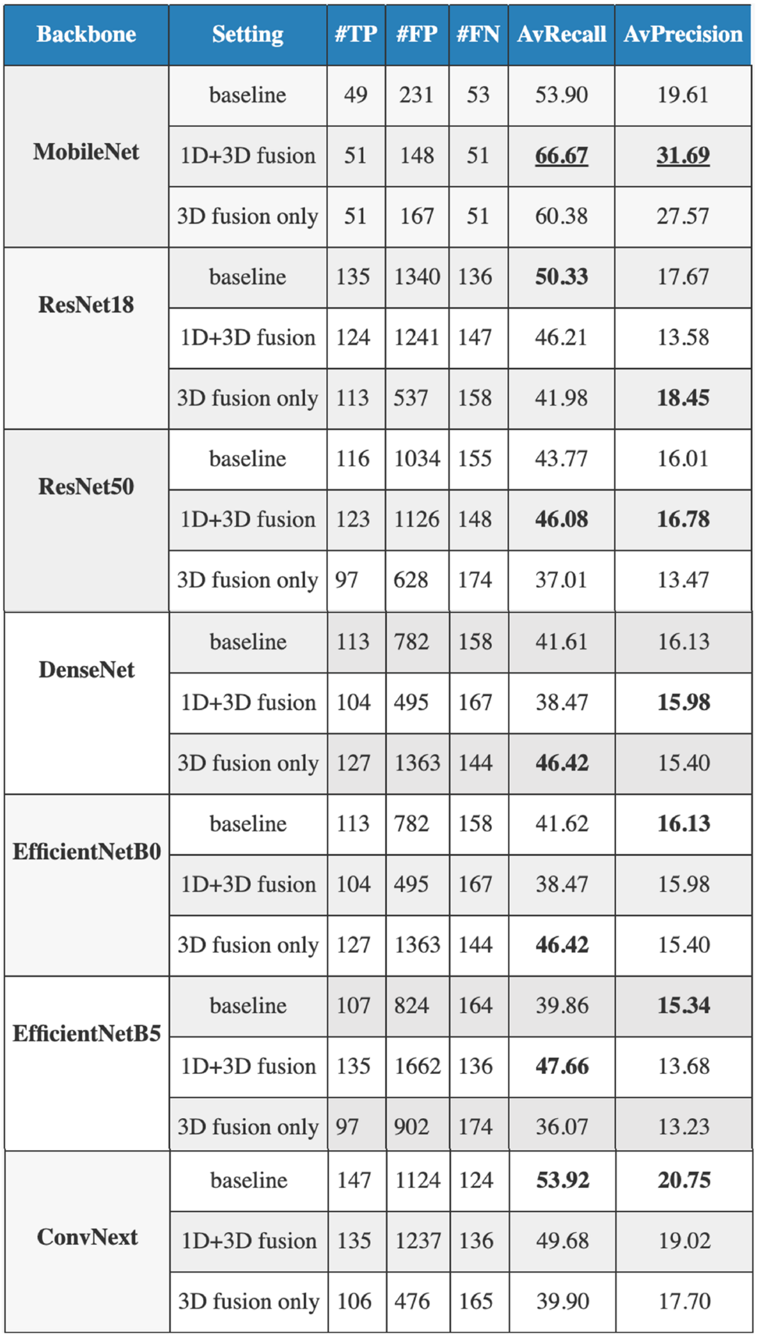
c) Results

Disease	Mask R-CNN (Baseline)		MSF-Net (3D)		MSF-Net (1D)		MDF-Net	
	AP	AR	AP	AR	AP	AR	AP	AR
Enlarged cardiac silhouette	54.8306	**100.0000**	69.1240	94.4444	50.5422	83.3333	**70.3607**	**100.0000**
Atelectasis	14.5784	40.0000	16.3322	42.8571	17.2963	**54.2857**	**24.4309**	48.5714
Pleural abnormality	**16.2893**	42.8571	13.2027	**52.3810**	10.0843	23.8095	16.0932	28.5714
Consolidation	3.1212	20.0000	**16.9346**	**40.0000**	17.5518	30.0000	14.2857	30.0000
Pulmonary edema	9.2211	66.6667	22.2601	**72.2222**	15.4344	55.5556	**33.2543**	66.6667
Overall	19.6081	53.9048	27.5707	**60.3810**	22.1818	49.3968	**31.6850**	54.7619

Figure 2. (**a**) The groundtruth and predictions from Mask R-CNN (Baseline) and MDF-Net. In this figure, MDF-Net successfully identified the presence of pulmonary edema, a condition for which our consulting radiologists indicated the beneficial influence of clinical data in its detection. (**b**) Confusion matrix per abnormality (Note: The number of true negative cases is infinite in the object detection task). (**c**) Evaluation results (Score threshold = 0.05, IoBB threshold = 0.5).

To evaluate the effectiveness of each fusion method proposed in our MDF-NET, we created two Multimodal Single Fusion Networks (MSF-Net), which only apply 1-D or 3-D fusion to conduct ablation studies. Figure 2c presents the performance of Mask R-CNN (Baseline), MDF-Net and the two MSF-Nets. Figure 6b presents the evaluation results across different IoBB thresholds.

Secondly, we investigated the impact of employing different sets of clinical features. Recognizing that each clinical feature may contribute variably to the diagnostic process, we selected varying combinations of these



Backbone	Setting	#TP	#FP	#FN	AvRecall	AvPrecision
MobileNet	baseline	49	231	53	53.90	19.61
	1D+3D fusion	51	148	51	**66.67**	**31.69**
	3D fusion only	51	167	51	60.38	27.57
ResNet18	baseline	135	1340	136	**50.33**	17.67
	1D+3D fusion	124	1241	147	46.21	13.58
	3D fusion only	113	537	158	41.98	**18.45**
ResNet50	baseline	116	1034	155	43.77	16.01
	1D+3D fusion	123	1126	148	**46.08**	**16.78**
	3D fusion only	97	628	174	37.01	13.47
DenseNet	baseline	113	782	158	41.61	16.13
	1D+3D fusion	104	495	167	38.47	**15.98**
	3D fusion only	127	1363	144	**46.42**	15.40
EfficientNetB0	baseline	113	782	158	41.62	**16.13**
	1D+3D fusion	104	495	167	38.47	15.98
	3D fusion only	127	1363	144	**46.42**	15.40
EfficientNetB5	baseline	107	824	164	39.86	**15.34**
	1D+3D fusion	135	1662	136	**47.66**	13.68
	3D fusion only	97	902	174	36.07	13.23
ConvNext	baseline	147	1124	124	**53.92**	**20.75**
	1D+3D fusion	135	1237	136	49.68	19.02
	3D fusion only	106	476	165	39.90	17.70

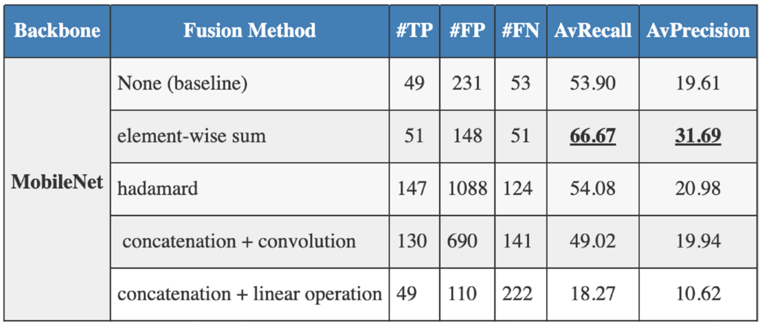
Figure 3. Ablation study results for different backbone architectures in the MDF-Net. This table provides the average precision (AP) and average recall (AR) values obtained using MobileNet, EfficientNet, ResNet, DenseNet, and ConvNextNet and also the overall number of True Positives (TP), False Positives (FP), and (False Negatives).

features to feed into the model. By observing how the model's performance fluctuated with different feature sets, we gained insight into the salience of specific clinical features and their role in accurate abnormality detection (Fig. 5).

Impact of different backbones

To thoroughly evaluate the performance of our proposed MDF-Net model and its robustness across different architectures, we conducted an ablation study focusing on different backbones.

