MAYO CLINIC

Clinical Validation and Implementation of an Artificial Intelligence Model for Digital Analysis of Ki-67 Biomarker in Breast Cancer

Kristina del Rosario[1], Leslie Stoy[1], Sami Blom[2], Niina Vaheri[2], Eric Korman[1], Amy Plagge[1], Christian Welke[1], Saba Yasir[1], Zongming Eric Chen[1]

[1]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN; [2]Aiforia Technologies®, Helsinki, Finland

OBJECTIVES

- Successfully implement an AI model for Ki-67 IHC in breast cancer
- Establish a validation template and limitations for AI models on biomarker IHC slides

METHODS

- An AI model comprised of 4 convolutional neural networks for Ki-67 IHC was trained in Aiforia Create using scanned images from Leica GT450s and a defined ground truth
- Over 200 WSI were used for refinement of ground truth and training of the AI model, and for validation/verification in the Aiforia Clinical platform
- Four accuracy (table 1) and five precision studies (table 2) were performed in a clinical validation to ensure satisfactory performance of tumor detection and cell classification layers
- Scoring criteria used during visual review for accuracy of tumor detection (Figure 3):
  - 0=no significant errors
  - 1= minor errors that would not significantly change the Ki-67 result, minor edits may be necessary
  - 2=moderate errors that have the possibility to significantly change the Ki-67 result, manual edits change result but does not change clinically significant category
  - 3=severe errors or errors that would significantly change the Ki-67 result, manual edits change the result category

RESULTS

ACCURACY

Table 1. Summary of Accuracy Studies

Study Description	r^2	Pearson	F1	Average (0-3 scale)
Accuracy of results vs previously validated test (n=62) (Fig 2)	0.95	0.97	-	-
Accuracy of AI tumor detection vs manual identification (n=25 areas) (Fig 1)	-	-	0.87	-
Accuracy of tumor detection vs whole slide visual review (n=62) (Fig 3)	-	-	-	0.38
Accuracy of cell classification vs manual scoring (n=24)	0.95	0.97	-	-

- Clinical verification was also performed and achieved an r^2=0.99 for both accuracy (n=20) and precision (n=10).

Figure 1. Validation in Aiforia to determine F1 score for tumor identification (F1=0.87). F1 was calculated using manually drawn areas as gold standard (dotted lines) and compared areas the AI model identifies as tumor (darker purple areas).

