Supplementary Table 1. Comparison of AUC between the average of radiologists and the AI system on each dataset and after pooling the data.*

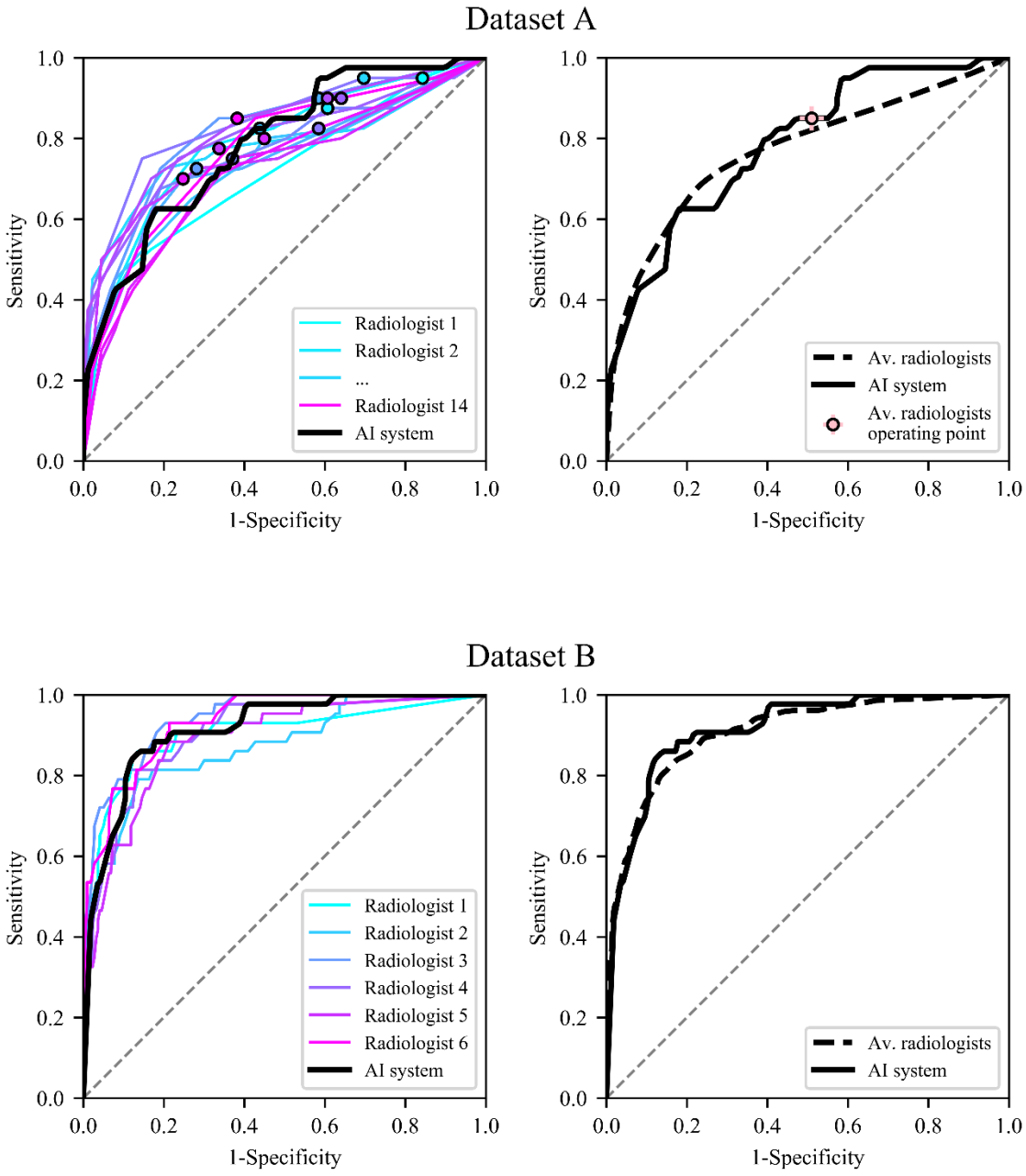
Dataset	Average AUC of radiologists (95% CI)	AUC AI system	AUC AI system – average AUC of radiologists (95%CI)	AUC best radiologist
A	0.769 (0.698, 0.840)	0.783	+0.014 (-0.058, 0.086)	0.834
B	0.907 (0.854, 0.961)	0.915	+0.008 (-0.035,0.051)	0.943
C	0.858 (0.814, 0.901)	0.879	+0.021 (-0.0287, 0.0715)	0.909
D	0.815 (0.767, 0.864)	0.850	+0.036 (-0.015, 0.087)	0.856
E1	0.787 (0.732, 0.841)	0.825	+0.038 (-0.012,0.088)	0.861
E2	0.803 (0.763, 0.843)	0.796	-0.007 (-0.057, 0.043)	0.872
F	0.860 (0.831, 0.889)	0.852	-0.008 (-0.038, 0.022)	0.869
G	0.808 (0.752, 0.859)	0.817	+0.009 (-0.038, 0.054)	0.839
H	0.841 (0.785, 0.897)	0.861	+0.020 (-0.035, 0.075)	0.869
Overall	0.814	0.840	+0.026 (-0.003,0.055)	--

