

SUPPLEMENTARY MATERIALS 6

Effect of Threshold Modification on Diagnostic Performance

Table 3 illustrates the result of per-lesion binary classification based on the following thresholds: 1) The class with the highest DLLD confidence score was considered to be the decision for the lesion; 2) lesions with a reader confidence score ≥ 3 were considered metastasis. In binary classification, threshold modification causes a change in the performance of the DLLD and reader.

Supplementary Table 3 demonstrates the results based on the following thresholds: 1) lesions with a DLLD metastasis confidence score ≥ 80 were considered metastasis; 2) lesions with a reader confidence score ≥ 4 were considered metastasis. The sensitivity of DLLD was not significantly different from that of the abdominal radiologists and radiology residents (69.70%, CI [60.64, 78.75]; 68.01%, CI [62.68, 73.35], $p = 0.750$; 64.31%, CI [58.83, 69.79], $p = 0.332$, respectively). The FPP of DLLD was higher than that of the abdominal radiologists and radiology residents (0.541, CI [0.433, 0.649]; 0.126, CI [0.085, 0.166], $p < 0.001$; 0.247, CI [0.194, 0.300], $p < 0.001$, respectively).

Supplementary Table 4 illustrates the results based on the following thresholds: 1) lesions with a DLLD metastasis confidence score of 100 were considered metastasis; 2) lesions with a reader confidence score of 5 were considered metastasis. The sensitivity of DLLD was not significantly different from that of the abdominal radiologists and radiology residents (37.37%, CI [27.84, 46.90]; 47.81%, CI [42.10, 53.53], $p = 0.068$; 37.04%, CI [31.51, 42.56], $p = 0.951$, respectively). The FPP of DLLD was not significantly different as compared to that of the abdominal radiologists and radiology residents (0.094, CI [0.032, 0.156]; 0.039, CI [0.015, 0.063], $p = 0.109$; 0.063, CI [0.033, 0.093], $p = 0.384$, respectively).