

Supplementary Information:
MEAI : An Artificial Intelligence Platform For Predicting Distant
And Lymph Node Metastases Directly From Primary Breast
Cancer

Contents

1 Supplementary formula description

2 Supplementary tables

Supplementary formula description

The equations used to calculate each evaluation indication in this article:

The confusion matrix contains information about four components:

True negative (TN), called the true negative rate, indicates the number of samples that are actually negative samples predicted to be negative samples.

False positive (FP), called the false positive rate, which indicates the actual number of negative samples predicted as positive samples.

False negative (FN), known as the false negative rate, indicating that the actual number of samples predicted to be negative samples is positive.

True positive (TP), called the true positive rate, indicates that the actual number of positive samples predicts the number of positive samples.

Accuracy:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}$$

Sensitivity:

$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity:

$$\text{Spec} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

PPV:

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

NPV:

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

AUROC: Area under receiver operating characteristic curves.

AUPRC: Area under precision-recall curve

Supplementary Tables

Supplementary Table 1. Performance of AI systems at the various slice sizes. We consider the effect of distinct patch sizes on model prediction accuracy. We report ACC, AUROC, specificity, sensitivity, PPV and NPV and 95% confidence intervals for the AI system taking different slice sizes on the test set (n = 187 WSIS).

Slide-size	ACC	AUC	Sensitivity	Specificity	PPV	NPV
Testing set (n = 187)						
Patch-size1 (128 × 128)	0.861 (0.853–0.869)	0.910 (0.898–0.922)	0.783 (0.773–0.793)	0.883 (0.871–0.895)	0.898 (0.890–0.906)	0.740 (0.730–0.750)
Patch-size2 (224 × 224)	0.898 (0.885–0.910)	0.934 (0.922–0.945)	0.909 (0.872–0.917)	0.879 (0.866–0.891)	0.907 (0.895–0.918)	0.882 (0.869–0.894)
Patch-size3 (256 × 256)	0.885 (0.872–0.898)	0.930 (0.914–0.946)	0.878 (0.853–0.903)	0.896 (0.882–0.910)	0.912 (0.900–0.924)	0.848 (0.824–0.872)
Patch-size4 (512 × 512)	0.886 (0.874–0.898)	0.931 (0.918–0.944)	0.914 (0.901–0.927)	0.843 (0.826–0.860)	0.881 (0.866–0.896)	0.893 (0.875–0.911)
Patch-size5 (1024 × 1024)	0.891 (0.878–0.904)	0.941 (0.936–0.946)	0.888 (0.871–0.905)	0.897 (0.875–0.919)	0.913 (0.906–0.920)	0.860 (0.843–0.877)

Supplementary Table 2. Additional metrics for the performance of no augmentation network. We validated the utility of our presented augmentation network module on the test set (n = 187 WSIs) of when removing the augmentation network from the AI system. All the prediction metrics of the system decreased.

Model / Pathologist	ACC	AUC	Sensitivity	Specificity	PPV	NPV
Testing set (n = 187)						
Model	0.898 (0.885–0.910)	0.934 (0.922–0.945)	0.909 (0.872–0.917)	0.879 (0.866–0.891)	0.907 (0.895–0.918)	0.882 (0.869–0.894)
Model without Augmentation	0.856 (0.843-0.869)	0.897 (0.880-0.914)	0.833 (0.818–0.848)	0.865 (0.843–0.887)	0.892 (0.872-0.912)	0.735 (0.713-0.757)

Supplementary Table 3 | Experience of Pathologists who participated in the Pathologist evaluation. We concluded the pathologists evaluated on the Hybrid set in terms of the number of years of experience in pathology research.

Pathologist	Years of experience
Pathologist 1	1
Pathologist 2	3
Pathologist 3	4
Pathologist 4	7
Pathologist 5	10
Pathologist 6	15

Supplementary Table 4 | Evaluation performance of pathologists. We calculated the specificity and sensitivity of the AI system, splitting its probabilistic predictions in two to match the pathologist-assessed sensitivity and specificity, respectively. We report the observed values and 95% confidence intervals for AUROC, ACC, specificity, sensitivity, PPV, and NPV achieved by the AI system and pathologists on a mixed set (n = 187 WSIs).