*AI = Artificial Intelligence, AUC = Area Under the Curve, CI = Confidence Interval.

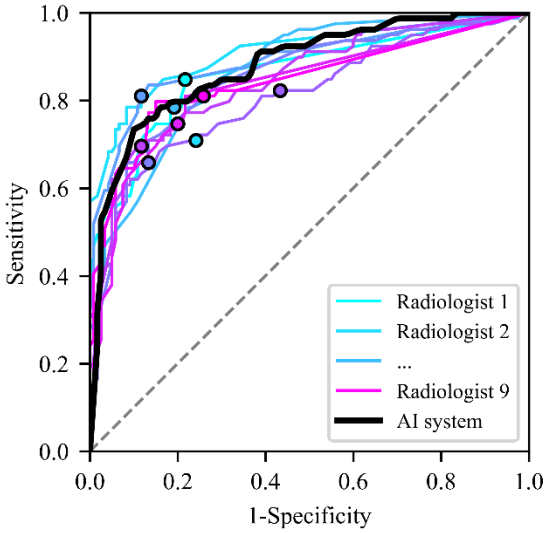
Supplementary Table 2. Average operating point of the radiologists using a case recall / no recall (BI-RADS >3) as threshold.*

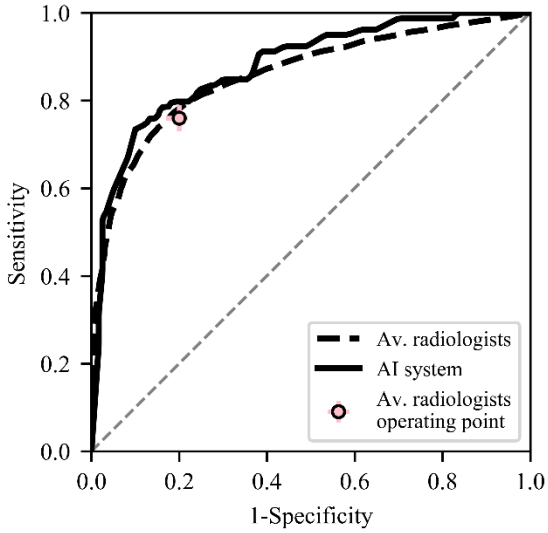
Dataset	Average specificity of radiologists (95% CI)	Average sensitivity of radiologists (95% CI)	Sensitivity AI at radiologists' specificity (95% CI)	Difference sensitivity AI - radiologists (95% CI)	AI system score threshold
A	0.49 (0.40,0.61)	0.84 (0.76,0.92)	0.85 (0.73,0.97)	+0.01 (-0.091,0.116)	7.32/10
B	NA	NA	NA	NA	NA
C	0.79 (0.73,0.86)	0.77 (0.70,0.83)	0.80 (0.70,0.90)	+0.03 (-0.062,0.126)	8.26/10
D	0.67 (0.62,0.73)	0.77 (0.67,0.87)	0.85 (0.77,0.93)	+0.08 (-0.009,0.175)	7.52/10
E1	0.54 (0.47,0.60)	0.82 (0.75,0.89)	0.86 (0.76,0.96)	+0.04 (-0.048,0.120)	7.12/10
E2	0.51 (0.46,0.57)	0.83 (0.78,0.88)	0.81 (0.73,0.89)	-0.02 (-0.088,0.051)	7.52/10
F	0.68 (0.61,0.76)	0.84 (0.77,0.90)	0.86 (0.80,0.92)	+0.02 (-0.088,0.051)	7.32/10
G	0.75 (0.67,0.82)	0.76 (0.67,0.85)	0.75 (0.65,0.85)	-0.01 (-0.107,0.075)	8.01/10
H	0.73 (0.65,0.80)	0.83 (0.76,0.89)	0.81 (0.73,0.89)	-0.02 (-0.112,0.070)	8.01/10

* For the AI system, the reported sensitivity is evaluated at the closest operating point to the average radiologist in terms of specificity (which corresponds to the specified threshold level). 95% CI are shown in parentheses, unless stated otherwise. AI = Artificial Intelligence, BI-RADS = Breast Imaging Reporting and Data System, CI = Confidence Interval, NA = Not Available.

