

The Supplementary File for Large-scale Long-tailed Disease Diagnosis on Radiology Images

Qiaoyu Zheng^{*,1,2}, Weike Zhao^{*,1,2}, Chaoyi Wu^{*,1,2}, Xiaoman Zhang^{1,2}, Lisong Dai^{1,3},
Hengyu Guan^{4,5}, Yuehua Li^{1,3}, Ya Zhang^{1,2}, Yanfeng Wang^{1,2,†} and Weidi Xie^{1,2,†}

^{*}Equal contributions. [†]Corresponding author.

¹Shanghai Jiao Tong University ²Shanghai AI Laboratory

³Shanghai Sixth People’s Hospital, Affiliated to Shanghai Jiao Tong University

⁴Department of Reproductive Medicine, Ren Ji Hospital, Shanghai Jiao Tong University School of
Medicine

⁵Shanghai Key Laboratory for Assisted Reproduction and Reproductive Genetics

<https://qiaoyu-zheng.github.io/RP3D-Diag>

Email addresses: {three-world, zwk0629, wtzxxxwcy02, weidi}@sjtu.edu.cn,

A Implementation Details

At training time, we consider two diagnosis tasks in different granularities: Disorder-level classification (5569 classes) and ICD10-level classification (931 Classes). The image input will all be resized to $256 \times 256 \times D$ in height, width, and depth respectively. For 3D scans, the depth is treated as a factor for ablation study, and will be discussed in the experiment section. In the vision encoder, we adopt two separate modules for normalization and a shared module to compute the final embedding. The detailed architecture design will also be further discussed in our ablation study. In fusion module, we use a 2-layer, 4-head transformer encoder with a learnable “[CLS]” token for final prediction.

For augmentation, we adopt Gaussian Noise, Contrast Adjustment, Affine Variation, and Elastic Deformation, implemented from the MONAI [1] package. For optimization, we utilize the AdamW optimizer with a cosine learning rate curve, setting the maximum learning rate to $lr = 1 \times e^{-5}$, with an adjustable batch size ranging from 4 to 32 depending on the input image depth and model scale to avoid out-of-memory error. The total training duration spans 100 epochs, with the initial 5 epochs for warm-up. At fine-tuning stage, we adopt a similar model architecture and optimization setting.

B Fusion Strategy

In this part, we will discuss the fusion strategy for baseline implementation in detail and this will be adopted for all the compared methods that are without fusion module.. Since most existing architectures cannot perform case-level diagnosis, we propose three parameter-free methods to give case-level prediction with a scan-level classifier:

- **Random Picking.** We randomly select an image scan a case. It is based on the belief that each scan in a case should encompass the essential information required for an accurate.
- **Max Pooling.** We adopt max pooling on all the images in a case to get the final prediction. It is based on the principle that if there is an image within a case that can illustrate the presence of abnormality, we should consider it as an unhealthy case.
- **Mean Pooling.** Instead of max pooling, we replace it with mean pooling in this case. It is rooted in the notion that relying solely on the diagnosis from a single image is insufficient, and our goal is to comprehensively leverage all the scan image information in a case.

Subsequently, we will present a comparative analysis of the outcomes obtained through these strategies,

focusing on the two classification standards, ICD-10-CM and disorders. As shown in the Supplementary Table 1 and Supplementary Table 2, in most cases “Max Pooling” outperform others. Therefore, in the main paper, we adopt max pooling as our baseline for comparison.

Supplementary Table 1 | Results on ICD10-CM classes. We report the different scores for Head/Medium/Tail classes, “Rand”, “Max”. “Mean” represent the random picking, Max Pooling and Mean Pooling respectively.

Granularity	Split	Mode	Metrics						
			AUC	AP	F1	MCC	R@0.01	R@0.05	R@0.1
ICD-10-CM	Head	Rand	89.12	12.39	20.51	21.72	24.12	51.82	67.93
		Max	90.39	12.80	20.65	21.94	24.39	54.12	70.11
		Mean	85.49	13.05	20.25	21.52	24.31	49.92	65.28
	Medium	Rand	88.98	7.72	15.44	18.02	22.52	50.50	63.96
		Max	90.15	7.74	15.41	17.96	22.75	51.10	66.31
		Mean	85.05	7.69	15.03	17.53	22.58	47.41	64.52
	Tail	Rand	86.41	4.46	8.54	12.48	8.37	23.74	37.12
		Max	86.46	2.90	6.09	10.27	8.51	25.06	42.80
		Mean	81.50	2.68	5.42	9.04	6.15	20.13	30.64

Supplementary Table 2 | Results on Disorder classes. We report the different scores for Head/Medium/Tail classes, “Rand”, “Max”. “Mean” represent the random picking, Max Pooling and Mean Pooling respectively.

Granularity	Split	Mode	Metrics						
			AUC	AP	F1	MCC	R@0.01	R@0.05	R@0.1
Disorders	Head	Rand	91.32	14.08	22.93	24.61	30.68	58.78	73.79
		Max	91.91	13.56	22.08	23.66	30.88	60.50	74.70
		Mean	88.72	14.29	23.26	24.60	31.97	59.67	71.22
	Medium	Rand	92.63	9.47	17.62	20.61	28.18	57.04	71.42
		Max	93.62	9.19	17.16	20.42	28.24	59.32	73.80
		Mean	90.97	9.54	17.33	20.42	29.31	58.15	71.00
	Tail	Rand	87.61	4.85	8.66	13.26	8.58	24.25	41.39
		Max	88.55	4.34	8.03	12.96	9.13	27.46	41.43
		Mean	84.50	4.35	8.09	12.80	9.38	25.96	42.57

C Ablation Study

To explore the optimal model architecture and parameter configurations, including the fusion strategy, visual encoder architecture, the depth of 3D scans, and augmentation. The results on the visual encoder architecture can be found in the main body. In the default experiment setting, we set 512x512 as the image size, use 16 as the 3D scan depth, and ResNet-18+ResNet-34 as the visual backbone, without knowledge enhancement and augmentation strategy. While conducting an ablation study on certain component, we keep other settings unchanged.

