

Supplementary

1 Dataset statistics and annotations

We collected knee joint data from 5 hospitals (centers A–E) in different cities. We initially collected from center A, center B, and center C–E, respectively. We excluded a total of 248 patients (51, 140, and 57 patients from center A, center B, and center C–E, respectively) who underwent surgery before the MRI and 95 patients (76, 14, and 5 patients from center A, center B, and center C–E, respectively) who had poor MR image quality due to artifact.

All patients’ MR images were exported from the picture archiving and communication system (PACS) workstation and saved in Digital Imaging and Communications in Medicine (DICOM) format. As shown in Tab. 1, PDW sequences in three planes are collected in all centers. For center D, an additional sagittal T1W is collected and further interpolated to coronal T1W. For center E, an additional sagittal T2W sequence is collected. Details of knee MRI protocol are demonstrated in Tab. 2 and Tab. 3.

Table 1: Sequence sets in each center

	Center A	Center B	Center C	Center D	Center E
Sagittal PDW	✓	✓	✓	✓	✓
Coronal PDW	✓	✓	✓	✓	✓
Axial PDW	✓	✓	✓	✓	✓
Sagittal T2W	✓				✓
Coronal T1W	✓			✓*	

* Converted from sagittal T1W sequence

The datasets were annotated by two senior musculoskeletal radiologists (S.S.L. and Q.Y., with 10 and 16 years of experience respectively). They dichotomized each of the 12 types of knee abnormalities as follows: (a) meniscal tear or not; (b) cartilage damage or not; (c) anterior cruciate ligament tear or not; (d) posterior cruciate ligament injury or not; (e) medial collateral ligament injury or not; (f) lateral collateral ligament injury or not; (g) patellar retinaculum injury or not; (h) infrapatellar fat pad injury or not; (i) synovial plica or not; (j) cyst (including popliteal cyst, ganglion cyst, meniscus cyst,

and synovial cyst) or not; (k) bone contusion or not; and (l) joint effusion or not.

First, MR images were interpreted by radiologists, who were blinded to the knee arthroscopy information in this step. For controversial cases, two senior radiologists will discuss and their consensus was used as the final annotations. Then, radiologists review the knee arthroscopies performed by orthopedists. Arthroscopy was the gold standard for annotating meniscal tears, cartilage damage, and anterior cruciate ligament tear. For the other nine types, the MRI consensus of two senior radiologists was used as the gold standard. Positive lesions were scored as 1, and negative lesions were scored as 0.

Patient characteristics of the dataset are summarized in Tab. 4, and Tab. 5 shows the distribution of each type of knee abnormality in our dataset. Considering that the limited data may affect the reliability of the evaluation, we merged center B, C, and center D, E based on the availability of their sequences.

Table 2: MRI protocols. TR = repetition time. TE = echo time. FOV = field of view. ms = millisecond. mm = millimeter. PDwFS = proton density-weighted with fat suppression.

Center	Scanner	Sequence	TR/TE (ms)	Slice thickness (mm)	Gap (mm)	FOV (mm)	Matrix size
A	Philips 1.5T (Achieva)	Axial PDwFS	3194/30	4.0	0.4	100	336×264
		Coronal PDwFS	3000/30	3.0	0.4	113	384×224
		Sagittal PDwFS	3031/30	3.3	0.4	113	384×192
		Coronal T1WI	500/17	3.0	0.4	89	384×224
		Sagittal T2WI	2800/35	3.3	0.4	112	384×192
	Philips 3.0T (Ingenia)	Axial PDwFS	2780/65	4.0	0.3	100	256×224
		Coronal PDwFS	3731/30	3.5	0.3	90	334×224
		Sagittal PDwFS	2845/30	3.5	0.3	100	334×192
		Coronal T1WI	582/20	3.5	0.3	90	334×224
B	Philips 1.5T (Achieva)	Sagittal T2WI	2406/100	3.5	0.3	100	334×192
		Axial PDwFS	1639/60	3.0	0.7	160	256×224
		Coronal PDwFS	3000/30	3.0	0.6	180	334×224
		Sagittal PDwFS	3029/60	3.0	0.6	180	334×192
C	Philips 3.0T (Elition)	Axial PDwFS	2577/45	5.0	0.3	160	256×224
		Coronal PDwFS	2345/45	4.0	0.3	180	334×224
		Sagittal PDwFS	2345/25	4.0	0.3	180	334×192