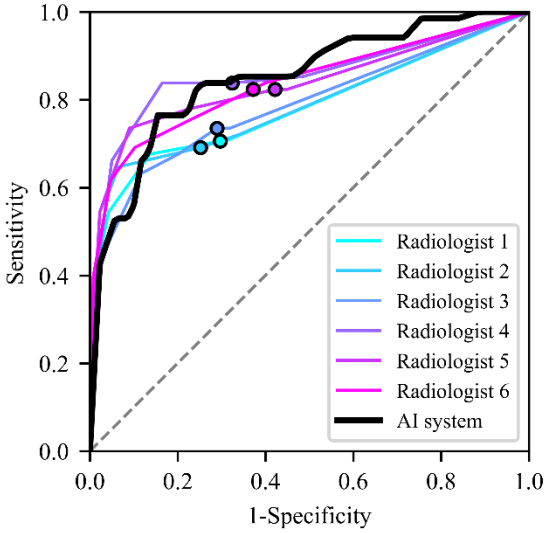

Dataset A
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Radiologist 1
Radiologist 2
...
Radiologist 14
AI system
Av. radiologists
AI system
Av. radiologists operating point
Dataset B
Radiologist 1
Radiologist 2
Radiologist 3
Radiologist 4
Radiologist 5
Radiologist 6
AI system
Av. radiologists
AI system

