Evaluating the Clinical Utility of Artificial Intelligence Assistance and its Explanation on the Glioma Grading Task

Supplemental S2: AI Model and Explanation, Additional Quantitative Results

Weina Jin* Mostafa Fatehi* Ru Guo Ghassan Hamarneh

Contents

*co-first author

1 Training AI Model and Generating Explanation

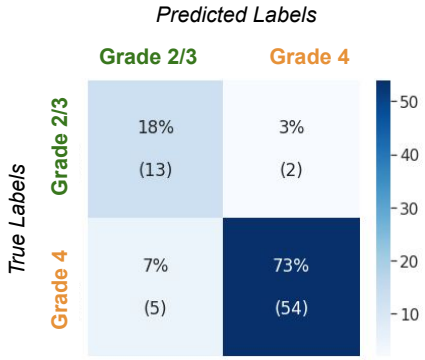
1.1 AI model and algorithmic evaluation on glioma grading task

We trained an AI model using the BraTS 2020 dataset to grade glioma MRIs. The AI model receives an MRI input and outputs a glioma grade of either a grade II/III glioma or a glioblastoma (GBM). The MRI input has the size of $4 \times 240 \times 240 \times 155$, which are the number of pulse sequences, height, width, and depth, respectively. The model architecture is a VGG-like [13] three-dimensional (3D) convolutional neural network (CNN), with six 3D CNN layers connected to two fully connected layers. During model training, to overcome the class imbalance issue, we used a weighted sampler to sample grade II/III glioma or GBM class, weighted by their inverse sample count. We used the cross-entropy loss function, and trained the model with an Adam optimizer, a learning rate of 0.0005, a batch size of 4, and 32 epochs.

We stratified split the BraTS dataset into 65% training (239 cases), 15% validation (56 cases), and 20% (74 cases) hold-out test set by keeping the same grade II/III glioma: GBM ratio in each set. There were no patient ID overlapping among the three datasets. We used the training data to train AI models, the validation to select the hyperparameters and best-performing model, and the hold-out test data to report AI model performance. The training, validation, and test accuracies of the AI model were 80.28%, 92.86%, and 90.54%, respectively. The fine-grained model performance metrics are in Fig. 1, which were also shown to participants in the clinical study.

We used PyTorch[1] and MONAI API[2] for model training, and Captum[3] to generate post-hoc color maps. To train the models and generate color maps, we used a computer with 1 GTX Quadro 24 GB GPU and 8 CPU cores, and a SLURM[4] based high performance computing cluster with jobs configured to use no more than a minimum of 1 GPU, and 8 cores CPU each with 128 GB RAM.

[1] http://pytorch.org
[2] http://monai.io
[3] http://captum.ai
[4] https://slurm.schedmd.com/overview.html



	Sensitivity	Specificity	Positive Predictive Value	F1 score	Number of class in the test set
Grade 2 or 3 glioma	0.87	0.92	0.72	0.79	*15*
Grade 4 glioblastoma	0.92	0.87	0.96	0.94	*59*

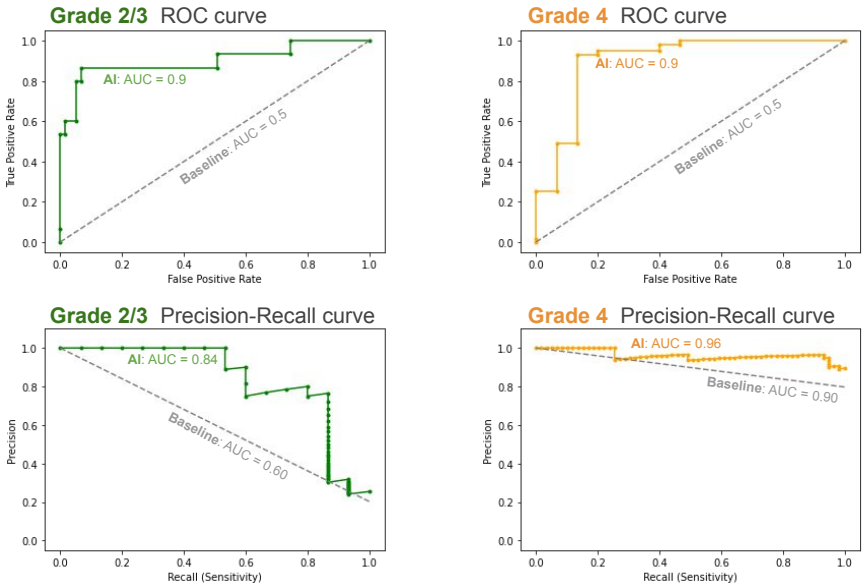


Figure 1: AI model performance metrics. A simplified version was also shown in the clinical study.

1.2 Generating and selecting the optimal AI explanation

The AI model we trained to grade glioma is a black-box CNN model. To explain the model decisions to physicians, we applied post-hoc XAI algorithms that act as a surrogate model to approximate the black-box AI model by probing the model parameters and/or input-output pairs. We aimed to select the optimal XAI algorithm to use in the clinical study from 16 post-hoc XAI algorithms: Gradient [12], Guided Back-Prop [15], Deconvolution [17], SmoothGrad [14], GradCAM [9], Guided GradCAM [9], Input$\times$Gradient [11], DeepLift [10], Integrated Gradients [16], Gradient Shap [7], Occlusion [17, 18], Feature Ablation [1], Feature Permutation [3], Lime [8], Shapley Value Sampling [2], and Kernel Shap [7]. These algorithms belong to the feature attribution method. They generate a feature attribution map or color map overlaid on the input image, and use the important image regions to explain model prediction. The selection criterion was to choose the most truthful XAI algorithm to the AI model decision process [5, 4]. Following the cumulative feature removal method to evaluate XAI truthfulness in Jin et al. [5, 6], we conducted a computational evaluation to calculate the ΔAUPC score from the cumulative feature removal method of the 16 XAI algorithms on the test set. The cumulative feature removal method iteratively removes the input features from the most to the least important features according to the color map explanation, and plots the relationship between gradual feature ablation and model accuracy. The evaluation metric ΔAUPC is to quantify the degree of performance deterioration by calculating the difference of area under the perturbation curve (AUPC) between an XAI algorithm $\mathcal{H}$ and its random feature removal baseline $\mathcal{H}_b$. ΔAUPC is in the range of $[-1, 1]$, with a high ΔAUPC indicating a more truthful explanation. SmoothGrad had the highest ΔAUPC score of 0.33, thus it was relatively the most truthful XAI method among the 16 XAI algorithms. We chose to use SmoothGrad as the optimal XAI method to generate AI model explanations in the clinical study. The result of the cumulative feature removal experiment is shown in Fig. 2.