Continued on next page

Table 3: continued from previous page

Center	Scanner	Sequence	TR/TE (ms)	Slice thickness (mm)	Gap (mm)	FOV (mm)	Matrix size
D	GE 3.0T	Axial PDwFS	2685/30	4.0	0.6	160	288×256
		Coronal PDwFS	2740/30	3.5	0.4	160	320×256
		Sagittal PDwFS	3138/30	3.5	0.4	180	320×256
		Sagittal T1WI	732/14	3.5	0.4	180	288×192
	SIEMENS 3.0T	Axial PDwFS	4200/38	3.0	0.3	160	256×205
		Coronal PDwFS	4190/40	3.0	0.3	180	320×224
		Sagittal PDwFS	3390/38	3.0	0.3	220	320×224
		Sagittal T1WI	500/10	3.0	0.3	180	320×224
	SIEMENS 1.5T	Axial PDwFS	4000/75	3.5	0.6	160	320×256
		Coronal PDwFS	4150/30	3.0	0.6	180	320×224
		Sagittal PDwFS	3310/32	3.0	0.6	180	320×224
		Sagittal T1WI	876/12	3.0	0.6	180	256×230
E	GE 3.0T	Axial PDwFS	2500/44	4.5	0.4	160	335×266
		Coronal PDwFS	2500/42	4.0	0.4	160	330×226
		Sagittal PDwFS	2500/43	3.5	0.4	180	330×244
		Sagittal T2WI	2500/73	3.5	0.4	180	352×236

Table 4: Patient characteristics. The Kruskal–Wallis test was used to compare the ages of patients in different centers, and the chi-square test was used to compare sex. A two-sided p value less than 0.05 was considered statistically significant. Data were analyzed using R statistical software (version 4.3.2).

	Center A	Center B	Center C	Center D	Center E	P Value
No. of patients	1,100	499	36	64	44	
No. of examinations	1,103	500	37	64	44	
Age						
Mean $\pm$ SD	39.42 $\pm$ 15.55	40.56 $\pm$ 16.14	52.83 $\pm$ 15.43	39.84 $\pm$ 15.07	32.68 $\pm$ 17.78	5.917×10^{-8}
Median [interquartile (%)]	35 (26, 50)	41 (27, 52)	56 (48.25, 65.50)	37 (27, 49)	26 (20, 51.25)	
Sex, n (%)						1.334×10^{-7}
	699 (63.5)	249 (49.9)	14 (38.9)	44 (68.8)	21 (47.7)	
	401 (36.5)	250 (50.1)	22 (61.1)	20 (31.3)	23 (52.3)	

Table 5: Class distribution in our dataset: The internal dataset (center A) is divided into training, validation, and testing subsets, while the external datasets (centers B-E) are only used as testing sets

		Center A			Center B	Center C	Center D	Center E
	Train	Val	Test	Total	Test	Test	Test	Test
Patient	773	110	220	1103	500	37	64	44
MENI	563	73	164	800	180	25	49	15
ACL	278	50	126	454	86	8	27	4
CART	319	25	47	391	115	23	25	16
PCL	114	16	72	202	62	9	13	5
MCL	148	24	85	257	93	10	13	7
LCL	114	18	69	201	128	7	9	4
EFFU	703	89	182	974	498	37	57	23
CONT	287	38	107	432	219	19	14	8
PLICA	488	59	59	606	278	8	0	14
CYST	146	17	16	179	103	11	13	9
IFP	305	40	75	420	257	13	0	9
PR	80	14	60	154	81	2	5	7

2 Experiment setup

2.1 Data pre-processing and parameter setting

In this study, the images were acquired from various centers, resulting in a variation in image quality. By considering a statistically significant P-value of ≤ 0.05 , we identified differences in scanning parameters, image spacing, and intensity distribution. To reduce the influence on model generalization, we perform the following steps for image pre-processing: 1. Resampling of isotropic voxels: all images are re-sampled to a spacing of $(0.28 \times 0.28 \times 3mm)$ and scale the images into 576×576 whilst keeping the number of slices unchanged; 2. Intensity standardization: we apply zero-mean normalization with its mean and standard deviation. 3. Knee area alignment: the images were cropped in an area of 448×448 according to the ROI segmented from the meniscus (a rough segmentation performed by U-Net as an upstream task) before being interpolated into the shape of 224×224 .