Model / Pathologist	ACC	AUC	Sensitivity	Specificity	PPV	NPV
Testing set (n = 187)						
Pathologist 1	0.708 (0.695–0.722)	0.769 (0.763–0.775)	0.744 (0.731–0.757)	0.646 (0.629–0.663)	0.752 (0.742–0.762)	0.635 (0.615–0.654)
Pathologist 2	0.726 (0.717–0.736)	0.788 (0.780–0.795)	0.751 (0.742–0.759)	0.684 (0.668–0.700)	0.772 (0.762–0.782)	0.653 (0.640–0.665)
Pathologist 3	0.748 (0.738–0.758)	0.822 (0.816–0.828)	0.770 (0.760–0.780)	0.709 (0.698–0.720)	0.788 (0.781–0.794)	0.682 (0.667–0.697)
Pathologist 4	0.735 (0.724–0.746)	0.807 (0.799–0.815)	0.755 (0.745–0.765)	0.700 (0.688–0.713)	0.782 (0.774–0.789)	0.661 (0.646–0.676)
Pathologist 5	0.771 (0.761–0.781)	0.830 (0.824–0.835)	0.792 (0.783–0.800)	0.735 (0.721–0.750)	0.804 (0.795–0.812)	0.716 (0.702–0.730)
Pathologist 6	0.786 (0.771–0.801)	0.850 (0.845–0.855)	0.806 (0.791–0.821)	0.751 (0.735–0.766)	0.813 (0.804–0.823)	0.740 (0.716–0.764)
Average	0.746 (0.726–0.769)	0.811 (0.788–0.834)	0.770 (0.750–0.790)	0.704 (0.674–0.734)	0.785 (0.767–0.803)	0.681 (0.649–0.713)
AI	0.898 (0.885–0.910)	0.934 (0.922–0.945)	0.909 (0.872–0.917)	0.879 (0.866–0.891)	0.807 (0.795–0.818)	0.782 (0.769–0.794)

Supplementary Table 5 | Performance of the Mixed diagnostic models. We presented the outcomes of the mixed diagnostic models, which combine the predictions of AI with each of the six pathologists on the mixed set (n = 208 WSIs), together with their values and 95% confidence intervals for ACC, AUROC, specificity, sensitivity, PPV, and NPV. The delta values represent the distinction between each mixed diagnostic model and its associated pathologist (P1–P6 with AI) and the 95% confidence intervals for each metric.