Image Dimension. Here, we aim to investigate the effect of input resolution, *i.e.*, increasing the depth of input volumes. Due to the constraints of GPU physical memory, we experiment with 16, 24, and 32 as the depths for 3D input volume. We employ trilinear interpolation to resample 3D scans to the same size. Supplementary Table 3 illustrates the performance by varying the depths of 3D images. It can be observed that an increase in depth can bring clear performance gain, suggesting that detailed depth information is critical to perform diagnosis. Consequently, with more slices, the model tends to yield more favorable results.

Augmentation. Here, we evaluate the effectiveness of augmentation on our dataset. Specifically, we adopt four augmentation strategies with 15% probability each, namely, Gaussian Noise, Contrast Adjustment, Affine Variation, and Elastic Deformation. As a result, about half of the training data will be applied at least one augmentation in each training batch. As shown in Supplementary Table 4, adopting data augmentation

Supplementary Table 3 | Ablation study on different 3D scan depth size settings. Considering the computational cost, we evaluated the commonly seen 16, 24, 32 depth sizes on our datasets. The value preceding ‘/’ represents the results from the subset with top 200 classes, while the value following ‘/’ denotes the results from the subset of 200 random classes.

3D Image Depth	AUC	AP	F1	MCC	R@0.01	R@0.05	R@0.1
16	87.06/82.44	11.27/6.52	17.36/10.15	19.23/13.28	21.48/12.33	44.38/30.76	61.54/45.09
24	87.49/83.78	11.41/6.97	17.84/10.48	19.39/13.96	21.76/13.14	45.72/31.89	62.53/46.51
32	88.13/84.23	11.98/7.52	18.02/11.86	19.67/14.49	22.84/15.19	47.46/33.65	63.91/48.07

shows notable performance improvement. This improvement is particularly evident on AUC and AP when employing the ViT as the visual backbone while, still, ResNet-based model performs better.

Supplementary Table 4 | Ablation on adopting augmentation strategy. In the table, we show the effectiveness of adopting reasonable data augmentation for our task. “ResNet” denotes the ResNet-based visual encoder architecture, “ViT” denotes the ViT-based one, and “ResNet-ViT” denotes the ResNet-ViT mixing one. The value preceding ‘/’ represents the results from the top subset, while the value following ‘/’ denotes the results from the random subset.

Augmentation	Visual Encoder	AUC	AP	F1	MCC	R@0.01	R@0.05	R@0.1
YES	ViT	84.22/81.26	8.69/6.87	16.36/9.92	17.18/13.43	19.83/13.90	38.70/29.17	54.04/42.66
	ResNet	88.96/84.47	13.92/8.66	22.04/12.80	23.08/15.93	25.75/16.52	51.61/35.18	66.47/52.15
	ResNet-ViT	90.61/87.39	14.87/9.83	24.10/14.77	24.94/18.07	26.68/18.35	53.26/38.21	68.57/54.61
NO	ViT	81.69/78.11	6.40/5.26	14.71/8.89	15.30/9.74	18.11/11.33	34.73/25.97	48.84/36.20
	ResNet	87.06/82.09	11.27/6.09	17.36/10.15	19.23/12.00	21.48/13.96	44.38/31.46	61.54/44.11
	ResNet-ViT	88.54/84.23	12.85/7.52	10.31/11.86	22.13/14.49	24.02/15.19	49.37/33.65	63.87/48.07

In conclusion, we find the ResNet-based model with augmentation strategy and unifying the 3D scan depth to 32 are the most suitable settings for our long-tailed case-level multi-modal diagnosis task.

D In-class Analysis

The experimental results on whether to use fusion strategy and knowledge enhancement strategy show that, statistically, models that combine both strategies tend to have a higher average AUC than those that do not, across the head, medium, and tail categories, with some categories showing noticeable increase, while some may remain same or slightly decreasing. Taking certain categories within the head category as examples from Supplementary Figure 1. The application of the fusion strategy do not result in improvements for the Meningioma class. Similarly, the implementation of the Knowledge Enhancement strategy do not lead to improvements for the Renal cell carcinoma class. However, we can observe significant improvements in more classes such as Normal, Tuberculosis, and Lymphoma. Consequently, the average metrics across classes shows a notable enhancement.

E Comparison with Current Diagnosis Datasets

In this section, we provide more detailed comparison with other current public available diagnosis datasets as shown in Supplementary Table 5. From the table, we can see our proposed dataset RP3D-DiagDS is the first diagnosis dataset considering whole body regions across different imaging modalities and cover a wide range of diseases.