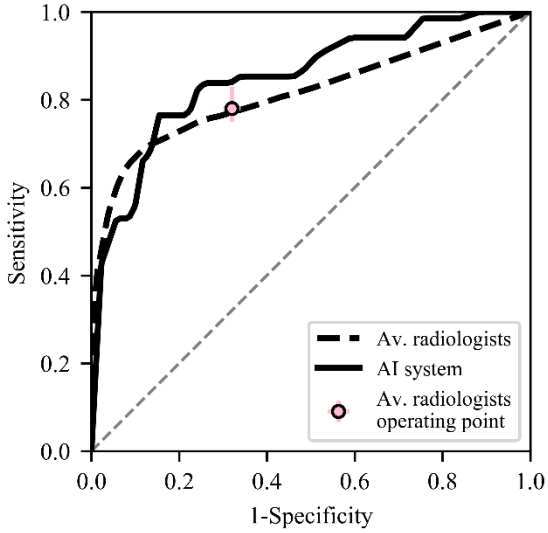
Dataset C

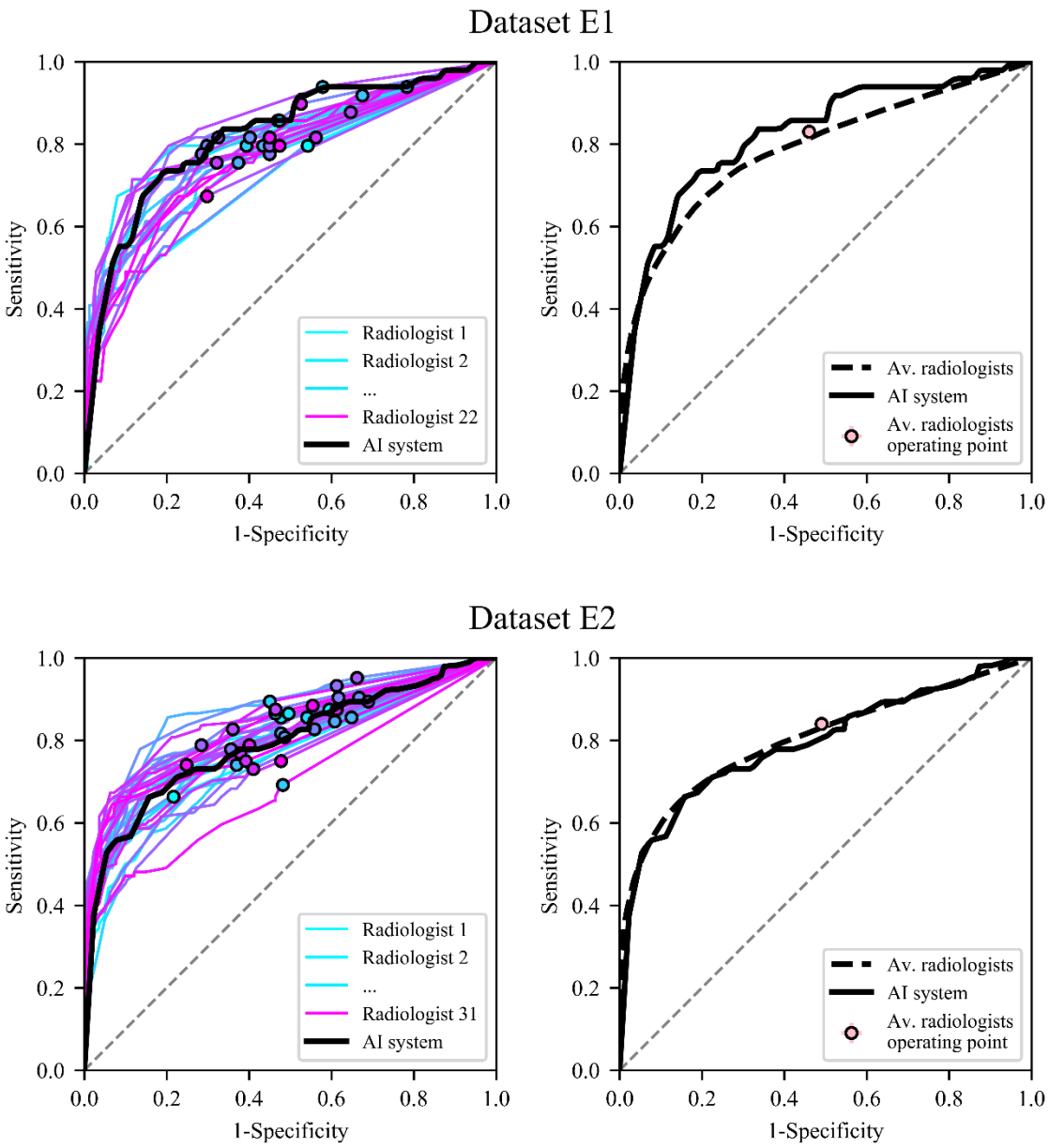




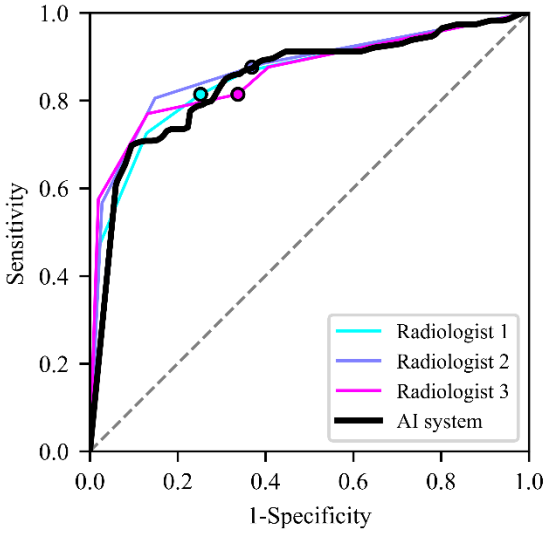
Dataset D

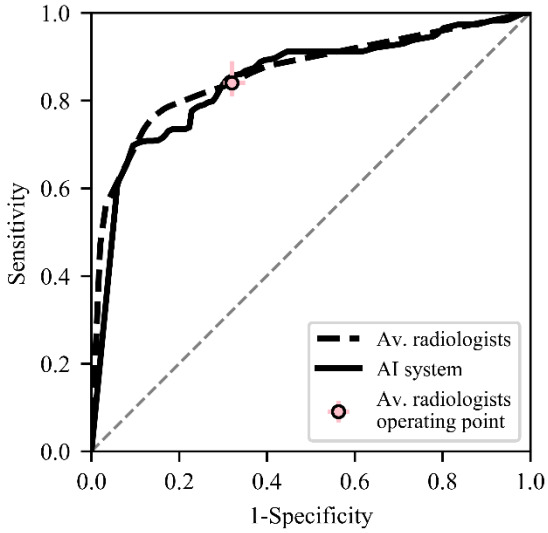




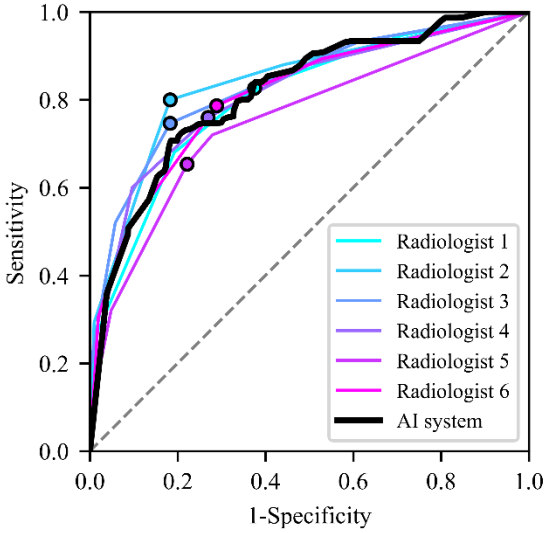

Dataset E1
Sensitivity
1-Specificity
Radiologist 1
Radiologist 2
...
Radiologist 22
AI system
Av. radiologists
AI system
Av. radiologists operating point
Dataset E2
Sensitivity
1-Specificity
Radiologist 1
Radiologist 2
...
Radiologist 31
AI system
Av. radiologists