Model/ Pathologist	ACC	AUC	Sensitivity	Specificity	PPV	NPV
Mixed set (n = 208)						
Pathologist 1	0.778 (0.870–0.885)	0.813 (0.902–0.924)	0.780 (0.873–0.887)	0.735 (0.821–0.850)	0.804 (0.895–0.912)	0.716 (0.802–0.830)
Δ	0.069 (0.055–0.083)	0.044 (0.029–0.058)	0.036 (0.025–0.047)	0.089 (0.077–0.102)	0.052 (0.043–0.060)	0.081 (0.061–0.102)
Pathologist 2	0.782 (0.771–0.794)	0.818 (0.812–0.824)	0.793 (0.786–0.800)	0.745 (0.736–0.753)	0.813 (0.804–0.823)	0.741 (0.725–0.758)
Δ	0.056 (0.048–0.064)	0.030 (0.017–0.043)	0.043 (0.034–0.051)	0.061 (0.044–0.077)	0.041 (0.036–0.046)	0.089 (0.076–0.101)
Pathologist 3	0.796 (0.783–0.809)	0.831 (0.821–0.841)	0.795 (0.788–0.802)	0.764 (0.749–0.778)	0.825 (0.816–0.834)	0.753 (0.742–0.764)
Δ	0.061 (0.050–0.073)	0.024 (0.014–0.034)	0.040 (0.033–0.047)	0.063 (0.041–0.086)	0.043 (0.033–0.054)	0.092 (0.075–0.109)
Pathologist 4	0.804 (0.793–0.816)	0.849 (0.838–0.860)	0.813 (0.803–0.822)	0.788 (0.773–0.803)	0.838 (0.833–0.844)	0.769 (0.762–0.775)
Δ	0.056 (0.041–0.071)	0.027 (0.015–0.039)	0.043 (0.036–0.050)	0.079 (0.061–0.096)	0.051 (0.042–0.059)	0.086 (0.069–0.104)
Pathologist 5	0.822 (0.816–0.829)	0.854 (0.849–0.858)	0.820 (0.811–0.830)	0.808 (0.793–0.822)	0.847 (0.840–0.854)	0.783 (0.773–0.793)
Δ	0.051 (0.038–0.065)	0.024 (0.015–0.033)	0.029 (0.018–0.039)	0.072 (0.055–0.089)	0.043 (0.031–0.056)	0.067 (0.046–0.089)
Pathologist 6	0.827 (0.819–0.834)	0.873 (0.866–0.880)	0.832 (0.824–0.840)	0.823 (0.806–0.840)	0.860 (0.850–0.870)	0.795 (0.788–0.702)
Δ	0.041 (0.021–0.60)	0.024 (0.012–0.035)	0.026 (0.006–0.045)	0.072 (0.047–0.098)	0.047 (0.036–0.058)	0.055 (0.024–0.085)
Average	0.802 (0.786–0.818)	0.840 (0.821–0.858)	0.806 (0.790–0.821)	0.777 (0.749–0.805)	0.831 (0.814–0.848)	0.760 (0.736–0.782)
Δ	0.056 (0.040–0.072)	0.029 (0.010–0.047)	0.036 (0.020–0.051)	0.073 (0.045–0.101)	0.046 (0.029–0.063)	0.078 (0.055–0.101)

AI	0.899 (0.883–0.915)	0.938 (0.919–0.956)	0.906 (0.890–0.921)	0.877 (0.849–0.905)	0.931 (0.914–0.948)	0.860 (0.836–0.882)
----	------------------------	------------------------	------------------------	------------------------	------------------------	------------------------

Supplementary Table 6 | Results of the analysis of various subgroups. We compared the number of correctly identified malignant lesions between AI systems and radiologists. In this analysis, we linked 100 WSIS (40 ME, 60 Non-ME) in the mixed set to biopsy-confirmed malignant findings. We stratified patients by metastatic site, age, and biomarker characteristics. For each stratum, we reported the total number of patients (n), identified by AI as the number of patients with distant and Lymph Node cancer metastases. We split the AI's probabilistic predictions by matching the radiologists' average specifications in the Pathologist evaluation.

Lesion characteristics	n	AI	Pathologists (mean ± std)
Ages			
<50	42	37	31.3 ± 2.0
>50	58	52	43.2 ± 1.2
Metastasis site			
Bone	5	4	2.8 ± 0.3
Lymph	10	8	7.4 ± 0.2
Lung	8	7	5.9 ± 0.4
Liver	8	7	5.4 ± 0.9
Other	9	8	6.7 ± 1.2
Non-metastasis	60	54	44.7 ± 3.5
Biomarkers of Invasive Cancers			
ER/PR-positive, HER2negative	58	52	43.2 ± 1.8
HER2-positive	32	26	23.8 ± 1.4
ER/PR/HER2-negative	7	5	4.2 ± 0.5

Supplementary Table 7. Differential prediction performance of artificial intelligence systems for distant versus lymph node metastases We tested the predictive capabilities of the AI system on the test set (n = 187 WSIS) for Lymph Node metastasis and distant metastasis of primary breast cancer, respectively. One test set contained 53 patients with distant metastasis and 22 patients with Lymph Node metastasis.

Classification	N	Mean	ACC
Lymph Node metastases	22	19.9	0.908 (0.898–0.918)
Distant metastases	53	46.0	0.869 (0.853–0.885)