For data augmentation, we perform crop and translate within 10 pixels, and rotate randomly across the slice plane within $\pm 25^\circ$. Finally, we adjust the contrast and brightness within a range of 1.5 times. Notably, in counter with the imbalance of positive and negative samples, we performed resample in some classes during training. The dataset was augmented by a factor of 4, resulting in a total of 3092 samples.

The experiments are conducted on a platform with Nvidia 3090. All the competing methods are trained using Adam optimizer with an initial learning rate of $1e-5$, and run 500 epochs with an early-stop strategy. In our experiment, we set the hyper-parameters empirically with $\alpha = 0.3$. We employ supervision to all predictions (i.e., predictions from three branches and the final prediction). The encoders that are pre-trained with relevant tasks [1] serve as the feature extractors to boost our training.

2.2 Adapting existing methods for comparison

We adapt the existing methods to our task for comparisons while keeping their functional network structures. To be specific, we expand the model into twelve individual networks for the methods that classify each abnormality separately. Besides, some of the existing methods are specially tailored for certain MRI sequences, in which we can only make a comparison with available sequences. First, MRNet [2] is set as the baseline, to preserve the simple-task structure, we build twelve individual models corresponding to each abnormality. We also compare our model to MPFuseNet [3] for validating our aggregation for three branches, the final layer of this network is expanded to fit twelve classes. The ELNet [4] takes a single sequence as input, so we choose to fit one model to each sequence and expend for twelve abnormalities. For a fair comparison, the models are trained with the same experiment settings (e.g., dataset, training parameters, hard devices). The input images are all cropped with ROI from meniscus segmentation.

3 Performance Evaluation

3.1 Patient-level evaluation

We further evaluate the patient-level performance of our model compared to the radiologists' in the internal dataset. We report the conventional multi-label classification metrics: mean average precision (mAP), category precision (CP), overall precision (OP), category recall rate (CR), overall recall rate (OR), category F1 score (CF1) and overall F1 score (OF1). The result is shown in Tab. 6.

We can see that even though our model performs better in all metrics than the junior radiologist, there is still a gap with the senior radiologist when considering the patient-level diagnosis. Given that our model performs closely to the senior radiologist in single-class accuracy, we assume that the senior radiologists have their own empirical prior knowledge of the diagnosis. Indeed, our probability matrix and multi-task structure enable the model to learn the plane-abnormality and inter-abnormality correlation positively, but it seems they are less likely to be found in limited training data. However, this knowledge is vital for AI models in making decisions, and may significantly affect the reliability during clinical implementation.

Table 6: Patient-level evaluation and comparison to radiologists. The conventional multi-label classification metrics are reported.

	mAP	CP	OP	CR	OR	CF1	OF1
CoPAS	0.622	0.651	0.702	0.692	0.671	0.610	0.686
Junior Radiologist	0.497	0.558	0.560	0.659	0.649	0.556	0.601
Senior Radiologist	0.638	0.710	0.754	0.765	0.762	0.715	0.758

*Highest scores are indicated in bold text.

3.2 Assisting radiologists with model prediction

To evaluate the capability of our model to provide diagnostic assistance in clinical settings, we involve 6 radiologists (2 senior and 4 junior) in the comparison experiments in the external dataset. We select 220 patients which keep the same distribution of positive samples as the internal testing set. The radiologists are asked to perform the diagnosis twice: first alone, then do it again with the predictions from our model as reference. The detail is described in the previous section. As shown in Fig. 1, we plot the ROC curve of our model's prediction and the (false positive rate (FPR), and true positive rate (TPR)) point sets of radiologists'.

The results of our model still surpass the junior radiologists in most cases. However, different from our observations on the internal dataset, the senior radiologist overwhelms our model and other radiologists with a high TPR and

low FPR. We assume this drawback of our model is caused by cross-center data shifting and the senior radiologist who has over 30 years of experience is less affected. The senior performs closely to others in CART and PLICA, which seems to be challenging for both humans and AI.

Notably, the accuracy of radiologists have a significant improvement when assisted by our model (with $p = 8.1 \times 10^{-11}$ by one-tail paired T-test), especially in MENI, CART, PLIA, CYST, and IFP. It also shows that in PCL, MCL, and LCL, the radiologist is certain of his decisions that the model's prediction has subtle effects, and thus has a high diagnostic accuracy. Additionally, we can see the performance of junior radiologists getting close to the seniors' in MENI and CART after the AI interferes, which shows our model can assist the junior radiologists to reach better performance even comparable to the senior ones.