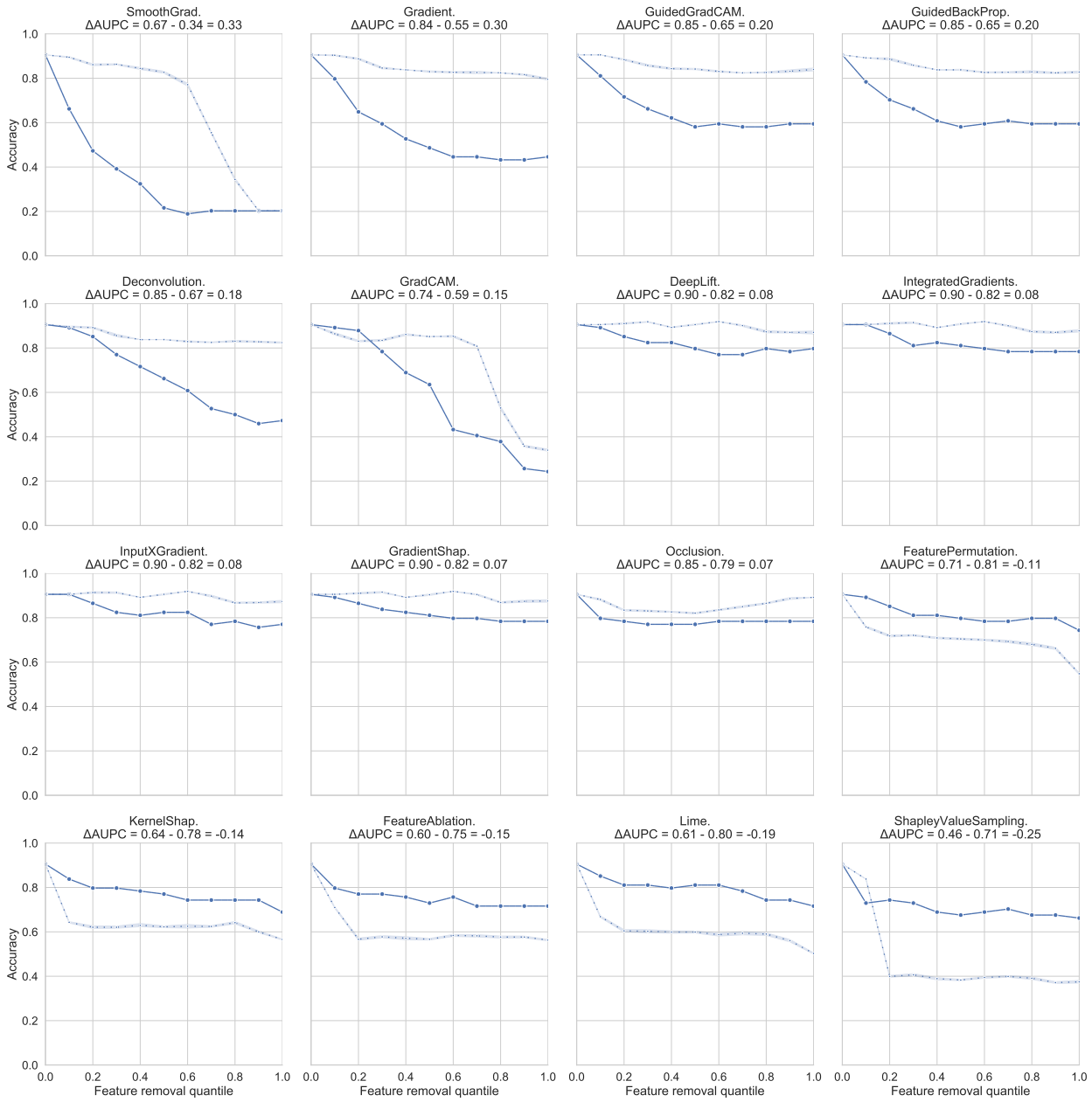


Figure 2: **Feature perturbation curve of cumulative feature removal experiment on the glioma task.** Each plot shows the feature perturbation curve of an XAI method ($\mathcal{H}$, solid line) and its random baseline ($\mathcal{H}_b$, dashed line). A bigger gap between the two curves indicates a higher ΔAUPC, thus a better performance on explanation truthfulness. Plots were arranged according to their ΔAUPC value ($\Delta\text{AUPC}(\mathcal{H}) = \text{AUPC}(\mathcal{H}_b) - \text{AUPC}(\mathcal{H})$, with numbers rounded to two decimal places) as indicated in the plot subtitle.

2 Additional Quantitative Results

2.1 Participants

DR ID	DR	DR + AI	DR + XAI	MRI num	Need XAI (%)	XAI Qual.	Position	Yr	AI Familiarity	AI Attitude
01	68.00	68.00	**72.00**	25	33.33	5.24 ± 1.24	Attending	13	hear of AI	Skeptical
02	66.67	**100.00**	100.00	2	100.00	5.00 ± 0.00	Attending			
03	84.00	84.00	84.00	25	16.00	1.04 ± 0.77	Attending	23	use AI in work/life	Not interested
04	80.00	**92.00**	92.00	25	100.00	8.68 ± 2.19	Resident	1	use AI in work/life	Interested
05	88.00	**92.00**	92.00	25	8.70	7.83 ± 1.03	Resident	3	hear of AI	Interested
06	76.00	**84.00**	**88.00**	25	12.00	5.25 ± 1.53	Attending	4	hear of AI	Skeptical
07	84.00	80.00	80.00	25	96.00	6.21 ± 2.24	Resident	4	hear of AI	Neutral
08	84.00	**96.00**	92.00	25	28.00	5.04 ± 2.07	Resident	7	hear of AI	Skeptical
09	87.50	87.50	**100.00**	8	25.00	7.25 ± 2.17	Resident			
10	80.00	**84.00**	84.00	25	20.00	6.60 ± 2.61	Attending	6		Excited
11	80.00	**84.00**	84.00	25	0.00	6.20 ± 1.77	Resident	4	can program, but not write AI code	Interested
12	80.00	**88.00**	88.00	25	96.00	7.80 ± 2.94	Fellow	7	can program, but not write AI code	Interested, Excited
13	84.00	84.00	84.00	25	13.04	6.56 ± 2.10	Attending	5	hear of AI	Interested, Excited
14	92.00	92.00	92.00	25	24.00	6.48 ± 1.70	Resident	0	can write AI code	Interested, Excited
15	88.00	**92.00**	92.00	25	16.00	2.56 ± 0.57	Attending	15	hear of AI	Skeptical
16	80.00	**84.00**	80.00	25	16.67	5.68 ± 3.40	Resident	1	use AI in work/life	Interested
17	80.00	**92.00**	92.00	25	12.50	9.00 ± 1.06	Resident	5	hear of AI	Interested
18	60.00	**76.00**	76.00	25	0.00	7.72 ± 2.46	Resident	4	hear of AI	Neutral
19	88.00	**96.00**	88.00	25	48.00	8.28 ± 1.43	Resident	4	hear of AI	Interested
20	80.00	80.00	80.00	25	12.00	8.42 ± 2.52	Attending	11	hear of AI	Interested
21	88.00	88.00	88.00	25	100.00	10.00 ± 0.00	Resident	4	can program, but not write AI code	Interested
22	80.00	80.00	**84.00**	25	88.00	5.68 ± 2.38	Resident	4	use AI in work/life	Excited
23	88.00	**92.00**	**96.00**	25	96.00	6.44 ± 1.50	Resident	7	use AI in work/life	Interested
24	68.00	**80.00**	**88.00**	25	16.00	5.92 ± 1.83	Resident		can program, but not write AI code	
25	82.35	**94.12**	94.12	17	17.65	5.62 ± 1.76	Resident			
26	84.00	84.00	84.00	25	20.00	6.40 ± 2.24	Resident	4	can program, but not write AI code	Interested, Excited
27	88.00	88.00	88.00	25	24.00	7.00 ± 2.83	Resident	4	can program, but not write AI code	Neutral
28	88.00	88.00	**92.00**	25	68.00	8.29 ± 1.51	Attending	5	hear of AI	Interested
29	100.00	100.00	100.00	3	66.67	2.00 ± 0.00	Fellow			
30	88.00	88.00	88.00	25	0.00	1.48 ± 0.77	Attending		hear of AI	
31	80.00	80.00	80.00	25	20.00	1.12 ± 0.43	Resident	8	never hear of AI	Not interested
32	100.00	100.00	100.00	2	0.00	7.50 ± 0.50	Attending			
33	88.00	84.00	**88.00**	25	25.00	5.71 ± 2.07	Attending	30	use AI in work/life	Excited
34	88.00	88.00	88.00	25	0.00	5.68 ± 2.29	Resident	9	can write AI code	Excited
35	66.67	**100.00**	100.00	2	50.00	9.50 ± 0.50	Resident			