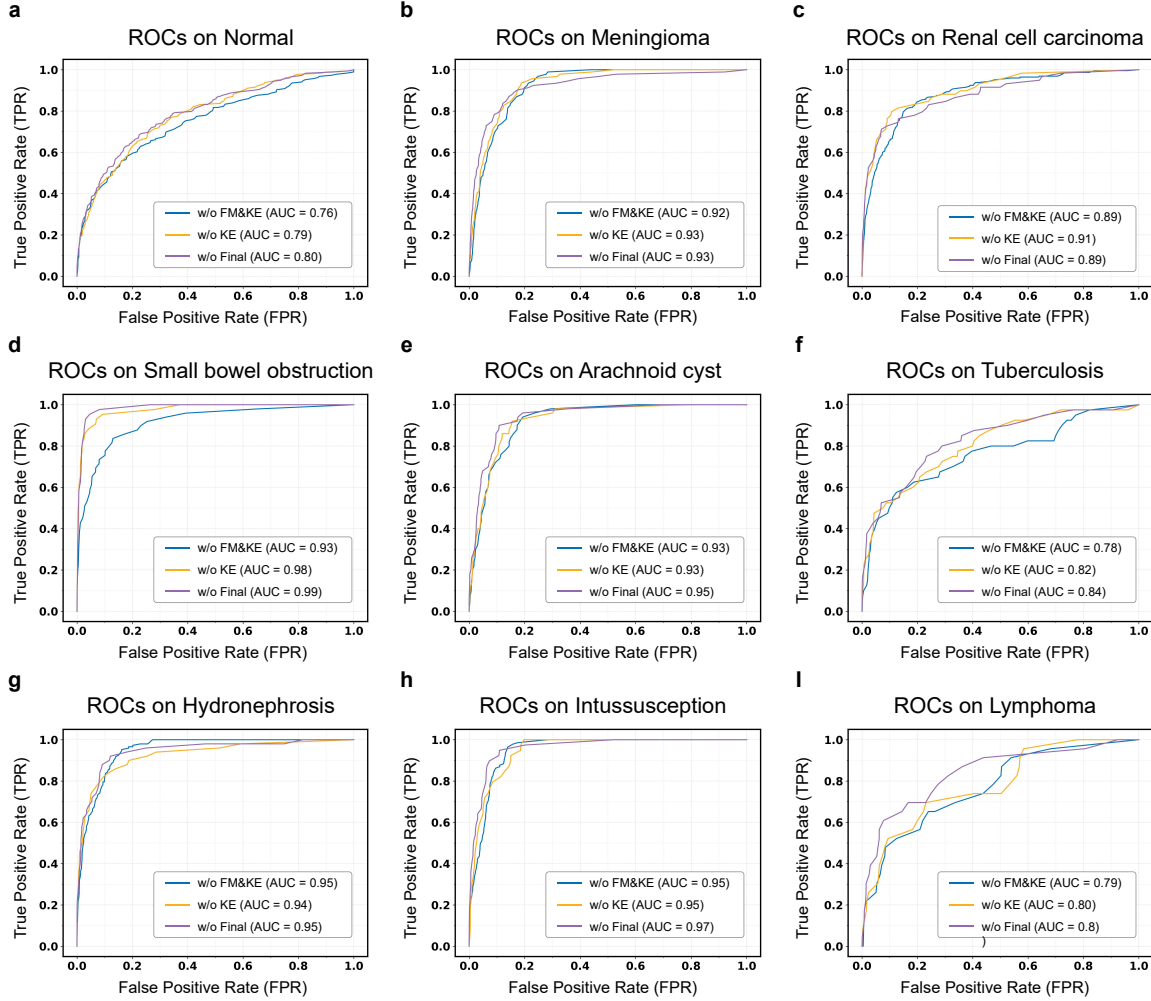
F More ROC Curves

In this section, we show more ROC curves for the classes in the head class category as shown in Supplementary Figure 1

Supplementary Table 5 | Comparison between RP3D-DiagDS and existing disease diagnosis datasets. #Image refers to the number of images in the dataset. #Disorders represents the number of categories of labeled disorders. 2D/3D indicates the dimension of images in the dataset, and the case-level column indicates whether the dataset contains multiple images per patient. Our dataset demonstrates a significant increase in both size and diversity.

Dataset	Anatomy	Modality	#Image	#Disorders	2D	3D	Case-level
ADNI [2]	Head and Neck	MRI	112k	4	✗	✓	✗
Brain-Tumor [3]	Head and Neck	MRI	7k	4	✓	✗	✗
CE-MRI [4]	Head and Neck	MRI	3k	3	✓	✗	✗
Brain-Tumor-17 [5]	Head and Neck	MRI	4.4k	17	✓	✗	✗
OASIS[6–8]	Head and Neck	MRI	3.2k	2	✓	✗	✗
ICH2020 [9]	Head and Neck	CT	82	5	✗	✓	✗
COVID19-CT-DB [10]	Chest	CT	1k	9	✓	✗	✗
MosMedData [11]	Chest	CT	1.1k	5	✗	✓	✗
LNDb [12]	Chest	CT	237	3	✗	✓	✗
NSCLC [13]	Chest	CT, Nuclear Medicine	211	4	✗	✓	✗
Covid-CXR2 [14]	Chest	X-ray	19k	3	✓	✗	✗
POCUS [15]	Chest	Ultrasound	1.1k	3	✓	✗	✗
MIA-COV19 [16]	Chest	CT	5.3k	2	✗	✓	✗
CC-CCII [17]	Chest	CT	3.8k	3	✗	✓	✗
RadChest [18]	Chest	CT	3.6k	83	✓	✗	✗
LIDC-IDRI [19]	Chest	CT	1.0k	3	✗	✓	✗
NIH Chestxray [20]	Chest	X-ray	100k	14	✓	✗	✗
PadChest [21]	Chest	X-ray	160k	193	✓	✗	✗
CheXpert [22]	Chest	X-ray	224k	14	✓	✗	✗
RSNA [23]	Chest	X-ray	30k	2	✓	✗	✗
SIIM-ACR [24]	Chest	X-ray	15k	2	✓	✗	✗
MIMIC-CXR [25]	Chest	X-ray	371k	14	✓	✗	✗
VinDr-CXR [26]	Chest	X-ray	18k	28	✓	✗	✗
VinDr-PCXR [27]	Chest	X-ray	9k	51	✓	✗	✗
ChestX-Det10 [28]	Chest	X-ray	3.5k	10	✓	✗	✗
COVID-19-Radio [29, 30]	Chest	X-ray	21k	4	✓	✗	✗
IU-Xray [31]	Chest	X-ray	7.4k	2	✓	✗	✗
ISPY1 [32]	Breast	MRI	755	5	✗	✓	✗
VinDr-Mammo [33]	Breast	Mammography	20k	10	✓	✗	✗
DDSM [34, 35]	Breast	Mammography	55k	5	✓	✗	✗
CMMD2022 [36]	Breast	Mammography	5.2k	2	✓	✗	✗
BUSI [37]	Breast	Ultrasound	780	3	✓	✗	✗
CT-Kidney [38]	Abdomen and Pelvis	CT	12k	4	✓	✗	✗
VinDr-SpineXr [39]	Spine	X-ray	10k	13	✓	✗	✗
Vertebrate-Xray [40]	Spine	X-ray	388	3	✓	✗	✗
MRNet [41]	Limb	MRI	1370	3	✓	✗	✗
MURA [42]	Limb	X-ray	15k	2	✓	✗	✗
KneeMRI [43]	Limb	X-ray	917	3	✓	✗	✗
DeepLesion [44]	Whole body	Radiology	32k	22	✓	✗	✗
RadImageNet [45]	Whole body	CT, MRI, Ultrasound	1.35M	165	✓	✗	✗
RP3D-DiagDS (ours)	Whole body	Radiology	192k	5568 (ICD-10-CM:930)*	✓	✓	✓

* We map disorders to ICD-10-CM for standardization.



Supplementary Figure 1 | ROC curves on nine classes in "head" category of disorders. FM, KE is short for Fusion Module and Knowledge Enhancement. Source data are provided as a Source Data file.

