Supplementary Materials

Figure 8 on the next page shows the correlation between difficulty as perceived by a pre-trained model and pathologist annotators. Table 4 on the next page lists examples of medical image classification tasks from prior work where we believe annotator agreement data to be accessible.

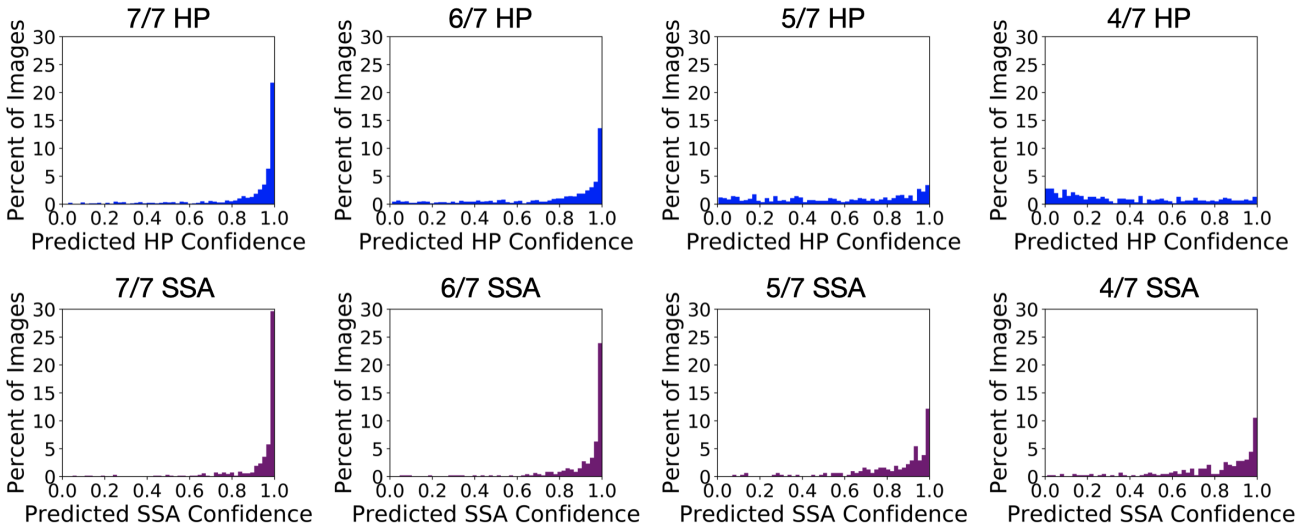

Figure 8: Distribution of predicted confidences by a pre-trained model fine-tuned on our dataset for different annotator agreement levels (indicated above each plot).

Dataset	Annotators	Resolution Method
Head CT Scans [18]	3	Majority Vote
Lymph Node Metastases [20]	3	Senior Expert Verification
Diabetic Retinopathy [23]	7	Majority Vote
Chest Radiograph [24]	3	Majority Vote
Lung Carcinoma [25]	3	Senior Expert Verification
Follicular Lymphoma Grading [27]	5	Majority Vote
Ulcer Recognition [28]	3	Senior Expert Verification
Colorectal Polyps [17]	5	Majority Vote
Breast Cancer [29]	3	Senior Expert Verification

Table 4: Examples of image classification tasks from prior work where annotator agreement is accessible in principle.