3.3 Performance evaluation in different subgroups

To evaluate the influence of bias resulting from different patient demographics, we divided the dataset (both internal and external) into several subgroups based on sex and age, as shown in Fig. 2. Specifically, we created male and female subgroups based on sex. Four age subgroups (Age0-Age3) are created according to median and interquartile range (shown in Tab. 4). With the average area under ROC curve (AUC) reported in Fig. 2, We can observe that the variation in sex and age has had a significant impact on the performance. The imbalanced sex subgroups have resulted in a performance drop from 0.80 to 0.76 in the internal dataset (center A) and have had a considerable influence on the performance in center E. Due to the limited number of data in center E, the performance variance is even greater compared to other centers. Similarly, in the age subgroups, the performance variance is amplified. For instance, center C only has 3 patients corresponding to the Age0 subgroup, resulting in a mean of 0.5 and a standard deviation of 0.3.

3.4 Clinical-model consistency quantification

We quantify the likelihood of accurately distinguishing a specific abnormality in a given plane using a numerical scale ranging from 0 to 1. For example, the sagittal plane is the preferred choice for clinicians to diagnose a meniscus tear (MENI), and we assign the value 0.9 to denote a high likelihood of observation. Conversely, the axial plane is considered the least preferred option, and we assign the value 0.1 to indicate a lower likelihood of observation. The prediction certainty is quantified by calculating the average distance of the output probability from 0.5, calculated by $Certainty = |p - 0.5|$, where a larger number means the model is more confident in this output. The quantification result is shown in Tab. 7, Tab. 8 and Tab. 9.

Table 7: Clinical Observation Probability

	MENI	ACL	CART	PCL	MCL	LCL	EFFU	CONT	PLICA	CYST	IFP	PR
Sagittal	0.90	0.90	0.90	0.90	0.60	0.60	0.90	0.90	0.60	0.90	0.90	0.10
Coronal	0.60	0.60	0.60	0.60	0.90	0.90	0.60	0.60	0.30	0.60	0.30	0.60
Axial	0.10	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.90	0.30	0.60	0.90

Table 8: Model Prediction Accuracy

	MENI	ACL	CART	PCL	MCL	LCL	EFFU	CONT	PLICA	CYST	IFP	PR
Sagittal	0.79	0.89	0.8	0.82	0.68	0.7	0.82	0.83	0.86	0.9	0.7	0.68
Coronal	0.68	0.76	0.72	0.7	0.75	0.82	0.78	0.76	0.82	0.59	0.33	0.67
Axial	0.45	0.88	0.73	0.66	0.62	0.72	0.79	0.64	0.75	0.68	0.52	0.78

Table 9: Model Prediction Certainty

	MENI	ACL	CART	PCL	MCL	LCL	EFFU	CONT	PLICA	CYST	IFP	PR
Sagittal	0.41	0.49	0.48	0.49	0.35	0.43	0.41	0.28	0.23	0.49	0.33	0.43
Coronal	0.26	0.39	0.39	0.35	0.35	0.43	0.44	0.26	0.21	0.37	0.14	0.43
Axial	0.22	0.49	0.27	0.27	0.37	0.40	0.33	0.32	0.24	0.35	0.19	0.34

4 Ablation study

The effectiveness of the proposed modules is verified by ablation studies on the Centre A dataset, the results are illustrated in Tab. 10. The performance gain is mainly attributed to two novel modules that leverage the attention mechanism (the cross-plane attention (CPA) and the cross-sequence attention (CSA)), and the correlation mining strategy from the plane-aware probability matrix. As mentioned before, we chose MRNet as the baseline model for its high performance in knee abnormality classification tasks. Hence, we build four variations of models (iv, v, vi, vii) in the ablation study.

In the first two models (ii and iii), we aim to figure out how meniscus ROI and multi-task may affect performance. Model (ii) is a baseline model with whole images as input. In model (iii), we modify the last FC layer to classify 12 abnormalities with one baseline network. Next, in model (iv), we test the attention modules by adding the cross-plane volumes for each branch in the basis of model (ii). In model (v), we add T1W and T2W sequences and multi-contrast integration modules. Lastly, in model (vi), we modify the output of three branches to form the plane-aware probability matrix and introduce the correlation mining module (CRM). In fact, model (vi) is the final version of our model. Then, we modify CoPAS into single-task structure and duplicate 12 times corresponding to 12 abnormalities, as is shown in (vii).