Figure 3 presents the obtained results. Backbone architectures play a pivotal role in deep learning models as they are responsible for the feature extraction process. For this study, we incorporated well-known architectures, including MobileNet, EfficientNet, ResNet, DenseNet, and ConvNextNet, which exhibit diverse architectural designs. We systematically analyzed and compared their performance when incorporated as the backbone of



Backbone	Fusion Method	#TP	#FP	#FN	AvRecall	AvPrecision
MobileNet	None (baseline)	49	231	53	53.90	19.61
	element-wise sum	51	148	51	**66.67**	**31.69**
	hadamard	147	1088	124	54.08	20.98
	concatenation + convolution	130	690	141	49.02	19.94
	concatenation + linear operation	49	110	222	18.27	10.62

Figure 4. Ablation study results showcasing the impact of different fusion methods on the performance of the MDF-Net model. The table compares the performance metrics, namely, average precision and average recall, for the four fusion methods examined: element-wise sum, concatenation followed by a linear operation, concatenation followed by a convolution operation, and the Hadamard product.

our MDF-Net model for different settings: baseline (image only), using 3D fusion only, and using both 3D and 1D fusion. This analysis aims to demonstrate the versatility of our proposed approach and to identify the backbone that can further enhance the performance of MDF-Net in disease localization tasks using chest X-rays and patients' clinical data. The ablation results indicate that MobileNet outperformed all other backbones with an increase of performance of 10.94% AP and 12.75% AR when compared to the second best-performing backbone, the ConvNext net. The characteristics of MobileNet, such as its reduced complexity and efficient learning with small datasets, were likely key contributors to its success in our MDF-Net model over other backbone candidates.

Impact of different fusion methods

Fusion methods play an instrumental role in multimodal deep learning models, acting as a bridge that intertwines the information derived from different data modalities. To scrutinize the effectiveness and compatibility of various fusion methods within our proposed MDF-Net, we conducted an ablation study where we tested different fusion strategies. The strategies assessed included element-wise sum, concatenation followed by a linear operation, concatenation followed by a convolution operation, and the Hadamard product. Each method amalgamates information in distinct ways, carrying unique assumptions about the interplay between the features derived from the image and clinical data. The results of our ablation study illustrate that the element-wise sum fusion method yielded the best performance within our MDF-Net model. This method, which combines features by adding them together element by element, seemingly outperforms state-of-the-art methods at fusing the information from chest X-ray images and clinical data, thus improving the model's ability to localize disease in chest X-rays accurately.

Figure 4 presents our results where one can see that the element-wise outperforms the baseline by 24% in terms of average Recall and average Precision by 62%. Compared to the second-best fusion method, the Hadamard product, the element-wise sum still achieved an increase of 23% in terms of average Recall and 51% in terms of average Precision. The element-wise sum fusion method is likely to have successfully harnessed the complementary information available in our study's chest X-ray images and clinical data, allowing for a more effective fusion and hence better localization of disease in chest X-rays. Its ability to retain original feature information and its computational efficiency further enhance its suitability for this task.

Impact of different clinical features

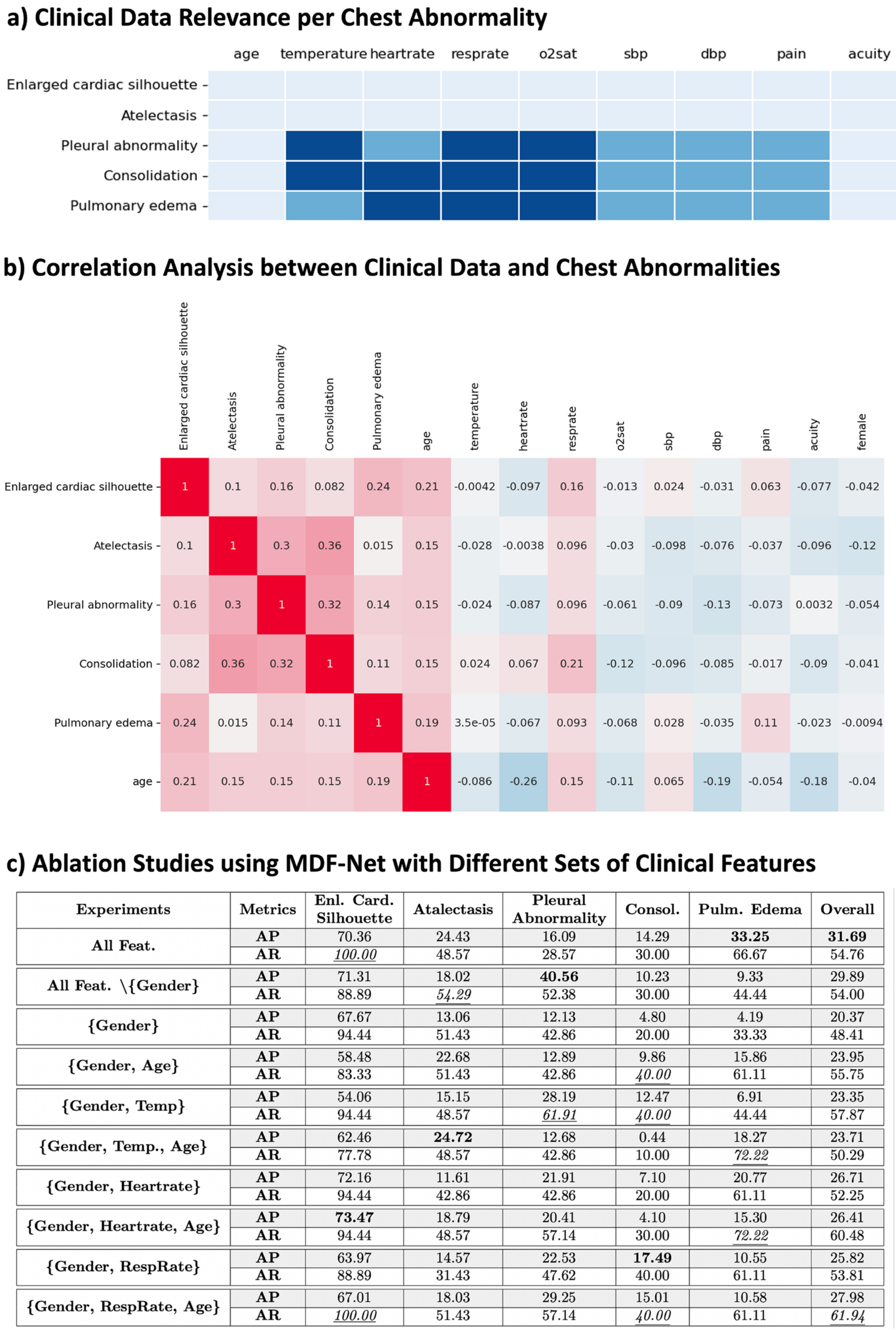
This section of the study examines the role different sets of clinical features play in the context of our MDF-Net architecture.

Firstly, we sought the expertise of radiologists, leveraging their vast domain knowledge to dissect the relationships between specific clinical features and their associations with certain chest abnormalities. For instance, there was a consensus among radiologists that clinical features such as body temperature could serve as a significant indicator of particular diseases which typically cause infections, including abnormalities like consolidation or pleural anomalies.

This initial consultation with radiologists allowed us to gain an intuitive understanding of these relationships, as illustrated in Fig. 5a. However, in order to solidify these findings and corroborate them with empirical evidence, we extended our investigation.

Secondly, we endeavoured to discern if this domain-specific knowledge mirrored the patterns present within our dataset. To this end, a correlation analysis was conducted that evaluated the associations between the chest abnormalities manifest in individual patients and their corresponding clinical data. These findings are graphically demonstrated in Fig. 5b, providing a clear, data-driven correlation between specific clinical features and disease manifestations, such as the correlation between the patient's age and the probability of developing an enlarged cardiac silhouette (older patients are more susceptible to heart problems and therefore to enlarged hearts), or pulmonary diseases such as atelectasis, consolidations or pleural abnormalities.