Table 1: Participants' demographics and their accuracies (%) in three conditions: 1) DR: without AI assistance, 2) DR+AI: with the assistance of AI prediction, and 3) DR+XAI: with the assistance of both AI prediction and explanation. We mark the accuracy in bold in DR+AI column if it is higher than DR; similarly, we mark the accuracy in bold in DR+XAI column if it is higher than DR+AI. MRI num column indicates the number of MRIs a participant interpreted in the survey. Need XAI is the percentage of participants needing to check AI explanation for the MRI case. XAI Qual. is the mean±std rating from participants on the explanation quality for each color map explanation on a [0, 10] scale. Participants' demographics, including their position, years of experience in neurosurgery (Yr), familiarity with AI, and attitude toward AI are also listed.

Data ID	GT	AI Pred.	DR	DR + AI	DR + XAI	DR num	Need XAI (%)	XAI Qual.	MRI link
BraTS20_Training_221	1	1	100.00	100.00	100.00	35	50.00	7.23 ± 2.75	vimeo.com/558775183
BraTS20_Training_208	1	1	100.00	100.00	100.00	30	26.67	6.79 ± 2.58	vimeo.com/558775334
BraTS20_Training_116	1	1	100.00	100.00	100.00	30	16.67	6.66 ± 2.69	vimeo.com/558785220
BraTS20_Training_114	1	1	93.55	93.55	93.55	31	35.48	6.55 ± 2.47	vimeo.com/558795281
BraTS20_Training_112	1	1	93.55	**100.00**	100.00	31	25.81	6.93 ± 2.72	vimeo.com/558795007
BraTS20_Training_099	1	1	19.35	**35.48**	**41.94**	31	76.67	3.90 ± 2.68	vimeo.com/558764148
BraTS20_Training_094	1	1	80.65	**93.55**	90.32	31	30.00	6.39 ± 2.87	vimeo.com/558764221
BraTS20_Training_093	1	1	100.00	100.00	100.00	30	24.14	7.13 ± 2.39	vimeo.com/558768608
BraTS20_Training_289	0	0	86.21	**93.10**	**96.55**	29	37.93	5.50 ± 2.80	vimeo.com/558897027
BraTS20_Training_076	1	1	96.77	**100.00**	100.00	31	26.67	6.52 ± 2.89	vimeo.com/558768466
BraTS20_Training_075	1	1	96.67	**100.00**	96.67	30	26.67	6.61 ± 2.79	vimeo.com/558768604
BraTS20_Training_070	1	1	80.00	**90.00**	**96.67**	30	30.00	6.90 ± 2.59	vimeo.com/546342295
BraTS20_Training_325	0	0	87.10	**93.33**	93.33	30	37.93	3.13 ± 2.85	vimeo.com/558895720
BraTS20_Training_064	1	1	100.00	100.00	100.00	30	20.00	6.93 ± 2.50	vimeo.com/558762879
BraTS20_Training_063	1	1	48.39	**64.52**	**67.74**	31	50.00	6.10 ± 2.72	vimeo.com/558762829
BraTS20_Training_060	1	1	89.66	**96.55**	96.55	29	37.93	6.38 ± 2.59	vimeo.com/558758233
BraTS20_Training_056	1	1	96.67	**100.00**	100.00	30	23.33	7.27 ± 2.64	vimeo.com/558758468
BraTS20_Training_053	1	1	100.00	100.00	100.00	30	20.69	7.20 ± 2.50	vimeo.com/558758697
BraTS20_Training_270	0	1	0.00	0.00	0.00	29	17.24	6.65 ± 2.64	vimeo.com/558784455
BraTS20_Training_277	0	0	41.94	**61.29**	61.29	31	43.33	6.13 ± 2.58	vimeo.com/546658075
BraTS20_Training_269	0	0	96.67	**100.00**	100.00	30	30.00	5.63 ± 3.04	vimeo.com/558775144
BraTS20_Training_264	0	0	100.00	100.00	100.00	29	24.14	5.29 ± 2.76	vimeo.com/558775501
BraTS20_Training_280	0	0	83.87	**90.32**	90.32	31	34.48	5.74 ± 2.88	vimeo.com/558774958
BraTS20_Training_171	1	0	83.33	70.00	**76.67**	30	65.52	5.07 ± 2.82	vimeo.com/558769028
BraTS20_Training_212	1	0	83.33	70.00	66.67	30	58.62	4.13 ± 2.88	vimeo.com/558786569

Table 2: We list the 25 MRIs used in the study with the ground-truth label (the column GT) of grade II/III glioma (label 0) or GBM (label 1), the predicted label from AI (AI Pred.), and the participants' accuracies (%) in three conditions: 1) DR: without AI assistance, 2) DR+AI: with the assistance of AI prediction, and 3) DR+XAI: with the assistance of both AI prediction and explanation. We mark the accuracy in bold in DR+AI column if it is higher than DR; similarly, we mark the accuracy in bold in DR+XAI column if it is higher than DR+AI. DR num is the number of collected responses from participants. Need XAI is the percentage of participants needing to check AI explanation for the MRI case. XAI Qual. is the mean±std rating from participants on the explanation quality for each color map explanation on a [0, 10] scale. The MRI and its color map explanation can be viewed by following the video links in the MRI link column.