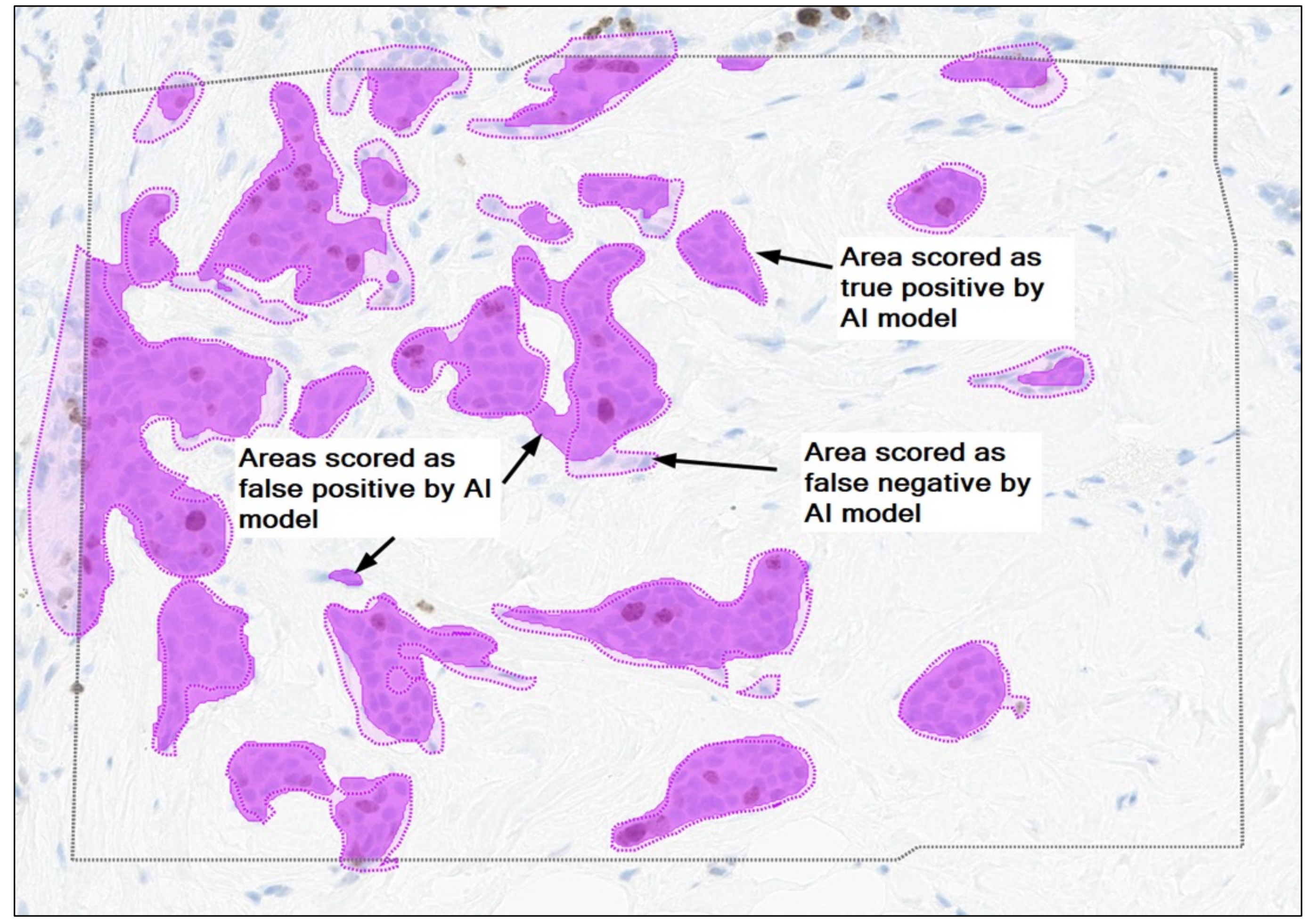

Area scored as true positive by AI model
Areas scored as false positive by AI model
Area scored as false negative by AI model