Finally, we focused on the impact of these clinical feature sets on the performance of the learning model itself. By applying MDF-Net to varied assortments of these clinical features, we could observe how each set influenced



a) Clinical Data Relevance per Chest Abnormality

b) Correlation Analysis between Clinical Data and Chest Abnormalities

c) Ablation Studies using MDF-Net with Different Sets of Clinical Features

Experiments	Metrics	Enl. Card. Silhouette	Atalectasis	Pleural Abnormality	Consol.	Pulm. Edema	Overall
All Feat.	AP	70.36	24.43	16.09	14.29	**33.25**	**31.69**
	AR	*100.00*	48.57	28.57	30.00	66.67	54.76
All Feat. \{Gender}	AP	71.31	18.02	**40.56**	10.23	9.33	29.89
	AR	88.89	*54.29*	52.38	30.00	44.44	54.00
{Gender}	AP	67.67	13.06	12.13	4.80	4.19	20.37
	AR	94.44	51.43	42.86	20.00	33.33	48.41
{Gender, Age}	AP	58.48	22.68	12.89	9.86	15.86	23.95
	AR	83.33	51.43	42.86	*40.00*	61.11	55.75
{Gender, Temp}	AP	54.06	15.15	28.19	12.47	6.91	23.35
	AR	94.44	48.57	*61.91*	*40.00*	44.44	57.87
{Gender, Temp., Age}	AP	62.46	**24.72**	12.68	0.44	18.27	23.71
	AR	77.78	48.57	42.86	10.00	*72.22*	50.29
{Gender, Heartrate}	AP	72.16	11.61	21.91	7.10	20.77	26.71
	AR	94.44	42.86	42.86	20.00	61.11	52.25
{Gender, Heartrate, Age}	AP	**73.47**	18.79	20.41	4.10	15.30	26.41
	AR	94.44	48.57	57.14	30.00	*72.22*	60.48
{Gender, RespRate}	AP	63.97	14.57	22.53	**17.49**	10.55	25.82
	AR	88.89	31.43	47.62	40.00	61.11	53.81
{Gender, RespRate, Age}	AP	67.01	18.03	29.25	15.01	10.58	27.98
	AR	*100.00*	51.43	57.14	*40.00*	61.11	*61.94*

Figure 5. Panel (**a**) The importance of clinical features for detecting each abnormality as reported by our radiology partners. Seven clinical features are highlighted for detecting pleural abnormality, consolidation, and pulmonary edema. Four out of seven highlighted features are considered the most significant. Panel (**b**) Correlation matrix between clinical features and abnormalities. Panel (**c**) Results of several ablation studies with different sets of features. Bold values indicate the best-performing model in terms of AP, and the underlined values indicate the best-performing model in terms of AR.

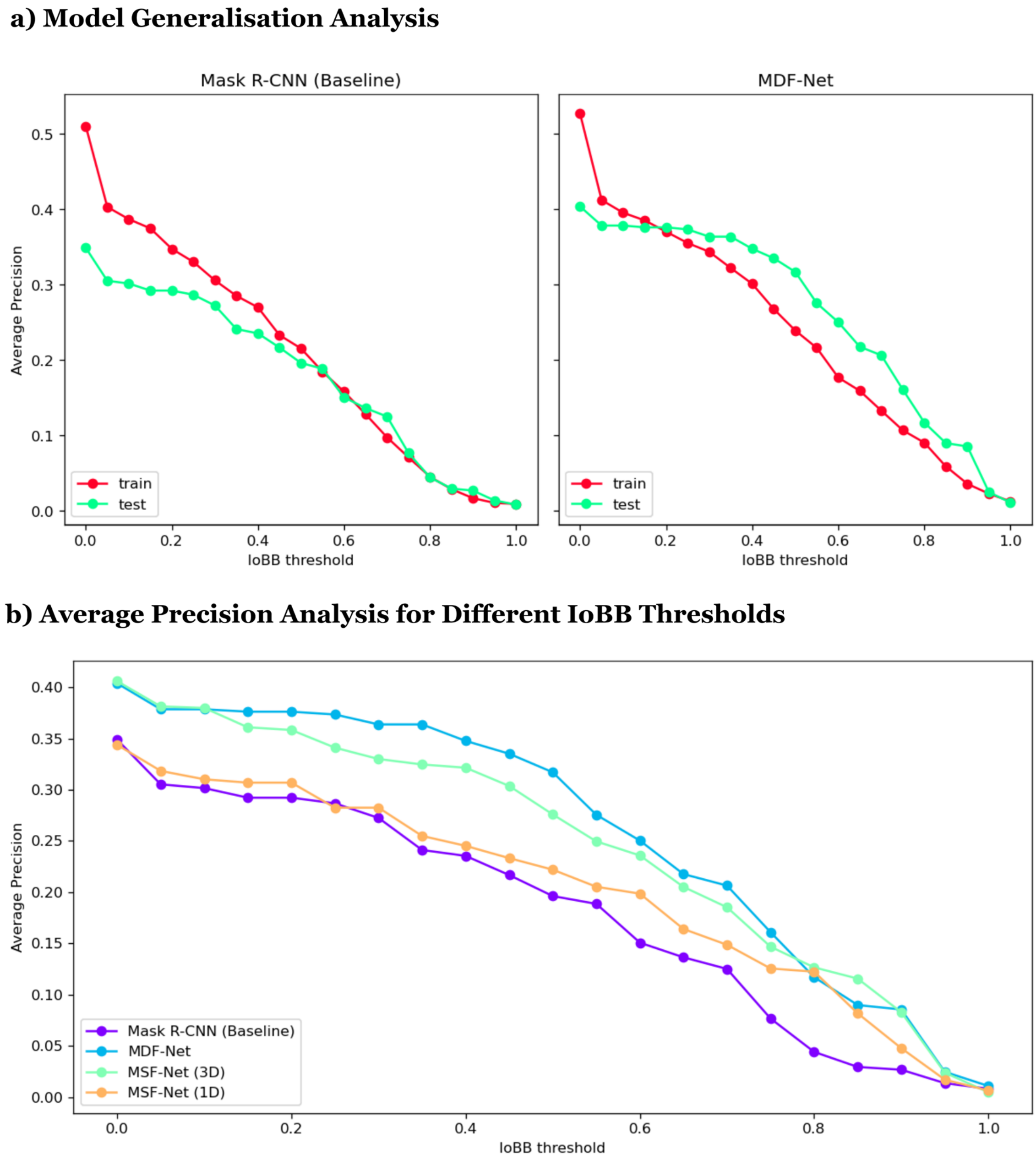


Figure 6. Panel (**a**) Model generalisation analysis. This graph shows that clinical data and MDF-Net improve the generalization ability on the test set. Panel (**b**) Average precision analysis for different IoBB thresholds: This chart shows MDF-Net reached the best performance when using both fusion methods. Using 1-D or 3-D fusion methods alone can already show some improvements in the model among all IoBB thresholds.

the model's learning capabilities. The performance of the MDF-Net model under these different conditions is displayed in Fig. 5c. Results show that the set of all features contributes to the model's best performance, which reinforces the value of clinical data and how it can significantly influence our MDF-Net's capacity to detect and localize abnormalities within chest X-ray images accurately.

Discussion

The comprehensive evaluation elucidated in the prior section underpins the establishment of the following noteworthy observations. The proposed MDF-Net surpasses the standard Mask R-CNN (used as our baseline) in both Average Precision (AP) and Average Recall (AR) metrics. Specifically, MDF-Net demonstrates superior AP on four out of the five assessed lesions, with only a marginal shortfall (– 0.17% AP) relative to the baseline model when detecting the *Pleural Abnormality*.

In terms of global performance, MDF-Net contributes to an enhancement of + 12% AP. Panel (a) in Fig. 6 illustrates the comparative performance of the model on the training set (depicted in red) and the test set (represented in green). Despite the similar performance levels achieved on the training set by both models, MDF-Net demonstrates superior generalization on the test set in contrast to the Mask R-CNN (baseline).

Additionally, as represented in panel (b) of Fig. 6, MDF-Net (indicated in blue), along with the other two MSF-Nets markedly outperform the Mask R-CNN (illustrated in purple) across nearly all Intersection over Bounding Box (IoBB) thresholds. While Mask R-CNN (baseline) performance declines at higher IoBB thresholds, MDF-Net maintains a reasonable performance standard, reinforcing that MDF-Net can effectively pinpoint lesions irrespective of the IoBB threshold.

Collectively, these empirical observations substantiate the pivotal role of clinical data, contravening the prevailing notion that image data alone suffices for robust medical diagnosis.

Ablation study for different fusion processes

The Average Precision (AP) and Average Recall (AR) of each model are shown in Fig. 2c. In the previous subsection, we demonstrated the effectiveness of MDF-Net, and in this section, we use MSF-Net (1D) and MST-Net (3D) to conduct an ablation study. This experiment allowed us to test the effectiveness of each fusion method. The results are as follows. Figure 2c shows MSF-Net (3D) has a larger improvement in performance (+ 7.96%AP, + 6.47%AR) than MSF-Net (1D), which indicates the 3-D fusion is more effective than 1-D fusion. Compared to Mask R-CNN (Baseline), MSF-Net (3D) performs better on 4 out of 5 lesions, the same as MDF-Net. And, We also noticed this model has the highest AR compared to all other models.