G More Probability Distributions, Cases Examples, and Saliency Map Visualization.

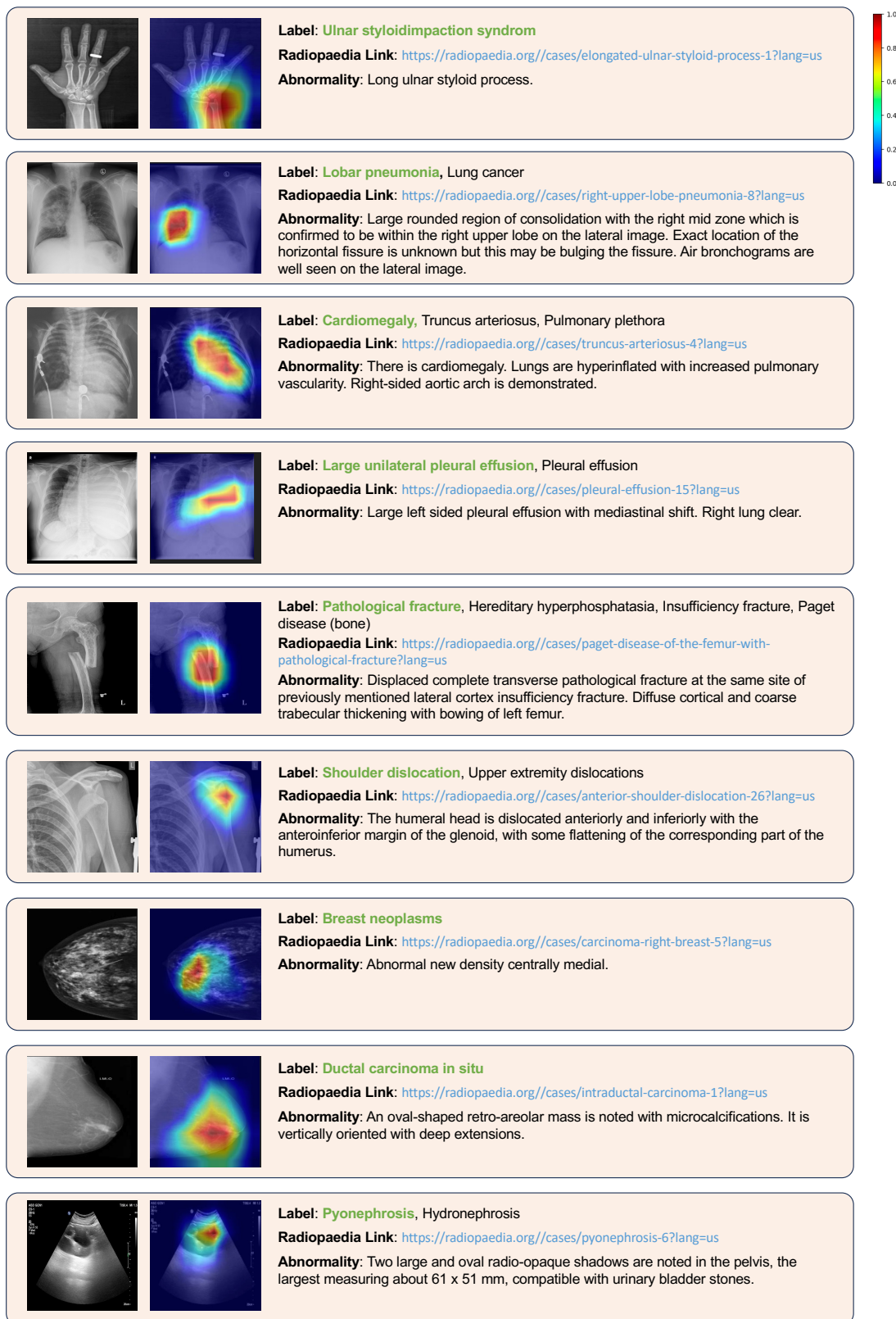
To provide a more comprehensive and clear demonstration of our model's performance, we included more detailed prediction probability distributions, cases, and visualization in the supplementary. First, for each disorder listed in Supplementary Table 6, we plot its corresponding probability distribution in Supplementary Figures 5, 6, 7, 8 and 9. Then in Supplementary Figure 2, 3 and 4, we show more cases and corresponding saliency maps. Each case block is attached with a link to the original website on Radiopaedia, along with all corresponding diagnostic labels and abnormality descriptions. During the diagnostic process, our model considers all images linked in the provided Radiopaedia references, while we only display keyframes for each case here.

Supplementary Table 6 | Modality distributions on some typical disorders. These disorders represent various body systems. The "Modality Distribution" columns show the proportion of imaging modalities used for each disorder in RP3D-DiagDS. As a single case may contain images from multiple modalities, the total count is the sum of images across all modalities for all cases. For each disorder, we present the overall proportions of five primary modalities: CT, MRI, X-ray, Ultrasound, and Mammography, with remaining minor modalities grouped under 'Others'.