2.2 Physicians' task performance with and without the assistance of AI and its explanation

In the manuscript, we have reported the performance using the accuracy metric. Here we also report results using other performance metrics, including F1, Matthews correlation coefficient (MCC), sensitivity, specificity, and positive predictive value (also called precision) in Table 3. Statistical tests show similar trends as the result reported using accuracy: using Friedman tests, there are statistically significant differences in task performance measured by a particular metric among the three conditions. Post-hoc analysis using Wilcoxon signed-rank tests with Bonferroni correction showed that DR+AI condition had a statistically higher performance compared to the DR condition; similarly, the DR+XAI condition had a statistically higher performance compared to the DR condition. However, the performances between DR+AI and DR+XAI conditions did not show a statistically significant difference.

Metric	AI	DR	DR+AI	DR+XAI	p-value: Friedman	p-value: DR vs. DR+AI	p-value: DR vs. DR+XAI	p-value: DR+AI vs. DR+XAI
Acc.	0.8800	0.8249 ± 0.0869	0.8770 ± 0.0733	0.8852 ± 0.0702	7.755e-06	0.001543	0.0004037	0.3454
F1 grade II/III	0.8000	0.6424 ± 0.2257	0.6997 ± 0.2356	0.7116 ± 0.2390	4.341e-05	0.002134	0.0006973	0.5456
F1 GBM	0.9143	0.8777 ± 0.0644	0.9126 ± 0.0559	0.9187 ± 0.0535	7.755e-06	0.002895	0.0008773	0.6383
MCC	0.7181	0.5448 ± 0.2274	0.6242 ± 0.2340	0.6415 ± 0.2365	4.341e-05	0.002134	0.0006985	0.5456
Sen. grade II/III/Spec. GBM	0.8571	0.6694 ± 0.2444	0.7184 ± 0.2510	0.7224 ± 0.2521	0.0001117	0.0178	0.01081	0.9519
Sen. GBM/Spec. grade II/III	0.8889	0.8800 ± 0.0961	0.9060 ± 0.0870	0.9156 ± 0.0834	0.001765	0.05405	0.008309	0.7735
PPV grade II/III	0.7500	0.6412 ± 0.2502	0.7041 ± 0.2608	0.7226 ± 0.2635	4.341e-05	0.006771	0.001361	0.7418
PPV GBM	0.9412	0.8845 ± 0.0751	0.9239 ± 0.0472	0.9263 ± 0.0455	2.901e-05	0.001361	0.0006973	0.6061

Table 3: Physicians' task performance with physicians alone (DR), with the assistance of AI prediction (DR+AI), and with the assistance of AI prediction and its explanation (DR+XAI), using multiple performance metrics. We also list AI performance using these performance metrics (AI), the p-value of the Friedman test on the performance differences among the three conditions (DR, DR+AI, DR+XAI), and the p-values using Wilcoxon signed-rank tests with Bonferroni correction on the pairwise conditions.

In addition to the task performance analysis on the participants' data as a whole, we performed subgroup analysis by dividing participants into two groups according to their clinical positions and experience: 1) attending physicians, and 2) resident and fellow physicians. The descriptive statistics of the task performance accuracies for each subgroup are shown in Table 4 and 5.

Condition	N	M±SD	Min	25% Q	Mdn	75% Q	Max
DR	12	82.56 ± 9.28	66.67	79.00	84.00	88.00	100.00
DR+AI	12	86.33 ± 8.61	68.00	84.00	84.00	89.00	100.00
DR+XAI	12	87.67 ± 7.90	72.00	84.00	88.00	92.00	100.00

Table 4: Descriptive statistics for attending physicians' task performance accuracy (%). N - number of participants, M - mean, SD - standard deviation, Q - quantile, Mdn - median.

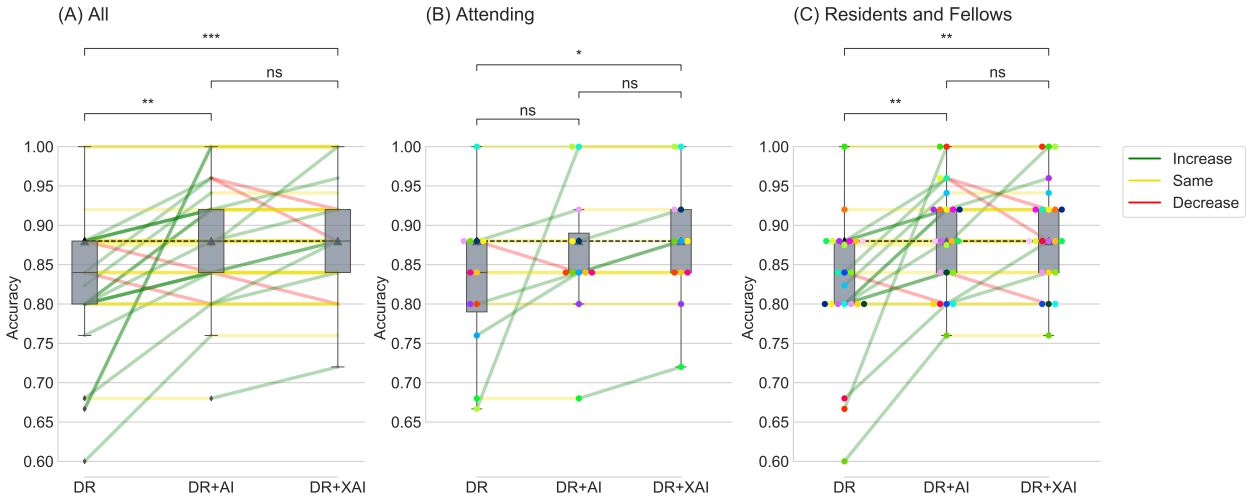


Figure 3: Participants' task accuracies on glioma grading in three conditions: 1) DR: Physicians performing the task alone; 2) DR+AI: Physicians performing the task with AI assistance (with predictions from AI); 3) DR+XAI: Physician performing the task with XAI assistance (with predictions and explanations from AI). All participants' data are visualized in panel (A), and the fine-grained attending physicians' or resident + fellow physicians' data are shown in panel (B) and (C) respectively. For each panel, we show box plots for the three conditions. The color of the dots indicates each participant's accuracy. The colored lines indicate a participant's accuracy change in between different conditions, with green lines indicating an accuracy increment, yellow indicating the same, and red indicating a decrement. The darkness of colored lines encodes the frequency of such a change. The horizontal dashed line indicates the AI accuracy of 0.88. ns: $p > 0.05$, $*$: $.01 \leq p \leq .05$, $**$: $.001 \leq p \leq .01$, $***$: $.0001 \leq p \leq .001$.