It is notable that the ROI from segmentation brings a significant improvement with an average AUC-ROC from 0.725 to 0.781. It may be because the abnormality symptoms take a small proportion compared to the whole image,

Table 10: Ablation study in Center A, the area under ROC curve (AUC-ROC) is reported as metric, our proposed modules improve the performance.

Method	MENI	ACL	CART	PCL	MCL	LCL	EFFU	CONT	PLICA	CYST	IFP	PR	Average
Centre A													
(i) MRNet [2] (Baseline)	0.659	0.950	0.778	0.759	0.802	0.763	0.789	0.820	0.759	0.819	0.732	0.744	0.781
(ii) Baseline w/o ROI	0.624	0.867	0.734	0.755	0.792	0.703	0.788	0.779	0.627	0.648	0.727	0.654	0.725
(iii) Baseline (Multitask)	0.630	0.890	0.762	0.747	0.732	0.739	0.724	0.781	0.723	0.750	0.714	0.739	0.744
(iv) CPA	0.696	0.909	0.769	0.765	0.782	0.751	0.787	0.808	0.773	0.768	0.742	0.732	0.774
(v) CPA+CSA	0.740	0.944	0.771	0.777	0.811	0.785	0.759	0.833	0.838	0.817	0.758	0.756	0.799
(vi) CPA+CSA+CRM (CoPAS)	0.763	0.949	0.790	0.800	0.782	0.816	0.787	0.821	0.877	0.862	0.734	0.767	0.812
(vii) CoPAS (single task)	0.745	0.923	0.793	0.792	0.816	0.801	0.789	0.826	0.856	0.833	0.756	0.766	0.808

*Highest scores are indicated in bold text.

which is challenging for the model to learn with limited data. Most importantly, cropping along ROI can align the spatial position across sequences. Besides, there is a performance drop (average AUC-ROC from 0.781 to 0.705) when the baseline was adapted into all twelve abnormality classifications in (iii). At the same time, comparing model (vi) and (vii), statistical analysis reveals significant differences ($P < 0.05$, Delong's test) among 10 out of the 12 classes, with the exception being EFFU and PR. Despite a marginal difference of only 0.004 in average AUC-ROC values, the multi-task model (vi) consistently outperforms the single-task model (vii) across a majority of the tasks. It proves our hypothesis that the model may be misled with similar features in a multi-task setting, especially given a small dataset. Nevertheless, considering the resources required for practical implementation, it may be advantageous to opt for a lighter multi-task structure, thereby making a trade-off.