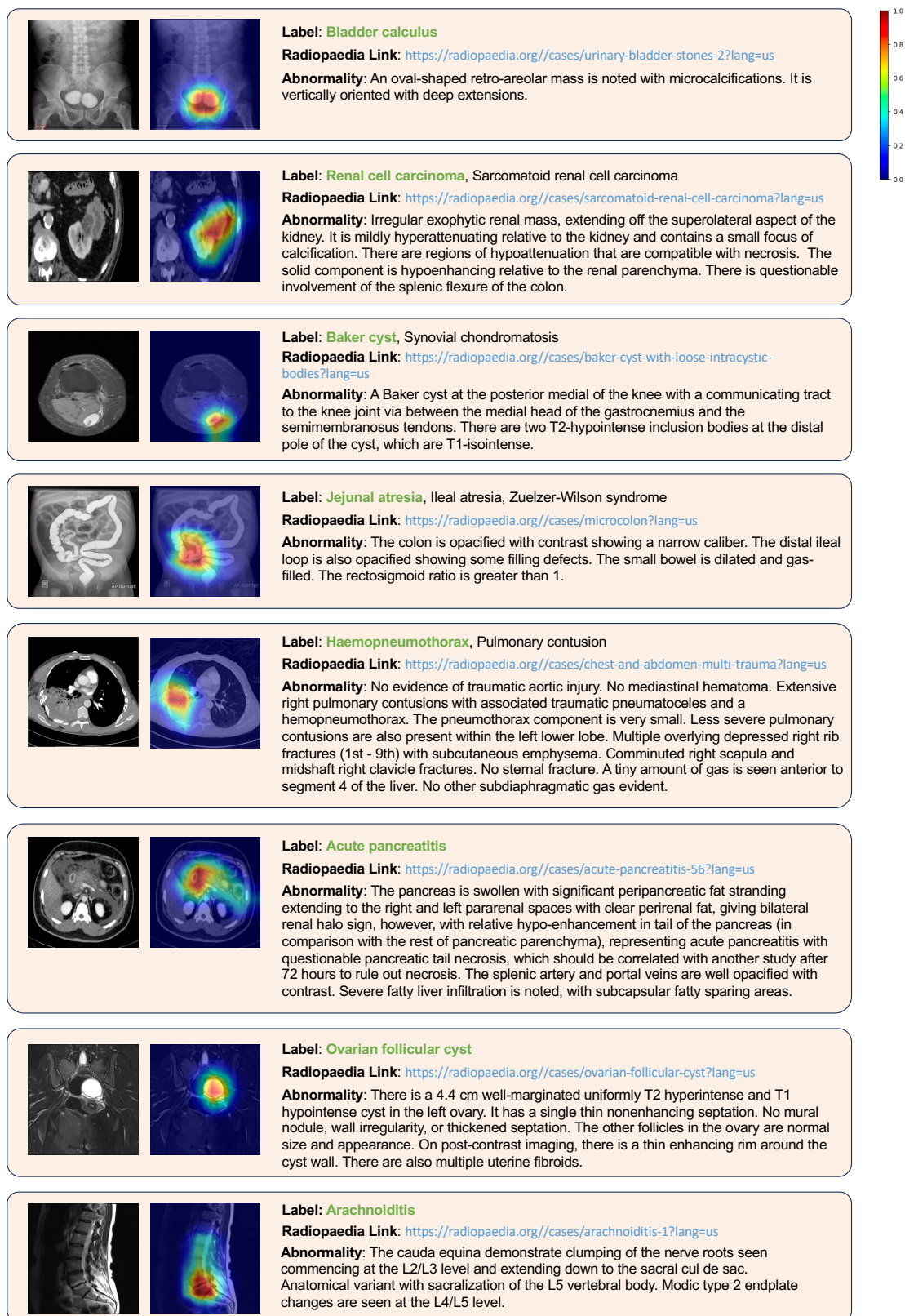
Anatomy	Disorder	Modality Distribution (%)					
		CT	MRI	X-ray	US	MG	Others
Head and Neck	Intracerebral Hemorrhage	82.76	12.93	0.43	-	-	3.88
	Subarachnoid Hemorrhage	72.22	10.42	2.08	-	-	15.28
	Obstructive Hydrocephalus	27.40	65.87	-	1.44	-	5.29
	Craniosynostosis	74.16	8.99	15.73	1.12	-	-
	Mesial Temporal Sclerosis	1.89	92.45	-	-	-	5.66
	Meningioma	27.04	68.87	0.99	0.42	-	2.68
	Papillary thyroid cancer	34.57	6.17	8.64	37.04	-	13.58
	Medulloblastoma	26.52	70.45	0.76	-	-	2.27
Chest	Pneumothorax	27.14	0.56	68.22	0.19	-	3.89
	Cardiomegaly	28.43	3.92	59.80	4.90	-	2.95
	Pulmonary Emphysema	52.69	-	41.94	-	-	5.37
	Pulmonary Fibrosis	54.00	-	44.00	-	-	2.00
	Pulmonary Embolism	66.07	1.43	16.07	4.64	-	11.79
	Usual Interstitial Pneumonia	64.37	-	29.89	-	-	5.74
	Coarctation of the Aorta	51.61	8.60	26.88	-	-	12.91
	Ventricular Septal Defect	47.56	-	31.71	13.41	-	7.32
	Pericardial Effusion	44.80	4.00	42.40	4.00	-	4.80
Abdomen	Cholelithiasis	27.92	43.15	2.03	12.69	-	14.21
	Hepatic Hydatid Infection	62.00	16.67	4.00	14.67	-	2.66
	Hydronephrosis	57.62	5.20	5.57	17.10	-	14.51
	Uterine Leiomyoma	21.83	51.78	7.11	13.20	-	6.08
	Testicular Torsion	-	2.90	-	94.20	-	2.90
	Small Bowel Obstruction	60.32	1.36	28.53	2.99	-	6.80
	Rectal Cancer	53.25	36.36	2.60	-	-	7.79
	Acute Pancreatitis	75.74	13.02	3.55	3.55	-	4.14
	Cirrhosis	64.08	16.02	1.46	13.59	-	4.85
Limb	Torus Fracture	1.12	2.25	94.38	-	-	2.25
	Metacarpal Fracture	1.32	2.63	96.05	-	-	-
	Tibial Plateau Fracture	37.80	7.32	47.56	1.22	-	6.1
	Shoulder Dislocation	5.26	11.40	76.32	2.63	-	4.39
	Elbow Joint Effusion	9.38	4.69	78.75	-	-	7.18
	Calcific Tendinitis	10.83	26.11	40.76	19.11	-	3.19
	Meniscal Tear	-	90.91	4.96	1.65	-	2.48
Spine	Cervical Spine Fractures	59.77	20.69	12.64	-	-	6.90
	Intramedullary Spinal Tumors	1.69	86.10	5.08	-	-	7.13
	Disc Herniation	17.65	69.12	1.47	1.47	-	10.29
	Spondylodiscitis	18.13	63.75	11.88	-	-	6.24
	Spondylolysis	34.18	20.25	37.97	1.27	-	6.33
Breast	Ductal Carcinoma in Situ	3.10	4.65	1.55	20.93	62.79	6.98
	Fibroadenoma (Breast)	-	8.20	1.64	39.34	49.18	1.64
	Breast Hamartoma	2.27	-	-	22.73	68.18	6.82
	Breast Neoplasms	21.64	13.15	6.85	23.01	30.41	4.94