For the accuracy of each condition in the attending physician subgroup, the non-parametric Friedman test showed a statistically significant difference in task accuracies among the three conditions, $\chi^2_F(2) = 8.27, p = .016$. We then conducted post-hoc analysis using Wilcoxon signed-rank tests with Bonferroni correction. The results showed that there was not a significant difference in accuracy between the **DR+AI** and **DR** conditions ($Z = 9.5, p = .51$), and between **DR+XAI** and **DR+AI** conditions ($Z = 0.0, p = .14$). However, the **DR+XAI** condition had a statistically higher accuracy compared to the **DR** condition ($Z = 0.0, p = .047$). We also calculated the effect size using common language effect size, and results showed a physician has a probability of 60.8% of having a higher accuracy when assisted by AI prediction (**DR+AI**) than performing the task alone (**DR**), a probability of 66.7% of having a higher accuracy when assisted by AI prediction and explanation (**DR+XAI**) than performing the task alone (**DR**), but only a probability of 56.3% of having a higher accuracy when assisted by AI prediction and explanation (**DR+XAI**) than assisted by AI prediction alone (**DR+AI**).

Condition	N	M±SD	Min	25% Q	Mdn	75% Q	Max
DR	23	82.46 ± 8.58	60.00	80.00	84.00	88.00	100.00
DR+AI	23	88.42 ± 6.67	76.00	84.00	88.00	92.00	100.00
DR+XAI	23	88.96 ± 6.66	76.00	84.00	88.00	92.00	100.00

Table 5: Descriptive statistics for resident and fellow physicians' task performance accuracy (%). N - number of participants, M - mean, SD - standard deviation, Q - quantile, Mdn - median.

For the accuracies in each condition in the resident and fellow physician subgroup, the non-parametric Friedman test showed a statistically significant difference in task accuracies among the three conditions, $\chi^2_F(2) = 16.98, p = .0002$. We then conducted post-hoc analysis using Wilcoxon signed-rank tests with Bonferroni correction. The results showed that the **DR+AI** condition had a statistically higher accuracy compared to the **DR** condition ($Z = 2.0, p = .004$); similarly, the **DR+XAI** condition had a statistically

higher accuracy compared to the **DR** condition ($Z = 2.0, p = .004$). However, the accuracies between **DR+AI** and **DR+XAI** conditions did not show statistically significant difference ($Z = 10.5, p = 1.6$). We also calculated the effect size using common language effect size, and results showed a physician has a probability of 70.2% of having a higher accuracy when assisted by AI prediction (**DR+AI**) than performing the task alone (**DR**), a probability of 73.2% of having a higher accuracy when assisted by AI prediction and explanation (**DR+XAI**) than performing the task alone (**DR**), but only a probability of 52.4% of having a higher accuracy when assisted by AI prediction and explanation (**DR+XAI**) than assisted by AI prediction alone (**DR+AI**).

To visualize the change in participants' task performance in each condition, we show the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve in Fig. 4 for AI and participants in each condition. The performance change as indicated by the arrows showed participants' performances had the tendency to point to the direction of better performance area (for the ROC plot, it is the upper left corner; for the PR plot, the upper right corner). This aligns with the main finding that AI assistance improved physicians' task performance. Such a performance boost was more prominent in participants whose initial task performance was inferior to AI performance (below the AI ROC or PR curve). There are rare participants who had arrows across the AI curve, and most arrow heads landed on or below the AI curve, indicating complementary doctor-AI performance can rarely be achieved for the study participants.