AI system
Av. radiologists operating point

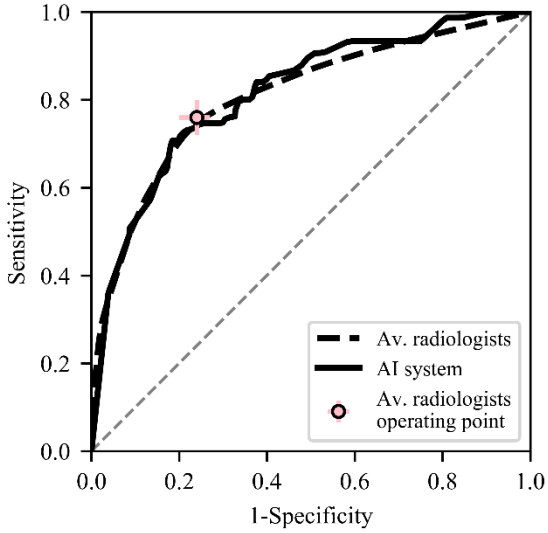
Dataset F

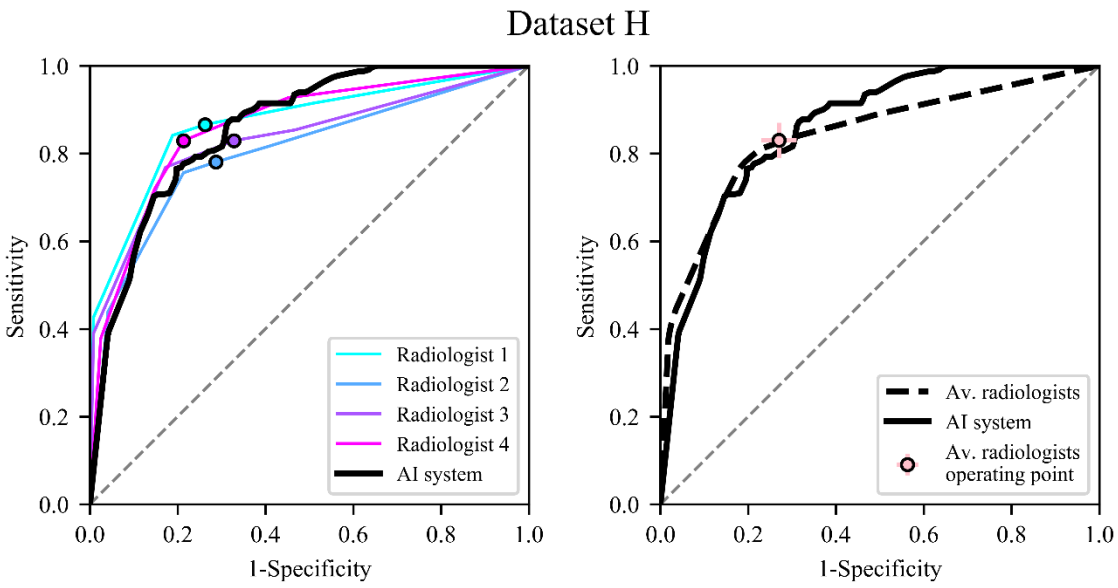




Dataset G







Supplementary Figure 1. For each dataset (A-H, see Table 1 for details), the receiver operating characteristic (ROC) curve of the artificial intelligence (AI) system is plotted against: (left) the ROC curves and operating points of each reader, and (right) the ROC curve and operating point of the average (Av.) of radiologists.