The MSF-Net (1D) has a slight improvement on AP by + 2.58% but loss − 4.508% on AR, as shown in Fig. 2c. In MDF-Net (1D), the clinical data are only used to be concatenated with flattened RoIs, which means the clinical data are not involved in deciding where the Regions of Interest are (RPN output). In other words, if we split Mask-RCNN into two stages, the first stage determines the regions of interest (RoIs); and the second stage identifies the lesions inside those RoIs. In 1-D fusion, the clinical data is only perceived by the second stage, so the clinical data is not used for identifying RoIs. Consequently, this 1-D fusion only helps the final classifier to filter out regions misclassified by RPN; hence, AP increased while AR decreased. When using both 1-D and 3-D fusions together, the 3-D fusion can help RPN to pick up abnormal regions better while 1-D fusion can help the final classifier to determine whether a lesion exists in the given region.

Ablation study for different backbones

Since the performance of our model is sensitive to the underlying deep learning mechanism, we examined the influence of different backbones in terms of the overall performance and robustness of the model. Backbone architectures play a central role in deep learning models since they are responsible for the feature extraction process and consequently influence the subsequent layers' performance. Figure 3 shows that MobileNet achieved the best results with an increase of performance of 10.94% AP and 12.75% AR when compared to the second best-performing backbone, the ConvNext net. MobileNet is likely to have achieved this performance due to several factors. Firstly, MobileNet is designed around the principle of depthwise separable convolutions, a technique that significantly reduces the model complexity and computational cost while preserving the performance level[46,47]. This design feature makes MobileNet particularly well-suited for scenarios where the training data are limited. In the presence of a smaller dataset, the lower complexity model mitigates the risk of overfitting, which often leads to better generalization performance when testing on unseen data. Secondly, since MobileNet is an architecture with fewer parameters and operations than other larger, more complex models like ResNet or DenseNet, this enables the model to learn meaningful representations even from a limited amount of data. Lastly, the computational efficiency and the smaller model size of MobileNet can potentially lead to faster convergence during training[46]. This might result in a more optimal model when the amount of training data and training iterations are limited, further explaining its superior performance in our study.

Ablation study for different fusion methods

Additionally, when investigating different fusion methods, we found that the element-wise sum achieved the best results (Fig. 4). The superior performance of the element-wise sum fusion method in our MDF-Net model can be attributed to several factors that are intrinsically linked to the nature of this method and the characteristics of the data we are working with, such as the complementarity of features, the information preservation, and the computational efficiency. For the complementarity of features, the element-wise sum operation assumes that the features from chest X-ray images and clinical data are largely complementary, meaning that the relevant information in each feature does not overlap significantly. This operation allows the model to preserve and combine the unique information contained in both sets of features, resulting in a more comprehensive feature representation. Additionally, the element-wise sum does not introduce additional transformation of the features apart from scaling. This means that it preserves the original feature values to a greater extent (information preservation). This simplicity could prevent the introduction of unnecessary complexity into the model, thereby reducing the risk of overfitting, which is particularly important when dealing with a small dataset such as ours. Finally, element-wise operations are computationally less demanding compared to operations such as matrix multiplication used in other fusion methods. This could lead to a more efficient learning process and potentially better optimization, contributing to improved performance. This is particularly suitable when combined with smaller backbones, such as MobileNet.

Ablation study for clinical features

In order to understand the contribution and significance of clinical features, we also conducted an ablation study by giving a different combination of clinical features to MDF-Net. The performance of different combinations is shown in Fig. 5c. When comparing the ablation result with the necessity table (Figs. 5a,c and correlation matrix (Fig. 5b, we found that radiologists stated that *heartrate* is less important than *temperature* and *resprate* in diagnosing a pleural abnormality (Fig. 5a). According to our ablation experiments (Fig. 5c), we found

$$AP_{\text{gender,heartrate}} < AP_{\text{gender,resprate}} < AP_{\text{gender, temperature}},$$

which follows the importance shown in the necessity table (Fig. 5a). However, when we introduce the *age* feature, we obtain the following inequalities,

$$AP_{\text{gender,age,temp}} < AP_{\text{gender,age,heartrate}} < AP_{\text{gender,age,resprate}},$$

which indicates that *temperature* and *resprate* did not bring more improvement to the model compared to heartrate when age is also used.

In terms of diagnosing pulmonary edema, radiologists consider *temperature* less important than *heartrate* and *resprate*. The same pattern as diagnosing pleural abnormality is shown. *Heartrate* and *resprate* have higher AP than *temperature* when *age* is not used.

Considering the effect of the feature age, (gender, heartrate) is the only combination that has a slight performance drop (− 0.298% AP) when *age* is introduced. The *age* improves the ability to detect Atelectasis but damages the performance of diagnosing Consolidation. In terms of overall performance, *age* gains improvement in most of the combinations, which is the same pattern shown in Fig. 5b that *age* has a higher correlation to most of the abnormalities. Moreover, in the same correlation matrix, we noticed the correlation between Consolidation and *resprate* is higher than *heartrate* and o2sat, and the ablation results also show *resprate* improves models the most. Lastly, although *resprate* has a high correlation with enlarged cardiac silhouette, the feature *heartrate* seems more important in determining this abnormality.

Conclusion

In this paper, we proposed a novel multimodal deep learning architecture, MDF-Net, and two fusion methods for multimodal abnormality detection, which can perceive clinical data and CXR images simultaneously. In MDF-Net, a spatialisation module is introduced to transform 1-D clinical data to 3-D space, which allows us to predict proposals with multimodal data. Additionally, the 1-D fusion is also used to provide clinical information to the final classifier in a residual manner. To test the performance of MDF-NET, we also propose a joining strategy to construct a multimodal dataset for MIMIC-IV. The experiments show that MDF-Net consistently and considerably outperforms the Mask R-CNN (Baseline) mode. Both fusion methods show significant improvements in Average Precision (AP) while applying them together can achieve the best performance.

Overall, our MDF-Net improves upon the baseline Mask R-CNN by enhancing its ability to localize diseases in chest X-ray images and extending its capabilities to incorporate vital clinical context, which is, at the moment, an important missing ingredient in the Deep Learning literature. This results in a more comprehensive and accurate diagnostic tool that can better support healthcare professionals and provide a better rationale for subsequent interpretations of the models.

In the future, we will explore the following two main directions:

In this work, we can only retrieve 670 instances for our multimodal dataset, which is considered small compared to other popular datasets used for X-ray diagnosis. If there are larger-scaled datasets with clinical data available in the future, our work should also be tested on them to have a more objective evaluation.

Our fusion methods can also be applied to other models, such as YOLO[48], SSD[49] and DETR[50]. We will incorporate other architectures to evaluate the effectiveness of our fusion methods.