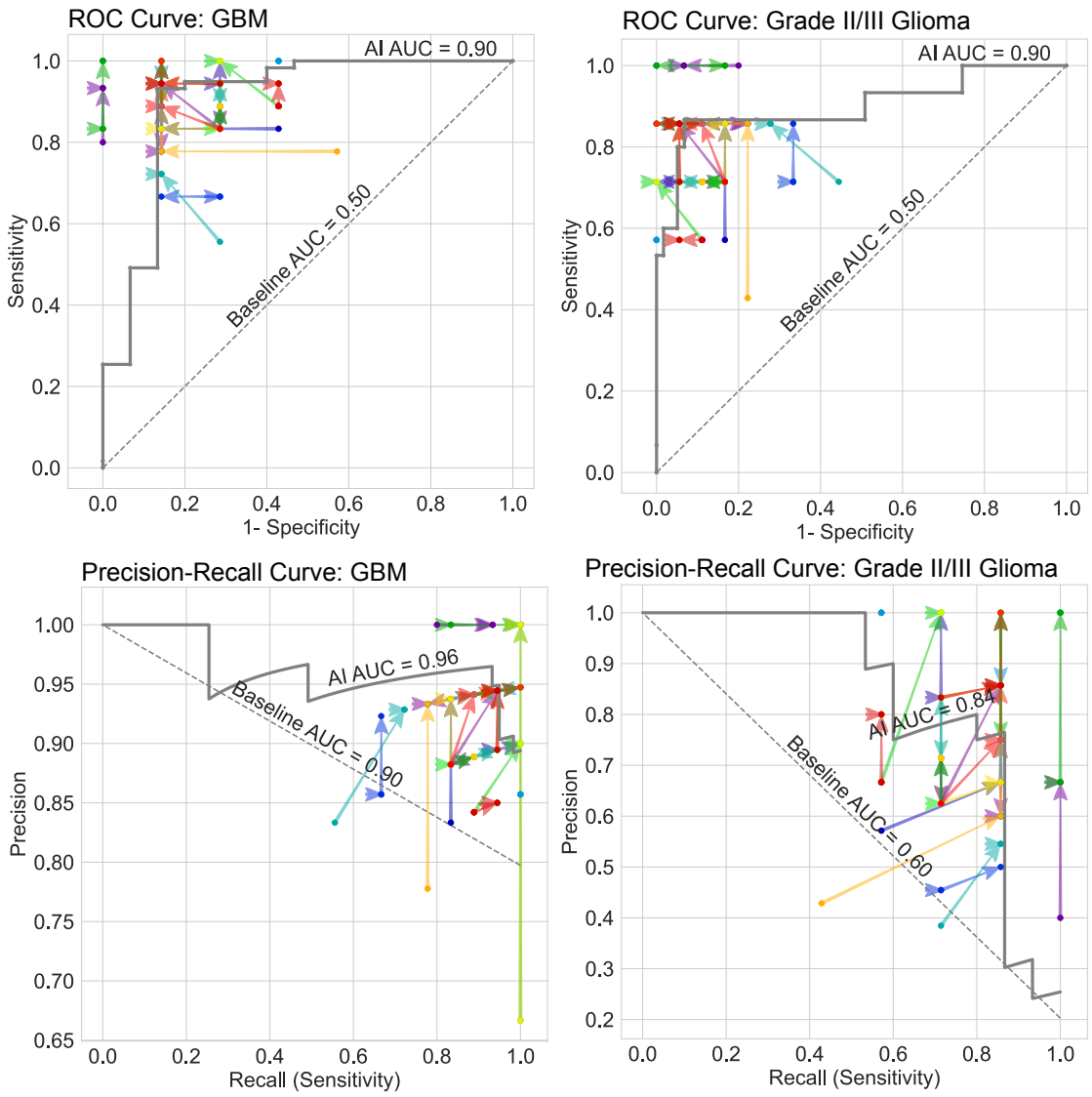


Figure 4: The receiver operating characteristic (ROC) curve (upper row) and precision-recall (PR) curve (lower row) to evaluate the performance on GBM (left) and grade II/III glioma prediction (right). In each plot, the AI model's performance is indicated as the gray curve, and the physicians' performance is indicated as dots, with different colors representing different participants. The arrows in between dots are the performance change between conditions from **DR** (doctor performing the task alone) to **DR+AI** (doctor assisted by AI), and from **DR+AI** to **DR+XAI** (doctor assisted by XAI). For the ROC curve, a better performance would be near the upper left corner. For the PR curve, a better performance would be near the upper right corner. We also indicate the AUC (area under the curve) value for the curve of AI and random guess baselines (the dashed gray line).

We also computed the correlation between physicians' task performance (measured using accuracy) improvement and their clinical experience, using the Spearman correlation coefficient. The results showed that there was a negative small correlation ($r = -0.23$, p-value =0.25) between physicians' years of practicing neurosurgery, and their performance improvement after AI prediction assistance (the performance difference between **DR+AI** and **DR**). This indicates physicians' clinical experience may be a weak indicator to predict AI assistant performance improvement, with junior physicians benefiting more from AI prediction assistance.

Furthermore, there was no correlation ($r = -0.05$, p-value =0.82) between physicians' years of practicing neurosurgery, and their performance improvement after AI prediction and explanation assistance (the performance difference between **DR+XAI** and **DR**). This may indicate physicians' performance improvement with both AI prediction and explanation assistance may be a more complex process, and physicians' clinical experience may not be a good single indicator to predict performance improvement during this process. We plot the correlation regression lines in Fig. 5.

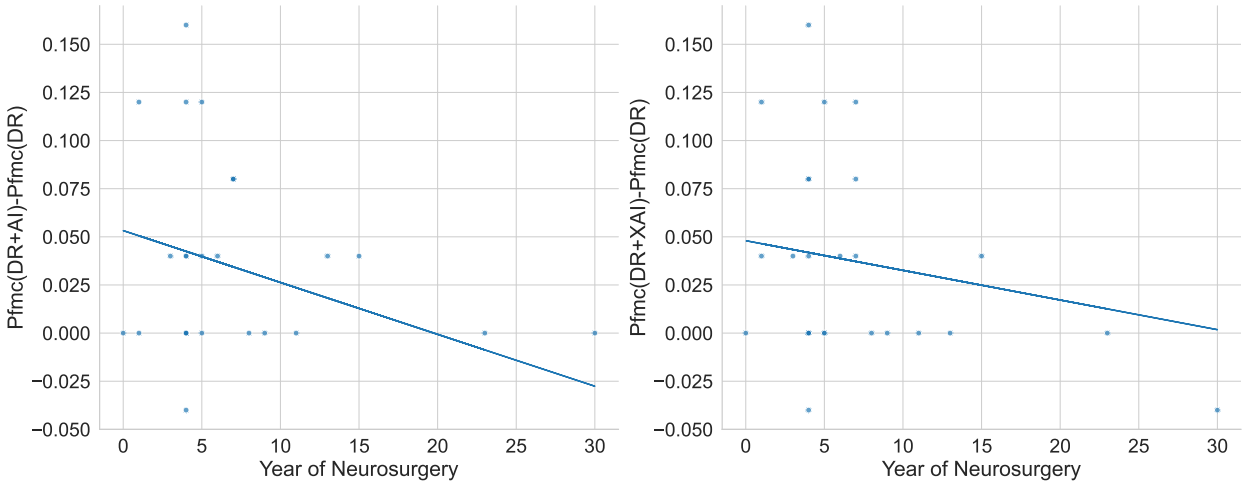


Figure 5: Scatter plot between participants' years of practicing neurosurgery (x axis), and their accuracy improvement (y axis) with the assistance of AI prediction (left), and with the assistance of AI prediction and explanation (right). We also show the regression lines for each plot, and the coefficients of the regression lines were -0.0027 and -0.0015, respectively.

2.3 Decision agreement and decision change

For the decision agreement pattern of the subgroups of attending vs. resident+fellow physicians, as shown in Fig. 6, as a baseline when physicians performed the task alone (**DR** condition), the decision agreement was 82.7% (210) for attending physicians, and 80.2% (405) for resident+fellow physicians. When physicians were assisted by AI prediction (**DR+AI** condition), the decision agreement increased to 85.4% (217) for attending physicians, and 87.5% (442) for resident+fellow physicians. When physicians were assisted by AI prediction and explanation (**DR+XAI** condition), the decision agreement was the same for attending physicians, 85.4% (217), and slightly increased to 88.1% (445) for resident+fellow physicians.

As shown in Fig. 6, physicians' subgroup decision change patterns had similar trends as the whole group analysis in the manuscript.

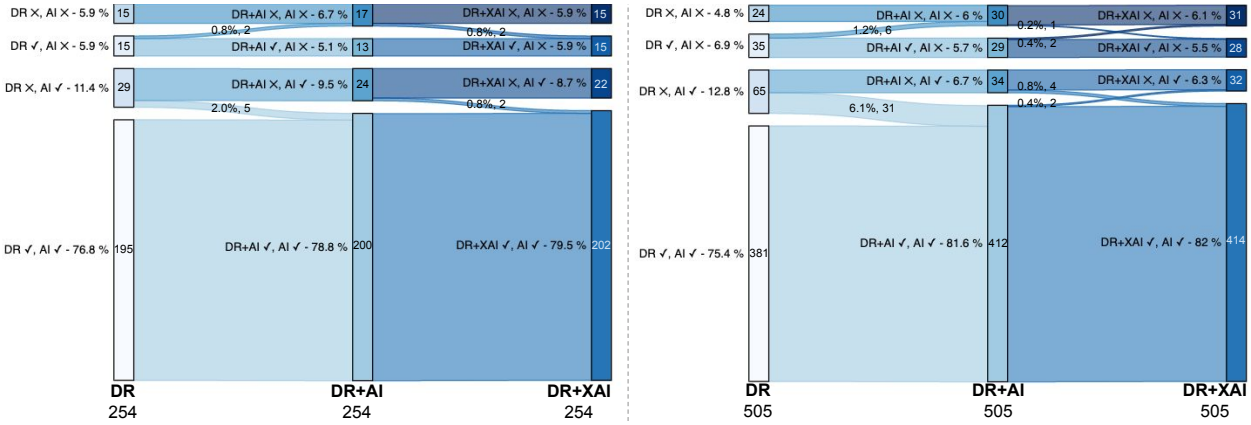


Figure 6: Participants' decision change stream plot for each error category for attending physicians (left), and resident+fellow physicians (right). The three columns represent the three conditions of DR, DR+AI, and DR+XAI, respectively. The rectangles in each column show the doctors' and AI's decision correctness. The decision agreement is the first and fourth rectangles, where doctors and AI made the same decisions (where both were incorrect or correct). The percentage and absolute number for each category are indicated. The total number of decisions was 254 and 505 for attending and resident+fellow physicians, respectively.