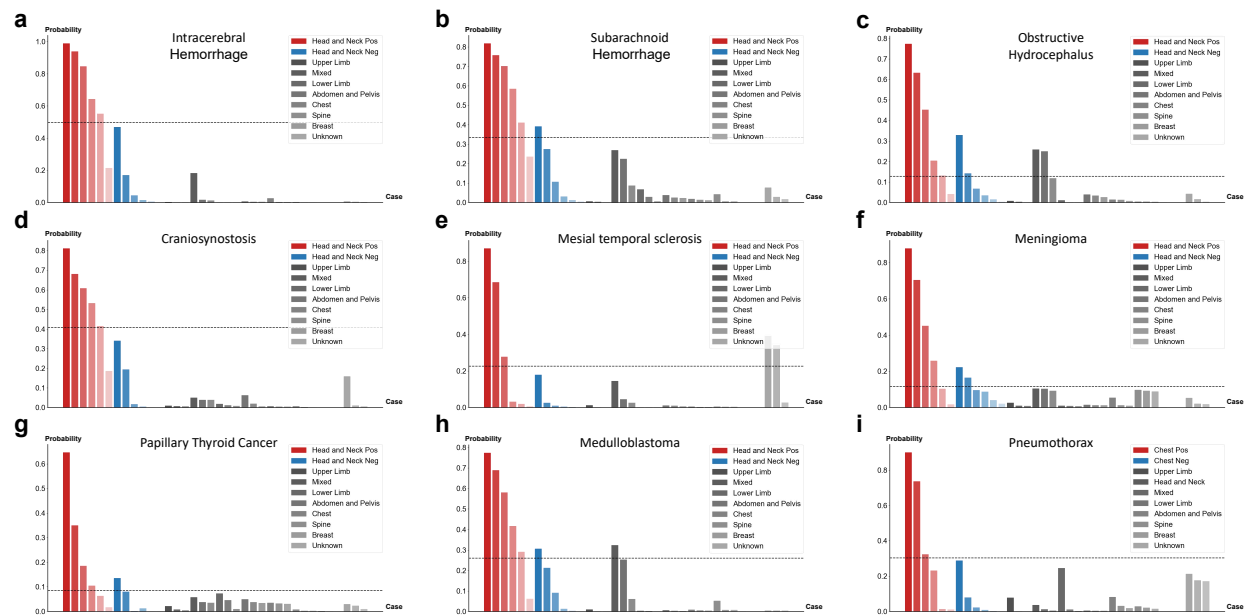
Supplementary Figure 2 | Saliency map of the key frames. Red indicates the areas that our model focuses on when inferring the corresponding disease category. Each case is attached with the corresponding disease labels, where the **green bold one** indicating the guiding label for the explainable image and others represent all diagnoses for the case, potentially from multiple modalities. The "Radiopaedia link" provides the source website. "Abnormality" refers to the anomalies in this case (which may include multiple 2D and 3D images, with only keyframes displayed here).



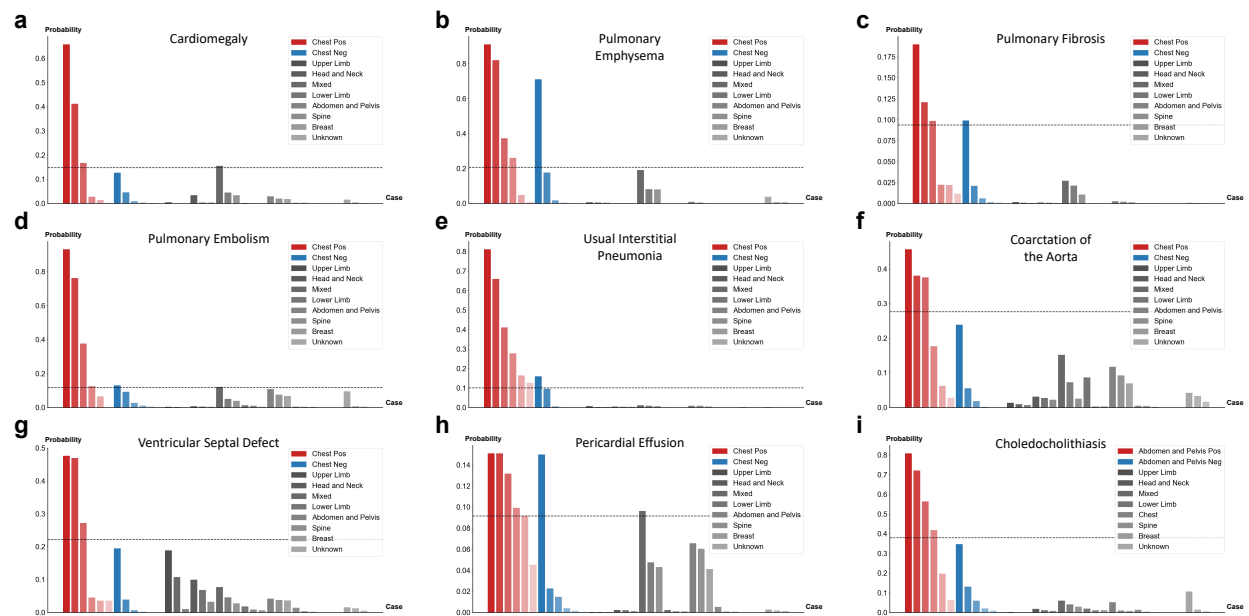
Supplementary Figure 3 | Saliency Map of the Key Frames. (Cont.)