Figure 2
Percentage of Positive Staining Ki-67 Nuclei

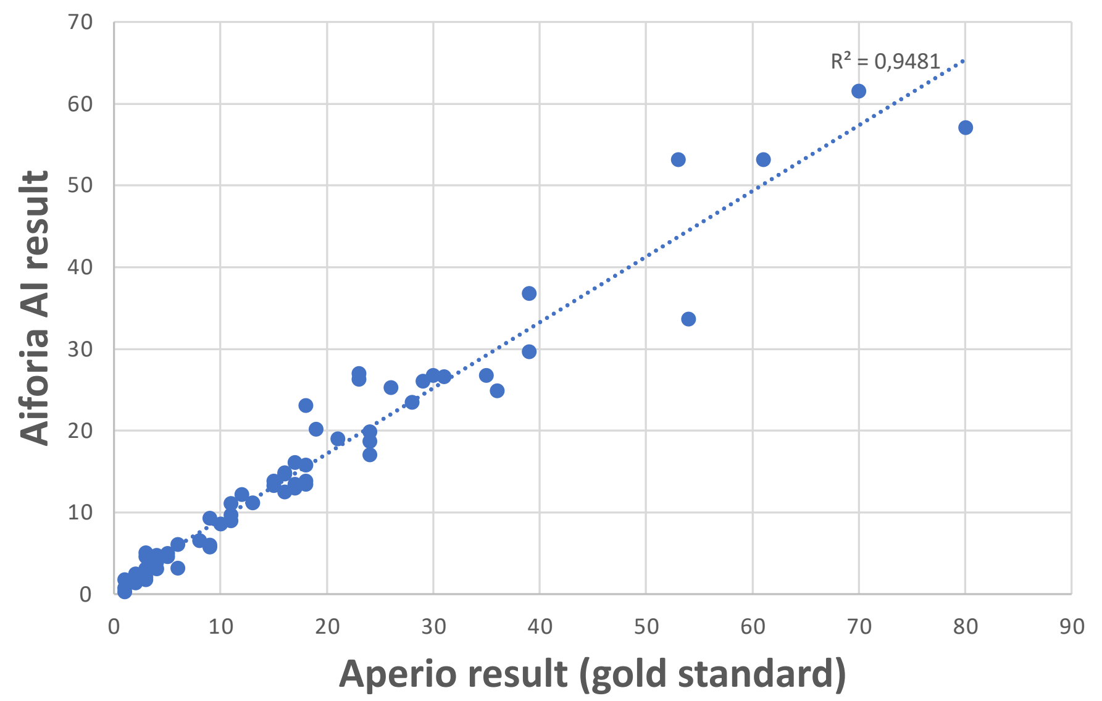
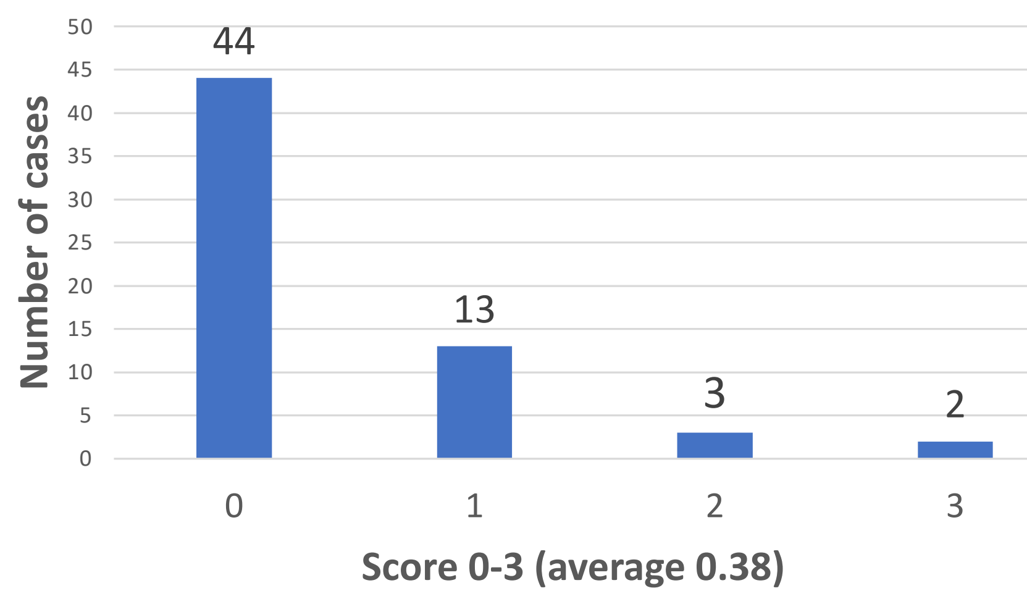
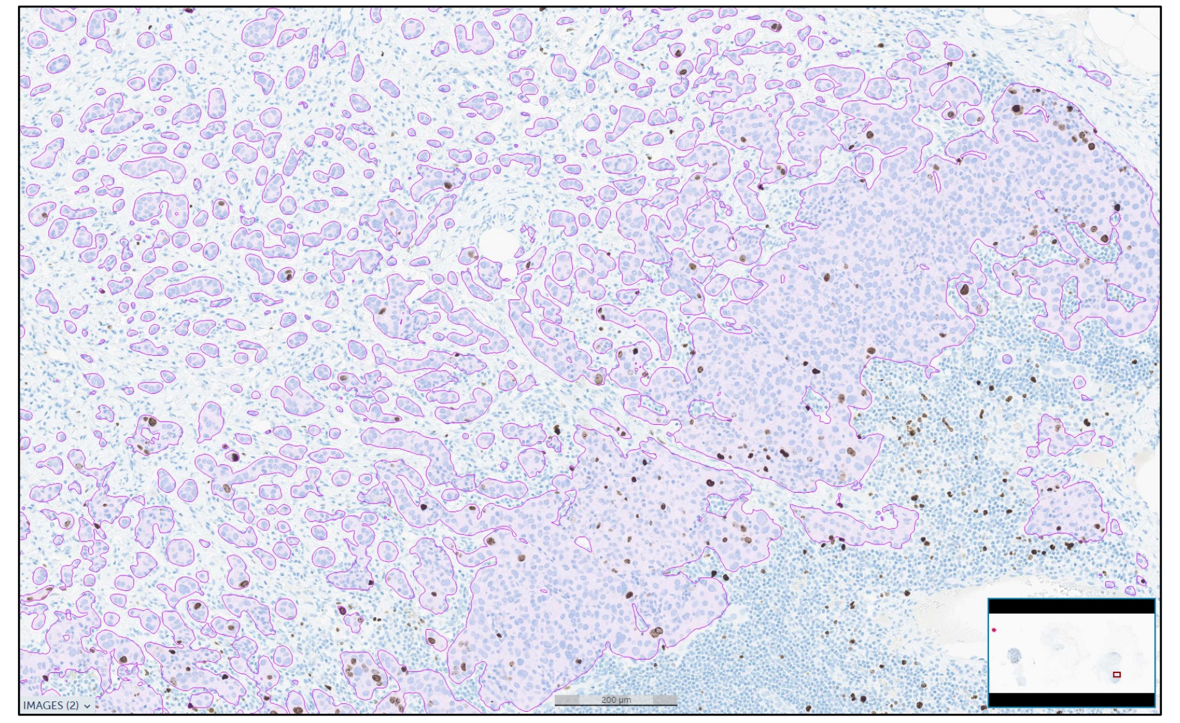