2.4 Trust and willingness to use AI

For participants' level of trust and willingness to use AI, we conducted a subgroup analysis based on two subgroups of participants' clinical positions: 1) attending physicians; 2) resident and fellow physicians. We used the non-parametric Friedman test to test if there were significant differences in participants' trust and willingness to use AI at three time points: 1) the initial baseline without knowing any information from AI; 2) after viewing AI performance metrics, and 3) after using AI predictions and explanations for the 25 MRIs.

For attending physicians, results did not show a statistically significant difference among the three time points for both trust in AI ($\chi^2_F(2) = 5.55, p = .062$), and willingness to use AI ($\chi^2_F(2) = 1.75, p = .416$). The descriptive statistics are shown in Table 6 and Fig. 7.

	Time point	N	M±SD	Min	25% Q	Mdn	75% Q	Max
Trust	Bsl	10	5.20 ± 2.35	2.00	3.50	5.00	5.75	9.00
	Pfm	10	6.80 ± 2.15	3.00	7.00	7.00	8.00	9.00
	Use	10	6.50 ± 2.84	2.00	3.75	7.50	9.00	9.00
Willingness	Bsl	10	4.00 ± 3.40	0.00	2.00	3.00	4.50	10.00
	Pfm	10	4.90 ± 3.28	0.00	2.50	5.00	7.50	10.00
	Use	10	3.80 ± 3.39	0.00	1.00	2.50	7.50	8.00

Table 6: Descriptive statistics for attending physicians' trust and willingness to use AI. N - number of participants, M - mean, SD - standard deviation, Q - quantile, Mdn - median. The three time points are: 1) Bsl: the initial baseline without knowing any information from AI; 2) Pfm: after viewing AI performance metrics, and 3) Use: after using AI predictions and explanations for the 25 MRIs.

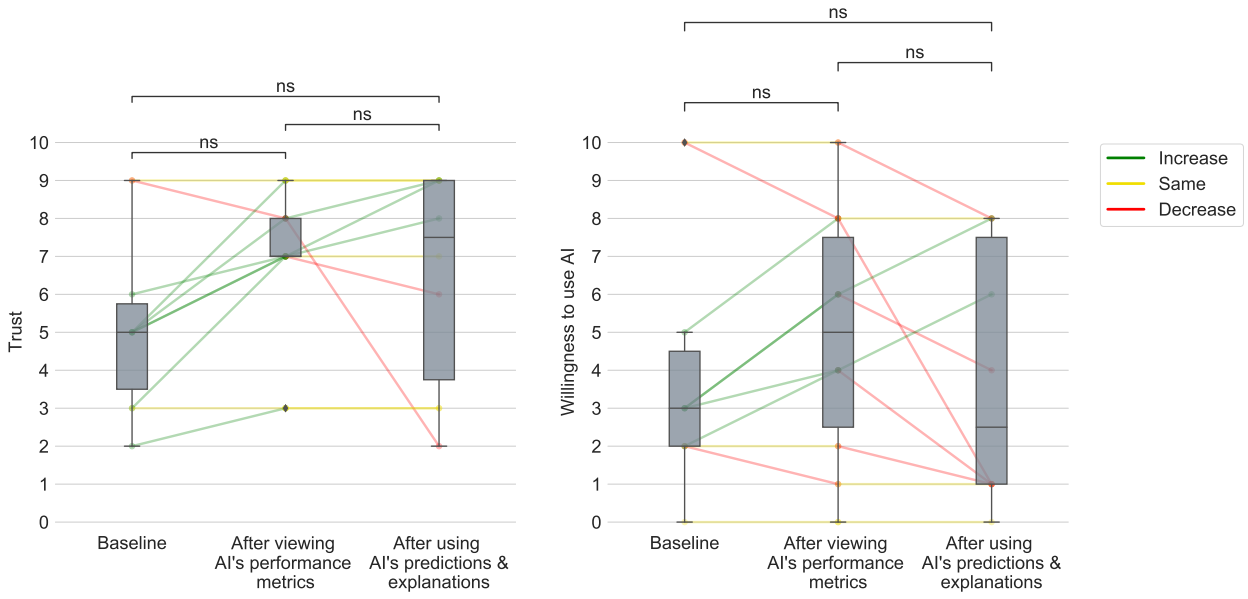


Figure 7: Box plots and changes of attending physicians' trust in the AI (left), and willingness to use AI (right) at the initial baseline, after viewing AI performance metrics, and after using the AI predictions and explanations. Both dependent variables are reported on a 0-10 point scale. The colored lines in between indicate change for each participant, with green indicating an increment, yellow indicating no change, and red indicating a decrement. The darkness of colored lines encodes the frequency of such a change. ns: $p > 0.05$, $*: .01 \leq p \leq .05$, $**: .001 \leq p \leq .01$, $***: .0001 \leq p \leq .001$.

For resident and fellow physicians, results showed a statistically significant difference among the three time points for both trust in AI ($\chi^2_F(2) = 11.47, p = .003$), and willingness to use AI ($\chi^2_F(2) = 7.86, p = .0196$).

We conducted post-hoc analysis using Wilcoxon signed-rank tests with Bonferroni correction to identify the significantly different pairs. For the level of trust in the AI system, resident and fellow physicians rated a statistically higher trust after viewing AI performance metrics compared with the initial baseline ($Z = 0.0, p = .00587$); but the trust level did not show a statistically significant difference after using AI predictions and explanations for the 25 MRIs compared with the initial baseline ($Z = 35.5, p = .085$); and there was no significant difference between the trust level after viewing AI performance metrics and after using AI predictions and explanations for the 25 MRIs ($Z = 53.0, p = 2.06$). Similarly, for the level of willingness to use AI, participants only rated a statistically higher willingness to use AI after viewing AI performance metrics compared with the initial baseline ($Z = 9.0, p = .026$); and the rest pairwise test did not show a statistically significant difference. The descriptive statistics are shown in Table 7, and the statistical test results are visualized in Fig. 8.