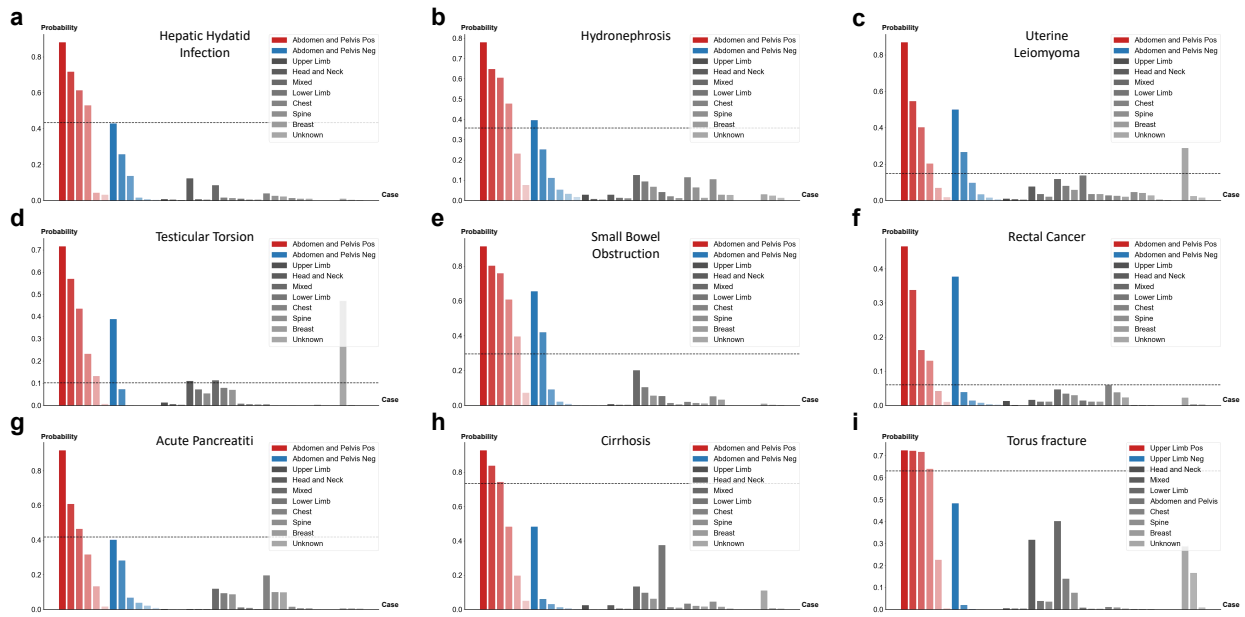
Supplementary Figure 4 | Saliency Map of the Key Frames. (Cont.)



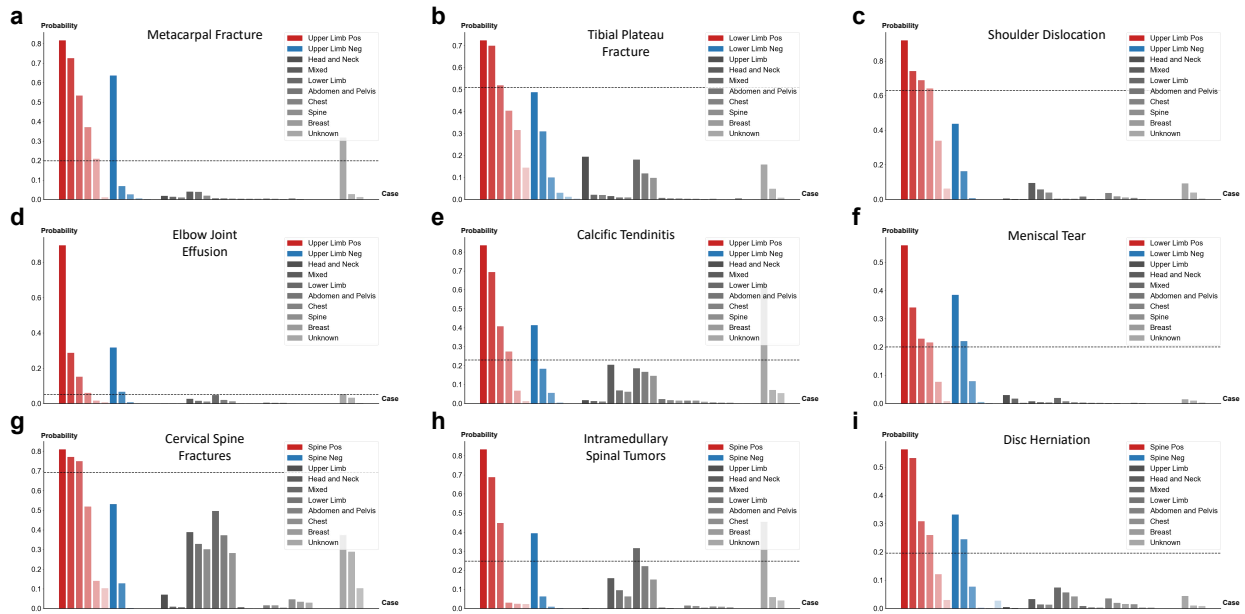
Supplementary Figure 5 | Prediction probability distribution on some representative disorders. Source data are provided as a Source Data file.



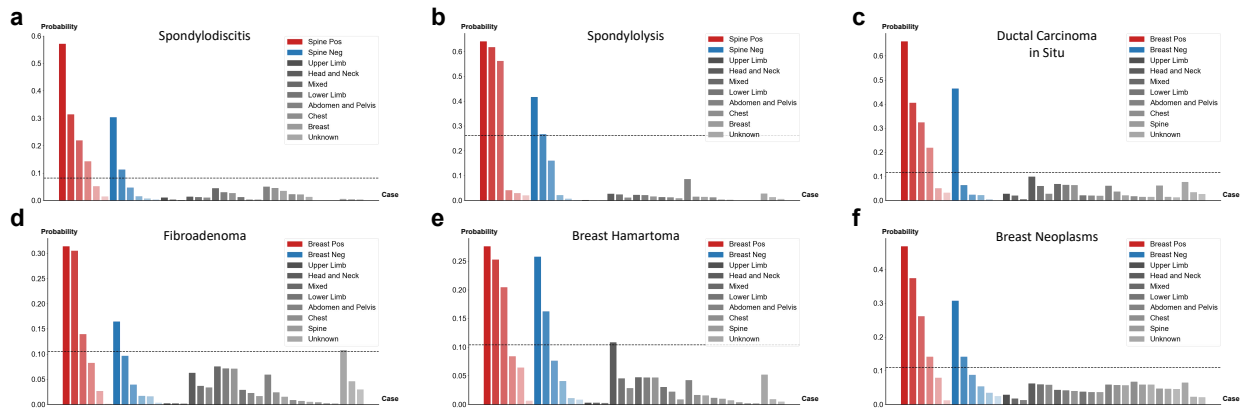
Supplementary Figure 6 | Prediction probability distribution on some representative disorders. Source data are provided as a Source Data file.