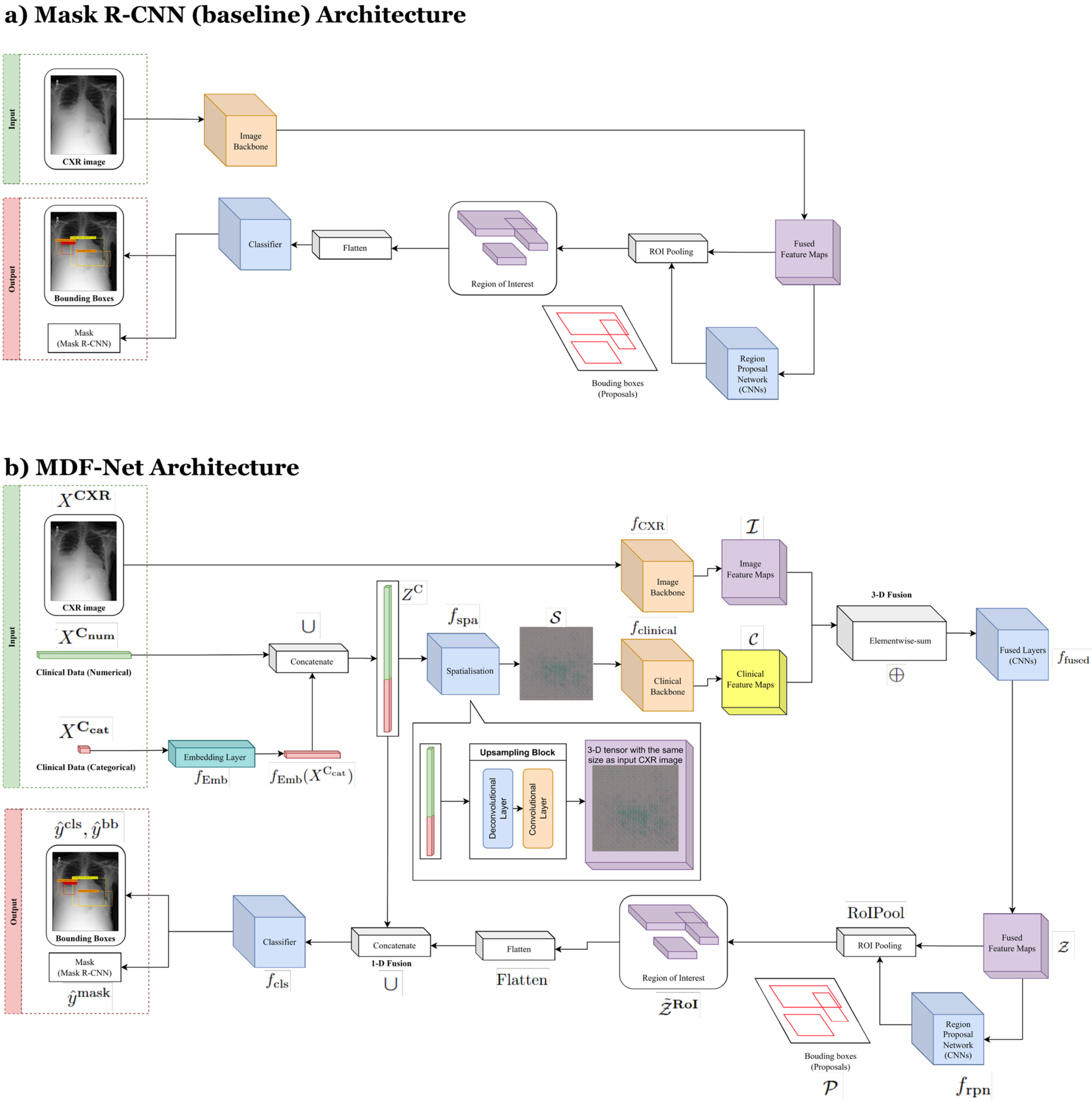
Methods

The methods employed in this research comprise a combination of image processing and machine learning techniques to achieve effective disease detection in chest X-ray images. We have chosen a robust and well-tested algorithm as our foundation and modified it to better suit our specific objectives. This has led to the development of an innovative model that takes advantage of the synergy between traditional image data and structured clinical data. This combined use of data sources significantly enhances the model's diagnostic performance by providing more contextual information for accurate disease localization. The following subsection provides a detailed explanation of the model architecture used in this study.

Model architecture

In this paper, we propose an extension of Mask R-CNN, MDF-Net. We chose Mask R-CNN as our baseline model for the following two reasons. First, Mask R-CNN is a simple state-of-the-art model for object detection and instance segmentation, with proven and established success in localizing and classifying objects within images in a variety of contexts (see for instance[51–53] that testifies its success with a wide range of applications). Given our task of disease localization in chest X-ray images, Mask R-CNN's ability to provide both the class and location of disease indications made it a suitable choice. Furthermore, Mask R-CNN's flexible and modular structure (that is, easy to train) allowed us to extend and modify it to better suit our specific task and incorporate our novel elements.

The main innovation in our MDF-Net is the dual-fusion architecture that allows for the integration of image and structured clinical data (i.e., tabular data). This is a significant departure from traditional Mask R-CNN models, which primarily work with or combine image and text data alone. By integrating clinical data, our model can consider the additional context that is crucial for accurate disease localization, thus making it more aligned with the actual diagnostic process of radiologists. Our model also includes a novel spatialization strategy, which converts clinical data into a 'pseudo-image' format that can be processed by the same convolutional layers as the image data. This significant innovation allows for a more seamless and effective integration of the two data types.



a) Mask R-CNN (baseline) Architecture

b) MDF-Net Architecture

Figure 7. Panel (**a**) Mask R-CNN architecture. Panel (**b**) Our proposed architecture and fusion methods.

Figure 7a shows the architecture of the original Mask R-CNN, which is the baseline model used in this work. The backbone can be any neural network that can extract feature maps from an image. The architecture of MDF-Net is shown in Fig. 7b. Considering the size of the dataset, a small pre-trained backbone model, MobileNetv3[47], is used in both Mask R-CNN (baseline) and MDF-Net to prevent overfitting.

Input layer

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

- a set of CXR images: $X^{\text{CXR}} \in \mathbb{R}^{W \times H \times C}$;
- a set of clinical data (numerical features): $X^{\text{C}_{\text{num}}} \in \mathbb{R}^{n_1 \times 1}$;
- a set of clinical data (categorical features): $X^{\text{C}_{\text{cat}}} \in \{0, 1\}^{n_2 \times 1}$.

In the proposed architecture, we set the dimensions of our image input space to $W = H = 512, C = 1$ (since the CXR image is grey-scale). In terms of the clinical data, our input space corresponds to the number of features: $n_1 = 9$ (continuous variables) and $n_2 = 1$ (categorical variable: gender).

Since the input contains modalities with different dimensions, to perform fusion, we need first to implement feature engineering in this data to have the same dimensions before attempting any localized object detection learning.

Feature engineering $\{X^{C_{\text{num}}}, X^{C_{\text{cat}}}\} \rightarrow \mathscr{C} \in \mathbb{R}^{W' \times H' \times D'}$
$X^{\text{CXR}} \rightarrow \mathscr{I} \in \mathbb{R}^{W' \times H' \times D'}$

The goal of this step is to transform both the image input data and the clinical data to the same shapes. To achieve this, we do the following:

(1) Image transform: We extracted feature maps from $\mathscr{I}$ using a CNN backbone, f_{CXR}, which in this case corresponds to MobileNetv3[47]:

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

where the resulting dimensional space is $\mathscr{I} \in \mathbb{R}^{W' \times H' \times D'}$. In our implementation, MobileNetv3 produced the features maps $W' = H' = 16, D' = 64$. We recognise that MobileNetv3 makes a significant reduction of our feature space; however, when we tried other CNN backbones, such as ResNet, we obtained worse results and overfitting. MobileNetv3 provided the best results in our preliminary tests, hence our reason for choosing this backbone.

(2) Clinical data encoding: In terms of the feature maps for clinical data, the goal is to concatenate numerical and categorical feature representations. To do so, we applied an embedding layer to the categorical features, f_{Emb}, so as to obtain a latent vector with the same dimensions of the numerical data. Next, we concatenate the resulting latent vector $f_{\text{Emb}}(X^{C_{cat}})$ with $X^{C_{num}}$ as follows:

$$Z^{\text{C}} = f_{\text{Emb}}(X^{C_{\text{cat}}}) \cup X^{C_{\text{num}}}, \text{ where } Z^{\text{C}} \in \mathbb{R}^{n \times 1}, \text{ with } n = n_1 + n_2,$$

where in our implementation $n = 64$, n_1 corresponds to the number of continuous features ($n_1 = 9$) and n_2 to the number of categorical features after being processed by an embedding layer ($n_2 = 55$); $\cup$ is the vector concatenation operation. Since the dimensionality $\mathscr{C}$ is different from the generated image feature maps, $\mathscr{I}$, this required an operation of transforming the dimensions of the clinical data from n *times* to $W' \times H' \times D'$. To achieve this, we propose a method of spatialisation of the clinical data.

(3) Spatialisation: We define a spatialisation layer, f_{spa}, as a deconvolutional layer[54] followed by a convolution operation. The deconvolution takes as input the $n \times 1$ dimensional clinical data vector and learns an upscaled representation using a sparse encoded convolution kernel[55,56]. This is given by

$$\mathscr{S} = f_{\text{spa}}(Z^C) = f^e_{L_{\text{spa}}}(Z^C), \tag{1}$$

where $f_{L_{\text{spa}}}(Z^C) = f_{\text{conv}}(f_{\text{deconv}}(Z^C))$, and e is the number of spatialised layers (in this work, we set $e = 9$). After applying spatialisation, the size of $\mathscr{S}$ will become $W \times H \times C$, which is the size of input image X^{CXR}. The primary purpose of using deconvolution, also known as transposed convolution, in this context is to facilitate the upsampling of clinical data to a dimensionality that aligns with the input chest X-ray images. Achieving parity of dimensions between these two data types is crucial as it enables us to acquire feature maps of the same size from both datasets, thereby paving the way for a more effective fusion operation in the subsequent stages of our model.

One of the significant benefits of deconvolution layers within this process is that they render the upsampling procedure trainable. In effect, this means that our model can learn the most efficient transformation strategy for converting the clinical data into a spatial 'pseudo-image', thereby enhancing its ability to integrate with the chest X-ray images and learn from the data concurrently.