	Time point	N	M±SD	Min	25% Q	Mdn	75% Q	Max
Trust	Bsl	19	5.37 ± 1.92	0.00	5.00	5.00	7.00	8.00
	Pfm	19	6.68 ± 1.42	4.00	5.00	7.00	8.00	8.00
	Use	19	6.68 ± 2.67	0.00	7.00	8.00	8.00	9.00
Willingness	Bsl	19	4.16 ± 2.52	0.00	3.00	5.00	5.00	8.00
	Pfm	19	5.16 ± 1.98	1.00	3.50	5.00	6.50	8.00
	Use	19	5.00 ± 3.04	0.00	2.50	5.00	7.50	9.00

Table 7: Descriptive statistics for resident and fellow physicians' trust and willingness to use AI. N - number of participants, M - mean, SD - standard deviation, Q - quantile, Mdn - median. The three time points are: 1) Bsl: the initial baseline without knowing any information from AI; 2) Pfm: after viewing AI performance metrics, and 3) Use: after using AI predictions and explanations for the 25 MRIs.

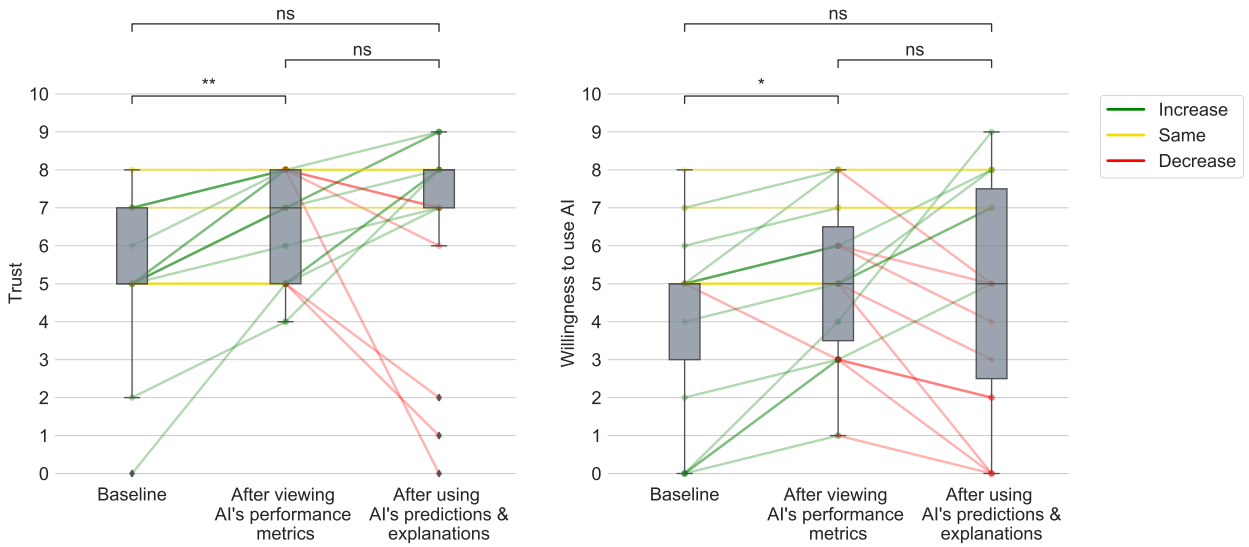


Figure 8: Box plots and changes of resident and fellow physicians' trust in the AI (left), and willingness to use AI (right) at the initial baseline, after viewing AI performance metrics, and after using the AI predictions and explanations. Both dependent variables are reported on a 0-10 point scale. The colored lines in between indicate change for each participant, with green indicating an increment, yellow indicating no change, and red indicating a decrement. The darkness of colored lines encodes the frequency of such a change. ns: $p > 0.05$, $* : .01 \leq p \leq .05$, $** : .001 \leq p \leq .01$, $*** : .0001 \leq p \leq .001$.

2.5 Clinical usage scenarios for AI explanation

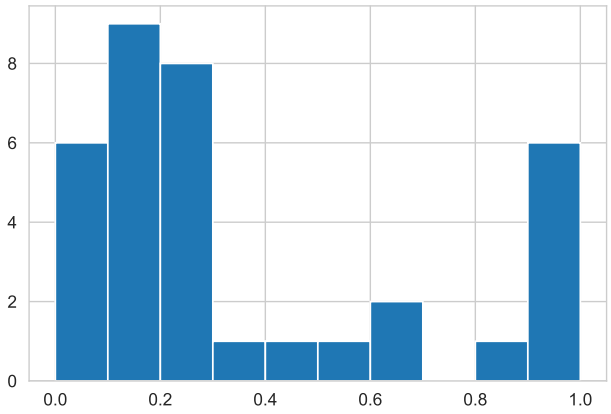


Figure 9: A histogram showing the distribution of 35 participants' need explanation degree. On the x-axis, 1.0 indicates a participant needs explanations for all the 25 MRI cases, and 0.0 indicates one needs explanations for none of the MRI cases.

3 A Note on Figure 1 in the Manuscript

We notice in panel (D) of Figure 1 in the manuscript, the tumor, with the data ID of BraTS20_Training_270, exhibits image features that resemble GBM: irregular tumor shape, enhanced tumor rim, and clearing tumor core indicating necrosis. In addition, as shown in Table 2, all 29 participants gave their judgment of GBM in the three conditions of DR, DR+AI, and DR+XAI. The AI model also predicted it as GBM with the explanation highlighting the enhanced rim of the tumor. However, the ground truth label for this case is grade II/III glioma. The ground truth label deviates from the common clinical knowledge of interpreting the glioma from MRI, and it may be caused by the following reasons: 1) This is a true grade II/III glioma. 2) There are noises in the ground truth labels of the dataset. 3) There may be errors in the histopathological process, such as sampling errors during biopsy. We can only raise our reasonable suspect but cannot confirm which reason was the case, as the biopsy was conducted previously and the case was anonymized when included in the BraTS public dataset.

We have verified that the following main study results did not change with the alternative ground truth label of BraTS20_Training_270 being GBM: 1) using Friedman test, there was still a significant difference in accuracy among the three conditions of DR, DR+AI, and DR+XAI; the accuracies of DR+AI and DR+XAI were significantly higher than DR, and there was no significant difference between DR+AI and DR+XAI. 2) complementary doctor-AI performance was not achieved for DR+AI or DR+XAI condition. This is because AI and the 29 participants had a uniform judgment in the three conditions for this case.

References

[1] Feature Ablation. Accessed: 2022-10-31.

[2] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009. Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

[3] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[4] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics.

[5] Weina Jin, Xiaoxiao Li, Mostafa Fatehi, and Ghassan Hamarneh. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Medical Image Analysis*, 84:102684, 2023.

[6] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. Evaluating explainable ai on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11945–11953, June 2022.

[7] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 1135–1144, New York, New York, USA, 2016. ACM Press.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[10] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org, 2017.

[11] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences, 2017.

[12] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[14] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017.

[15] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015.

[16] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

[17] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[18] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.