Supplementary Figure 7 | Prediction probability distribution on some representative disorders. Source data are provided as a Source Data file.



Supplementary Figure 8 | Prediction probability distribution on some representative disorders. Source data are provided as a Source Data file.



Supplementary Figure 9 | Prediction probability distribution on some representative disorders. Source data are provided as a Source Data file.

References

1. Cardoso, M. J. *et al.* Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701* (2022).
2. Jack Jr, C. R. *et al.* The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **27**, 685–691 (2008).
3. Nickparvar, M. *Brain Tumor MRI Dataset* 2021. <https://www.kaggle.com/dsv/2645886>.
4. Cheng, J. *et al.* Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS One* **10**, e0140381 (2015).
5. Feltrin, F. *Brain Tumor MRI Images 17 Classes* 2022. <https://www.kaggle.com/datasets/fernando2rad/brain-tumor-mri-images-17-classes>.
6. Koenig, L. N. *et al.* Select Atrophied Regions in Alzheimer disease (SARA): An improved volumetric model for identifying Alzheimer disease dementia. *NeuroImage: Clinical* **26**, 102248. ISSN: 2213-1582 (2020).
7. LaMontagne, P. J. *et al.* OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medrxiv*, 2019–12 (2019).
8. Marcus, D. S. *et al.* Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *Journal of Cognitive Neuroscience* **19**, 1498–1507. ISSN: 0898-929X (Sept. 2007).
9. Hssayeni, M. D. *et al.* Intracranial hemorrhage segmentation using a deep convolutional model. *Data* **5**, 14 (2020).
10. Shakouri, S. *et al.* COVID19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed COVID-19 diagnosis. *BMC Research Notes* **14**, 178 (2021).
11. Morozov, S. P. *et al.* Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465* (2020).
12. Pedrosa, J. *et al.* LNDdb: a lung nodule database on computed tomography. *arXiv preprint arXiv:1911.08434* (2019).
13. Bakr, S. *et al.* A radiogenomic dataset of non-small cell lung cancer. *Scientific Data* **5**, 1–9 (2018).
14. Pavlova, M. *et al.* COVID-Net cxr-2: An enhanced deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Frontiers in Medicine* **9** (2022).
15. Born, J. *et al.* POCOVID-Net: automatic detection of COVID-19 from a new lung ultrasound imaging dataset (POCUS). *arXiv preprint arXiv:2004.12084* (2020).
16. Kollias, D., Arsenos, A., Soukissian, L. & Kollias, S. *Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 537–544.
17. Zhang, K. *et al.* Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cell* **181**, 1423–1433 (2020).
18. Draelos, R. L. *et al.* Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical image analysis* **67**, 101857 (2021).
19. Armato III, S. G. *et al.* The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical Physics* **38**, 915–931 (2011).
20. Wang, X. *et al.* *Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 2097–2106.
21. Bustos, A., Pertusa, A., Salinas, J.-M. & de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* **66**, 101797 (2020).

22. 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* (2019), 590–597.
23. Shih, G., Wu, C. C., Halabi, S. S., Kohli, M. D., Prevedello, L. M., *et al.* Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence* **1**, e180041 (2019).
24. Filice, R. W. *et al.* Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset. *Journal of Digital Imaging* **33**, 490–496 (2020).
25. Johnson, A. E. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* **6**, 317 (2019).
26. Nguyen, H. Q. *et al.* VinDr-CXR: An open dataset of chest X-rays with radiologist’s annotations. *Scientific Data* **9**, 429 (2022).
27. Nguyen, N. H., Pham, H. H., Tran, T. T., Nguyen, T. N. & Nguyen, H. Q. VinDr-PCXR: An open, large-scale chest radiograph dataset for interpretation of common thoracic diseases in children. *medRxiv*, 2022–03 (2022).
28. Liu, J., Lian, J. & Yu, Y. ChestX-Det10: Chest X-ray Dataset on Detection of Thoracic Abnormalities. *arXiv preprint arXiv:2006.10550v3* (2020).
29. Chowdhury, M. E. *et al.* Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access* **8**, 132665–132676 (2020).
30. Rahman, T. *et al.* Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in biology and medicine* **132**, 104319 (2021).
31. Pavlopoulos, J., Kougia, V. & Androutsopoulos, I. *A survey on biomedical image captioning* in *Proceedings of the second workshop on shortcomings in vision and language* (2019), 26–36.
32. Hylton, N. M. *et al.* Neoadjuvant chemotherapy for breast cancer: functional tumor volume by MR imaging predicts recurrence-free survival—results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL. *Radiology* **279**, 44–55 (2016).
33. Nguyen, H. T. *et al.* VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *PhysioNet* (2022).
34. Heath, Michael and Bowyer, Kevin and Kopans, Daniel, and Moore, Richard and Kegelmeyer, W. Philip. *The Digital Database for Screening Mammography* in *Proceedings of the Fifth International Workshop on Digital Mammography* (Medical Physics Publishing, 2001), 212–218.
35. Lee, R. S. *et al.* A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data* **4**, 1–9 (2017).
36. Cai, H. *et al.* Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms. *Computational and Mathematical Methods in Medicine* **2019** (2019).
37. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data In Brief* **28**, 104863 (2020).
38. Islam, M. N. *et al.* Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. *Scientific Reports* **12**, 1–14 (2022).
39. Nguyen, H. T. *et al.* *VinDr-SpineXR: A deep learning framework for spinal lesions detection and classification from radiographs* in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24* (2021), 291–301.
40. Fraiwan, M., Audat, Z., Fraiwan, L. & Manasreh, T. Using deep transfer learning to detect scoliosis and spondylolisthesis from X-ray images. *Plos one* **17**, e0267851 (2022).
41. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Medicine* **15** (2018).
42. Rajpurkar, P. *et al.* Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957* (2017).

- 43. Štajduhar, I., Mamula, M., Miletić, D. & Uenal, G. Semi-automated detection of anterior cruciate ligament injury from MRI. *Computer methods and programs in biomedicine* **140**, 151–164 (2017).
- 44. Yan, K., Wang, X., Lu, L. & Summers, R. M. Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations. *arXiv preprint arXiv:1710.01766* (2017).
- 45. Mei, X. *et al.* RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence* **4**, e210315 (2022).