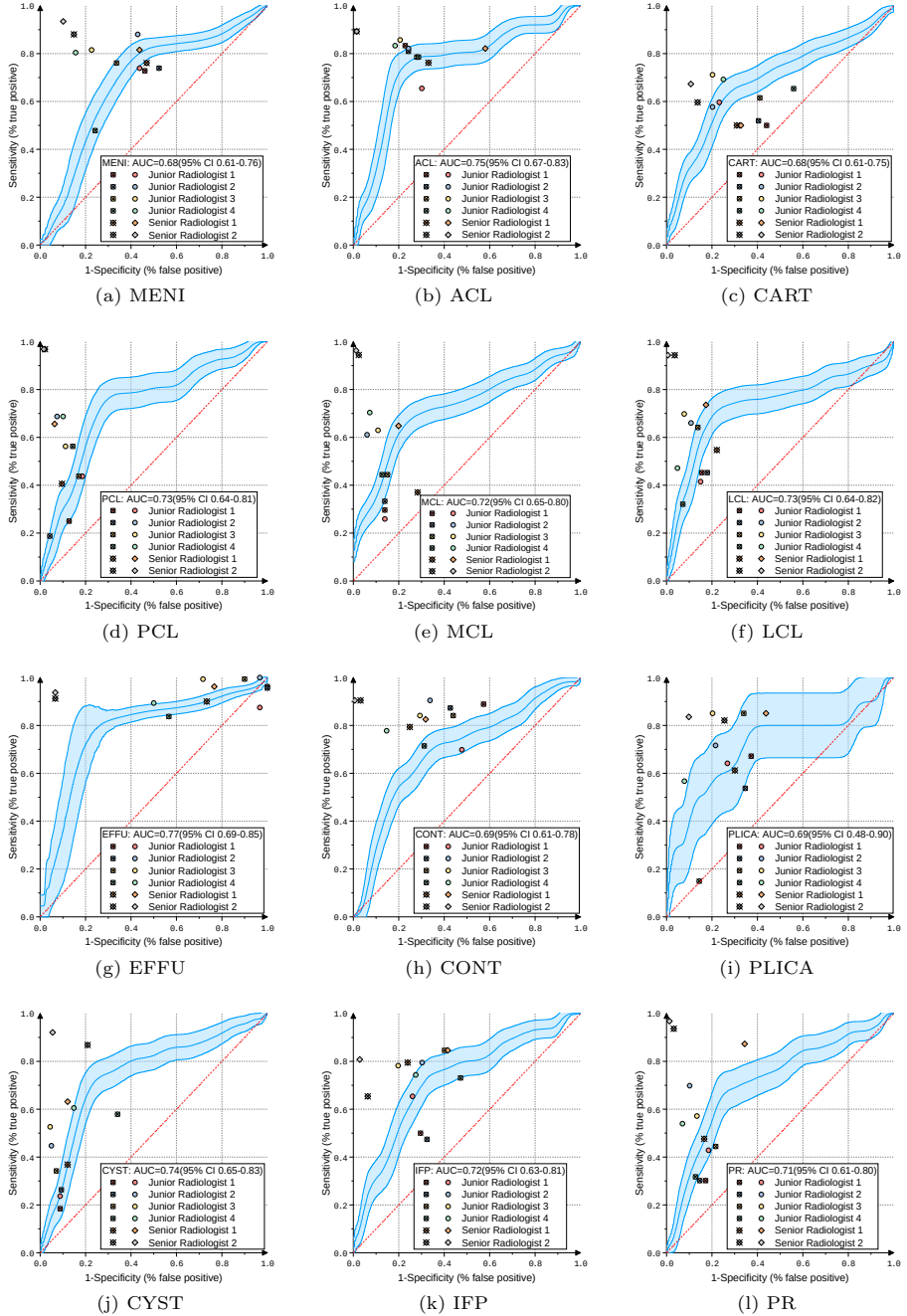


Fig. 1: ROC curve for each abnormality of our model's prediction and the (FPR, TPR) pair from the radiologists'. Each radiologist is denoted by points with and without cross indicating their results before and after AI assistance. Our model outperforms junior radiologists and makes significant improvements when assisting all radiologists. Please note that the points of Senior Radiologist 2 in (b) completely overlap. Source data are provided as a Source Data file.

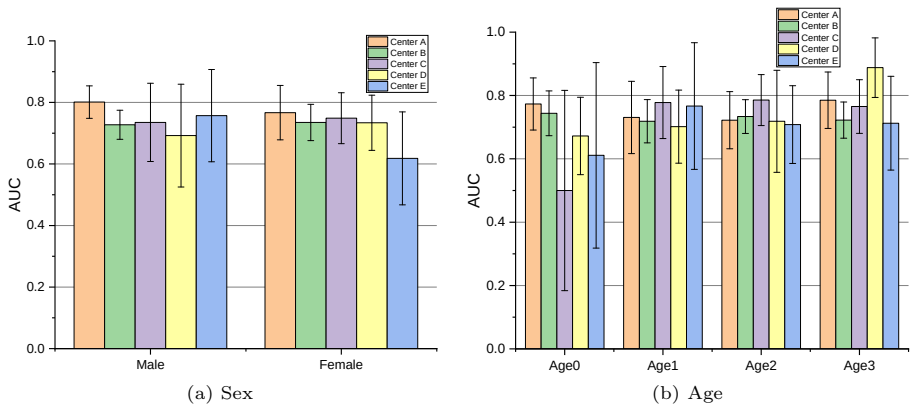


Fig. 2: Performance across different subgroups. Source data are provided as a Source Data file.

References

- [1] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer learning for 3D medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.
- [2] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Medicine*, 15(11):e1002699, 2018.
- [3] Niamh Belton, Ivan Welaratne, Adil Dahlan, Ronan T Hearne, Mísgina Tsighe Hagos, Aonghus Lawlor, and Kathleen M Curran. Optimising knee injury detection with spatial attention and validating localisation ability. In *Annual Conference on Medical Image Understanding and Analysis*, pages 71–86. Springer, 2021.
- [4] Chen-Han Tsai, Nahum Kiryati, Eli Konen, Iris Eshed, and Arnaldo Mayer. Knee injury detection using MRI with efficiently-layered network (elnet). In *Medical Imaging with Deep Learning*, pages 784–794. PMLR, 2020.