(4) Clinical data transform: The last step for clinical data is to extract the feature maps from spatialised clinical data $\mathscr{S}$. By applying a CNN f_{clinical}, which uses the same architecture as f_{CXR}, we can obtain the clinical feature maps by:

$$\mathscr{C} = f_{\text{clinical}}(\mathscr{S}).$$

In the end, with the proposed spatialisation operation f_{spa} and the following CNN f_{clinical}, the resulting $\mathscr{C} \in \mathbb{R}^{W' \times H' \times D'}$, which matches the dimensions of $\mathscr{I}$. From this step, one can proceed to the fusion of both modalities.

Input 3D fusion: $\{\mathscr{I}, \mathscr{C}\} \rightarrow \mathscr{Z} \in \mathbb{R}^{W' \times H' \times D'}$

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

$$\mathscr{Z} = f_{\text{fused}}(\mathscr{C} \oplus \mathscr{I}), z \in \mathbb{R}^{H' \times W' \times D'}, \tag{2}$$

where f_{fused} is another CNN module used for obtaining the features maps from the fused modalities, and $\oplus$ corresponds to the element-wise sum operation that is used for fusion. The final $\mathscr{Z}$ corresponds to the 3D feature map representation of the combined patient information. Next, we use this data representation as input to a Mask-RCNN architecture to perform abnormality detection.

Region proposal network: $\{\mathscr{X}\} \rightarrow \tilde{\mathscr{X}}^{\text{RoI}} \in \mathbb{R}^{W_r \times H_r}$

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

$$\mathscr{P} = f_{\text{rpn}}(\mathscr{X}), \forall p_i \in \mathscr{P} : p_i = (x_i, y_i, w_i, h_i, c_i^{\text{obj}}). \tag{3}$$

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

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

Output: $\{Z^C, \tilde{\mathscr{X}}^{\text{RoI}}\} \rightarrow \hat{y}$:

After learning the candidate RoIs, we flatten this data to serve as input to a normal dense neural network, which will perform the final classification. In order to emphasize the role of the clinical data in this classification process, we concatenate the clinical data representation, Z^C, with the flattened candidate RoIs, $\tilde{\mathscr{X}}^{\text{RoI}}$, before classification takes place. The role of the 1-Diffusion in the MDF-Net is to provide residual information to further pass the clinical data to deeper layers in our architecture. The final prediction $\hat{y}$ is then obtained by:

$$\hat{y} = f_{\text{cls}}(\text{Flatten}(\tilde{\mathscr{X}}^{\text{RoI}}) \cup Z^{\text{C}}), \tag{4}$$

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

Number of classes In this study we make object detection over five different classes: Enlarged Cardiac Silhouette, Atelectasis, Consolidation, Pleural Abnormality, and Pleural Edema. We chose to focus on five classes because they were the most representative in our dataset. We also took into account the constraints imposed by the dataset's class imbalance. The classes we have not included had insufficient examples, which would have posed significant challenges in training and evaluating our deep learning model. Training a deep learning model with insufficient data for some classes could lead to overfitting and poor generalization performance for those classes. Moreover, it could bias the model towards the classes with more data. Therefore, to ensure a reliable and robust model, we chose to focus on the five most representative classes.

Model complexity analysis

The overall computational complexity of the proposed architecture approximates the original Mask R-CNN model, which corresponds to the sum of the complexities of its components:

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

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

The Faster-R CNN is composed of 5 main parts as follows: (1) a deep fully convolutional network, (2) region proposal network, (3) ROI pooling and fully connected networks, (4) bounding box regressor, and (5) classifier.

The deep fully convolution network consists of five convolution layers. It is based on Zeiler and Fergus's *fast* (smaller) model[57]. From an image I, this step extracts $256 \times N \times N$ feature maps (given the five convolution layers[57]). This is the input of the RPN network and ROI pooling layer. In the RPN network, for each point of the feature map, there are, say K, anchors (or candidate window, typically 2000) with different scales and rations. Thus, there will be a total of N × N × K candidate widows. Non-maximum suppression allows us to obtain typically 2000[58] candidate windows. This yields a complexity of $O(N^2)$.

Using the candidate windows and the feature map above, the RoI pooling layer divides the varied size candidate windows into an $H \times W$ grid of sub-windows, then max-pooling the values in each sub-window into the corresponding output grid cell. The complexity of this process is $O(1)$.

Training

Once the model architecture has been established, the next crucial step involves training the model using a carefully designed loss function to achieve optimal performance. In the context of Mask R-CNN, the loss function plays a pivotal role in learning the optimal parameters of the model. A well-constructed loss function balances multiple objectives, including accurate classification of abnormalities, precise bounding box regression, and efficient object proposal.

In our training process, we incorporate five loss terms, each addressing a specific aspect of the model's learning objectives. These loss terms aim to guide the model toward achieving high precision in identifying and localizing diseases in chest X-ray images. In the following, we provide a detailed explanation of each of these loss terms.

- L_{cls}: Cross-entropy between groundtruth abnormality y^{cls} and predicted abnormalities $\hat{y}^{\text{cls}}$. This loss term requires the model to predict the class of abnormalities correctly in the output layer.
- L_{bb}: Bounding box regression loss between ground-truth bounding boxes y^{bb} and predicted bounding boxes $\hat{y}^{\text{cls}}$ calculated using smooth-L_1 norm:

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

β is a hyperparameter. In our implementation, $\beta = \frac{1}{9}$. To minimise this loss, the model has to locate abnormalities in the correct areas in the output layer.

- L_{mask}: Binary cross-entropy loss between ground-truth segmentation y^{mask} and predicted masks $\hat{y}^{\text{mask}}$, which requires the model to locate abnormalities at the pixel level.
- $L_{\text{obj}_{\text{rpn}}}$: Binary cross-entropy loss between ground-truth objectness y^{obj} and predicted objectness c^{obj} (confidence score), which requires RPN to correctly classify whether the proposals (candidate bounding boxes) contain any abnormality.
- $L_{\text{bb}_{\text{rpn}}}$: Proposal regression loss between proposals (candidate bounding boxes) $p : p \in \mathscr{P}$ and ground-truth bounding boxes y^{bb}, which is also calculated using the same smooth-L_1 norm function for L_{bb}. This loss term aims to improve RPN on localising abnormalities.

We used homoscedastic (task) uncertainty[59] to train the proposed model using these five loss terms by dynamically weighting them for better convergence. Let $\mathscr{L} = \{L_{\text{cls}}, L_{\text{bb}}, L_{\text{mask}}, L_{\text{obj}_{\text{rpn}}}, L_{\text{bb}_{\text{rpn}}}\}$, we used SGD (stochastic gradient descent) to optimise the overall loss function

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

where θ is the wights of MDF-Net, and α_l is a trainable parameter to weigh each task/loss.

The validation of this architecture required a multimodal dataset. In this study, we used medical data, more specifically chest X-ray images from MIMIC-CXR[44] and patient's clinical data from MIMIC-IV-ED[45]. However, these datasets are offered separately in the literature, and a thorough data integration had to be conducted before evaluating our architecture.

Dataset

Modern medical datasets integrate both imaging and also tabular data. The latter refers to medical history and lifestyle questionnaires, where clinicians have the responsibility to combine the above two sources of information. Note also that beyond diagnosis, multimodal data (i.e., comprising tabular and image data) is crucial to the advance and understanding of diseases, motivating the creation of the so-called biobanks. There are several examples of biobanks, for instance, German National Cohort[60] or the UK Biobank[61] that includes thousands of data fields from patient questionnaires including data from questionnaires, physical measures, sample essays, accelerometry, multimodal imaging, genome-wide genotyping. However, these datasets do not contain local image annotations of lesions. Therefore, we propose a strategy to combine MIMIC-IV[42] and REFLACX[62] to create our dataset, MIMIC-Eye, from scratch that meets the requirement of this work which can be accessed in physionet[43]. The Medical Information Mart for Intensive Care (MIMIC) IV dataset is from two in-hospital database systems, a custom hospital-wide EHR and an ICU-specific clinical information system, in Beth Israel Deaconess Medical Center (BIDMC) between 2011 and 2019. The MIMIC-IV database is grouped into three modules, including *core*, *hosp*, and *icu*. This work uses only the patient's data in the *core* module.

As well as the MIMIC-IV dataset, two other MIMIC-IV subsets, MIMIC-IV ED (Emergency Department)[45], and MIMIC-IV CXR (Chest X-ray)[44], are used to create the multimodal dataset. These two datasets can be linked to the MIMIC-IV dataset with *subject_id* and *stay_id*. The MIMIC-IV ED dataset was extracted from the Beth Israel Deaconess Medical Center emergency department. It contains data for emergency department patients collected while they are in the ED. The *triage* data of the MIMIC-IV ED dataset is one source providing patients' health condition in this work, such as *temperature*, *heart rate*, *resprate*, etc. MIMIC-CXR is another subset of MIMIC-IV consisting of 227,835 radiographic studies and 377,110 radiographs from BIDMC EHR between 2011 - 2016. In the original MIMIC-CXR dataset, the CXR images are provided in *DICOM* format, which allows radiologists to adjust the exposure during reading. However, the *JPG* file is preferred to train a machine learning model. The author of MIMIC-IV CXR then presented the MIMIC-CXR JPG dataset[41] to facilitate the training process. REFLACX dataset is another subset of MIMIC-IV ED, which provides extra data from different modalities, such as eye tracking data, bounding boxes, and time-stamped utterances. The bounding boxes in REFLACX are used as ground truth in this work. The aggregation of this dataset into a single multimodal dataset constitutes the MIMIC-EYE dataset[43].

In total, ten clinical features are used in this research. The MIMIC-IV Core *patients* data include only two clinical attributes, age, and gender. And the other eight clinical features are extracted from the MIMI-IV ED *triage* data. The explanations for these eight clinical features in the MIMIC-IV documentation are:

1. Temperature: The patient's temperature is in degrees Fahrenheit.
2. Heartrate: The patient's heart rate in beats per minute.
3. Resprate: The patient's respiratory rate in breaths per minute.
4. o2sat: The patient's peripheral oxygen saturation as a percentage.
5. sbp, dbp: The patient's systolic and diastolic blood pressure, respectively, were measured in millimetres of mercury (mmHg).
6. Pain: The patient self-reported pain level on a scale of 0-10.
7. Acuity: An order of priority. Level 1 is the highest priority, while level 5 is the lowest priority.

Before explaining the creation process, it is necessary to introduce some important IDs and data tables in the MIMIC-IV dataset. Four important IDs are used in MIMIC-IV to link the information across tables. They are:

- subject_id (patient_id): ID specifying an individual patient.
- stay_id: ID specifying a single emergency department stay for a patient.
- study_id: ID specifying a radiology report written for the given chest x-ray. It is rarely mentioned because we do not use the report as the ground truth label in this paper.
- dicom_id: ID specifying a chest x-ray image (radiograph).

And the following four tables in MIMIC-IV are used to create our multimodal abnormality detection dataset:

- MIMIC-IV Core patients: Information that is consistent for the lifetime of a patient is stored in this table, including age and gender.
- MIMIC-IV ED triage: This table contains information about the patient when they were first triaged in the emergency department, including temperature, heart rate, and more clinical data.
- MIMIC-IV Core edstays: Provides the time the patient entered the emergency department and the time they left the emergency department, which helps us to identify the *stay_id* for CXR images.
- MIMIC-IV CXR metadata: Contains the information about the CXR image (radiograph), including the time taken, height, and width.

Limitations and ethical considerations

Although our study presents promising results, several limitations should be acknowledged. As with all scientific research, our study contains inherent limitations, primarily revolving around data selection, fusion methods, and ethical considerations:

Dataset choice

The effectiveness of our proposed model relies heavily on the quality and comprehensiveness of the input information, both images and clinical data. While using publicly available and well-established datasets such as MIMIC-CXR, MIMIC IV-ED, REFLACX, and MIMIC-EYE minimizes the risk of data quality issues, these datasets may only partially represent diverse global populations. The performance of our model could vary when applied to different demographic groups, and potential biases in the datasets could influence the results.

Dataset size

Although many public datasets contain both CXR images and manual lesion annotations, unfortunately, to the best of our knowledge, we are unaware of any dataset that also contains the patients' clinical data. Due to privacy concerns, most publicly available medical image datasets do not include this kind of clinical information. Patient clinical data are sensitive and protected by strict privacy regulations. As a result, researchers often face significant challenges in obtaining datasets that combine imaging data with relevant clinical information. This policy limits the effectiveness and ability of our MDF-Net to generalize.

Limitations in fusion methods

For our study, we considered other fusion methods, such as the Laplacian pyramid and adaptive sparse representation[63], for the 3D fusion component in the proposed MDF-Net. However, these methods are not differentiable, which makes them incompatible with our end-to-end deep learning architecture. The backpropagation process used to train deep learning models requires the gradients (derivatives) of the loss concerning the model parameters. Non-differentiable operations disrupt this gradient flow, which could lead to suboptimal or untrainable models.

While the development and application of multimodal deep learning technologies have the potential to enhance disease diagnosis significantly, several ethical considerations must be addressed to ensure that such technologies are used responsibly and effectively.

Impact on healthcare professionals

While applying deep learning technologies may streamline diagnostic procedures and alleviate the workload of healthcare professionals, we must consider the potential impact on their roles and responsibilities. In this study, we align our work with the perspective that these technologies are tools that *can support, not replace*, healthcare professionals. Our interviews with radiologists highlighted the importance of integrating clinical data in the image diagnosis process, emphasizing the continued need for expert knowledge and more human-centred deep learning architectures as proposed in this paper.

Impact on trust

The black-box nature of deep learning models raises concerns about transparency and trust in AI decisions. Our multimodal DL architecture seeks to improve the interpretability of predictions by using clinical information alongside image data, providing a context for the decisions made by the model. Further development of these technologies must emphasize explainability, so healthcare professionals and patients can understand and trust the diagnoses provided by these models. By providing lesion detection of the predicted lesions, we are already promoting one layer of interpretability. For future work, we are already developing methods to translate the symbolic representation of identified lesion bounding boxes into human-level explanations incorporating domain knowledge.

Biases

Potential biases in the model's predictions, resulting from biased or unrepresentative training data, can lead to disparities in healthcare outcomes. Care must be taken to ensure that datasets used to train such models are representative of the diverse patient populations they will serve. In cases where data is imbalanced, techniques such as oversampling, undersampling, or synthetically augmenting the minority class should address this issue. However, this may reinforce potential selection biases in the dataset. For this reason, we restricted our study to the most frequent classes in the dataset (ending up in a much smaller dataset that impacted our model's performance, rather than introducing selection and sampling biases from data augmentations techniques. ‘

Data availability

The code for this work is available on GitHub at https://tinyurl.com/multimodal-abn-detection.

Received: 20 January 2023; Accepted: 27 August 2023
Published online: 23 September 2023

References

1. Haakenstad, A. *et al.* Measuring the availability of human resources for health and its relationship to universal health coverage for 204 countries and territories from, to 2019: A systematic analysis for the global burden of disease study 2019. *Lancet* **399**(2129–2154), 1990. https://doi.org/10.1016/s0140-6736(22)00532-3 (2022).
2. Maicas, G., Bradley, A. P., Nascimento, J. C., Reid, I. & Carneiro, G. Pre and post-hoc diagnosis and interpretation of malignancy from breast DCE-MRI. *Med. Image Anal.*https://doi.org/10.1016/j.media.2019.101562 *(2019).*
3. Shen, L. *et al.* Deep learning to improve breast cancer detection on screening mammography. *Sci. Rep.* **9**, 2045–2322 (2019).
4. Liu, X. *et al.* Deep learning-based automated left ventricular ejection fraction assessment using 2-d echocardiography. *J. Physiol. Heart Circ. Physiol.* **321**, H390–H399 (2020).
5. Medley, D. O., Santiago, C. & Nascimento, J. C. Cycoseg: A cyclic collaborative framework for automated medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 8167–8182. https://doi.org/10.1109/TPAMI.2021.3113077 (2022).
6. Pham, T.-C., Luong, C.-M., Hoang, V.-D. & Doucet, A. Ai outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function. *Sci. Rep.* **11**, 17485 (2021).
7. Haenssle, H. *et al.* Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **29**, 1836–1842 (2018).
8. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence* 590–597 (2019).
9. Rajpurkar, P. *et al.* Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning. CoRR **abs/1711.05225** 1711.05225 (2017).
10. Rajpurkar, P. *et al.* Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS Med.* **15**, 1–17. https://doi.org/10.1371/journal.pmed.1002686 (2018).
11. Yates, E., Yates, L. & Harvey, H. Machine learning "red dot": open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. *Clin. Radiol.* **73**, 827–831. https://doi.org/10.1016/j.crad.2018.05.015 (2018).
12. Moreira, C. *et al.* 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, 2023). https://doi.org/10.1145/3588015.3588403.
13. Rahimi, S., Oktay, O., Alvarez-Valle, J. & Bharadwaj, S. Addressing the exorbitant cost of labeling medical images with active learning. In *International Conference on Machine Learning in Medical Imaging and Analysis* (2021).
14. Guidotti, R. *et al.* A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**, 931–9342 (2018).
15. Lipton, Z. C. The mythos of model interpretability. *ACM Commun.* **61**, 36–43 (2018).
16. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. intell.* **267**, 1–38 (2019).
17. Moreira, C. *et al.* Linda-bn: An interpretable probabilistic approach for demystifying black-box predictive models. *Decis. Support Syst.* **150**, 113561 (2021).
18. Sindhgatta, R., Ouyang, C. & Moreira, C. Exploring interpretability for predictive process analytics. In *Proceedings of the 18th International Conference on Service Oriented Computing (ICSOC)* (2020).
19. Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C. & Jorge, J. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Inf. Fusion* **81**, 59–83 (2022).
20. Singh, A., Sengupta, S. & Lakshminarayanan, V. Explainable deep learning models in medical image analysis. *J. Imaging.* **6** (2020).
21. Zhuang, F. *et al.* A comprehensive survey on transfer learning. CoRR **abs/1911.02685** 1911.02685 (2019).
22. Yuan, Z., Yan, Y., Sonka, M. & Yang, T. Robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. CoRR **2012.03173** 2012.03173 (2020).
23. Tang, Y.-X. *et al.* Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ Digit. Med.*https://doi.org/10.1038/s41746-020-0273-z *(2020).*
24. He, K., Gkioxari, G., Dollár, P. & Girshick, R. B. Mask R-CNN. CoRR **1703.06870** (2017).
25. Liu, M. *et al.* Aa-wgan: Attention augmented Wasserstein generative adversarial network with application to fundus retinal vessel segmentation. *Comput. Biol. Med.* **158**, 106874 (2023).
26. Wu, P. *et al.* Aggn: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion. *Comput. Biol. Med.* **152**, 106457 (2023).
27. Li, H., Zeng, N., Wu, P. & Clawson, K. Cov-net: A computer-aided diagnosis method for recognizing covid-19 from chest X-ray images via machine vision. *Expert Syst. Appl.* **207**, 118029 (2022).
28. Bayoudh, K., Knani, R., Hamdaoui, F. & Mtibaa, A. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *Vis. Comput.*https://doi.org/10.1007/s00371-021-02166-7 *(2021).*
29. Wang, Y. *et al.* Deep multimodal fusion by channel exchanging. *Adv. Neural Inf. Process. Syst.* **33**, 4835–4845 (2020).
30. Badgeley, M. A. *et al.* Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digit. Med.* **2**, 1–10 (2019).
31. Lahat, D., Adali, T. & Jutten, C. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proc. IEEE* **103**, 1449–1477 (2015).
32. Ramachandram, D. & Taylor, G. W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Mag.* **34**, 96–108. https://doi.org/10.1109/MSP.2017.2738401 (2017).

33. Smilkov, D., Thorat, N., Kim, B., Viégas, F. B. & Wattenberg, M. Smoothgrad: removing noise by adding noise. CoRR **1706.03825** (2017).
34. Luís, A. *et al.* Integrating eye-gaze data into cxr dl approaches: A preliminary study. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* 196–199. https://doi.org/10.1109/VRW58643.2023.00048 (2023).
35. Azam, M. A. *et al.* A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput. Biol. Med.* **144**, 105253. https://doi.org/10.1016/j.compbiomed.2022.105253 (2022).
36. Moon, J. H., Lee, H., Shin, W., Kim, Y.-H. & Choi, E. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE J. Biomed. Health Inform.* 1–10 (2022).
37. Yan, B. & Pei, M. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence* (2022).
38. Chen, Z., Li, G. & Wan, X. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia* 5152–5161 (2022).
39. Oakden-Rayner, L. *et al.* Producing radiologist-quality reports for interpretable deep learning. In *IEEE 16th International Symposium on Biomedical Imaging* (2019).
40. Castillo, C., Steffens, T., Sim, L. & Caffery, L. The effect of clinical information on radiology reporting: A systematic review. *J. Med. Radiat. Sci.* **68**, 60–74. https://doi.org/10.1002/jmrs.424 (2021).
41. Johnson, A. E. W. *et al.* MIMIC-CXR-JPG: A large publicly available database of labeled chest radiographs. CoRR (2019). 1901.07042.
42. Johnson, A. *et al.* Mimic-iv. https://doi.org/10.13026/S6N6-XD98 (2021).
43. Hsieh, C. *et al.* Mimic-eye: Integrating mimic datasets with reflacx and eye gaze for multimodal deep learning applications. PhysioNet (version 1.0.0) (2023).
44. Johnson, A. E. W., Pollard, T., Mark, R., Berkowitz, S. & Horng, S. The mimic-cxr database. https://doi.org/10.13026/C2JT1Q (2019).
45. Johnson, A. *et al.* Mimic-iv-ed. https://doi.org/10.13026/77Z6-9W59 (2021).
46. Howard, A. G. *et al.* Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR **1704.04861**. abs/1704.04861 (2017).
47. Howard, A. *et al.* Searching for mobilenetv3. CoRR **1905.02244**. 1905.02244 (2019).
48. Redmon, J., Divvala, S. K., Girshick, R. B. & Farhadi, A. You only look once: Unified, real-time object detection. CoRR **1506.02640**. 1506.02640 (2015).
49. Liu, W. *et al.* Ssd: Single shot multibox detector. In *Computer Vision—ECCV 2016* (eds. Leibe, B., Matas, J., Sebe, N. & Welling, M.) 21–37 (Springer International Publishing, 2016).
50. Carion, N. *et al.* End-to-end object detection with transformers. CoRR **2005.12872**. 2005.12872 (2020).
51. Schweitzer, D. & Agrawal, R. Multi-class object detection from aerial images using mask r-cnn. In *2018 IEEE International Conference on Big Data (Big Data)* 3470–3477 (2018).
52. Liu, H. & Bhanu, B. Pose-guided r-cnn for jersey number recognition in sports. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 2457–2466. https://doi.org/10.1109/CVPRW.2019.00301 (2019).
53. Conrady, C. R., Şebnem, E., Attwood, C. G., Roberson, L. A. & de Vos, L. Automated detection and classification of southern African roman seabream using mask r-cnn. *Ecol. Inform.* **69**, 101593 (2022).
54. Zeiler, M. D., Krishnan, D., Taylor, G. W. & Fergus, R. Deconvolutional networks. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2528–2535. https://doi.org/10.1109/CVPR.2010.5539957 (2010).
55. Dumoulin, V. & Visin, F. A guide to convolution arithmetic for deep learning. https://doi.org/10.48550/ARXIV.1603.07285 (2016).
56. Albawi, S., Mohammed, T. A. & Al-Zawi, S. Understanding of a convolutional neural network. In *International Conference on Engineering and Technology (ICET)* 1–6.https://doi.org/10.1109/ICEngTechnol.2017.8308186 (2017).
57. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. *Comput. Vis. ECCV* **2014**, 818–833 (2014).
58. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, vol. 28 (Curran Associates, Inc., 2015).
59. Kendall, A., Gal, Y. & Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. CoRR **1705.07115** (2017).
60. German National Cohort (GNC) Consortium geschaeftsstelle@ nationale-kohorte. de. The german national cohort: Aims, study design and organization. *Eur. J. Epidemiol.* **29**, 371—382 (2014).
61. Sudlow, C. *et al.* Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **31**, e1001779 (2015).
62. Bigolin Lanfredi, R. *et al.* Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest X-rays. *Sci. Data* **9**, 350 (2022).
63. Wang, Z., Cui, Z. & Zhu, Y. Multi-modal medical image fusion by laplacian pyramid and adaptive sparse representation. *Comput. Biol. Med.* **123**, 103823 (2020).

Acknowledgements

This material is based upon 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.

Author contributions

C.H.: Conceptualisation, methodology, implementation, mathematical formulation, evaluation, writing. I.N.: Study idea and original insight, clinical knowledge, review. S.C.: Study idea and original insight, clinical knowledge, review. C. O.: Supervision, review, writing. M.B.: Supervision review, writing. J.N.: Conceptualization, methodology, writing, review. J.J.: Review, writing. C.M.: Conceptualization, methodology, mathematical formulation, writing, supervision, review.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.M.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.