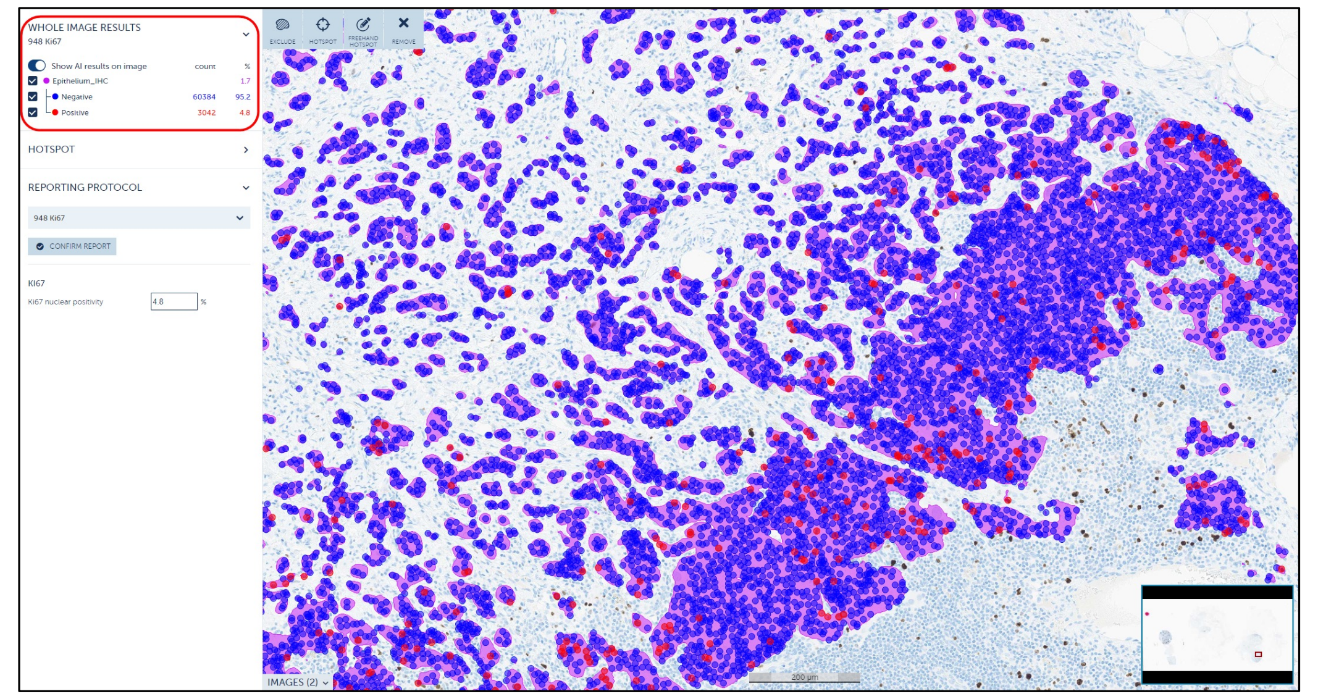
Figure 3
Visual Review of AI model's tumor detection



Tumor detection of AI model on metastatic breast cancer in a lymph node. Pink areas classified at tumor by the model.



Positive and negative cells identified by the AI model in tumor regions.



WHOLE IMAGE RESULTS
948 Ki67

		count	%
☑	Show AI results on image		
☑	Epithelium_IHC		1.7
☑	Negative	60384	95.2
☑	Positive	3042	4.8

HOTSPOT

REPORTING PROTOCOL

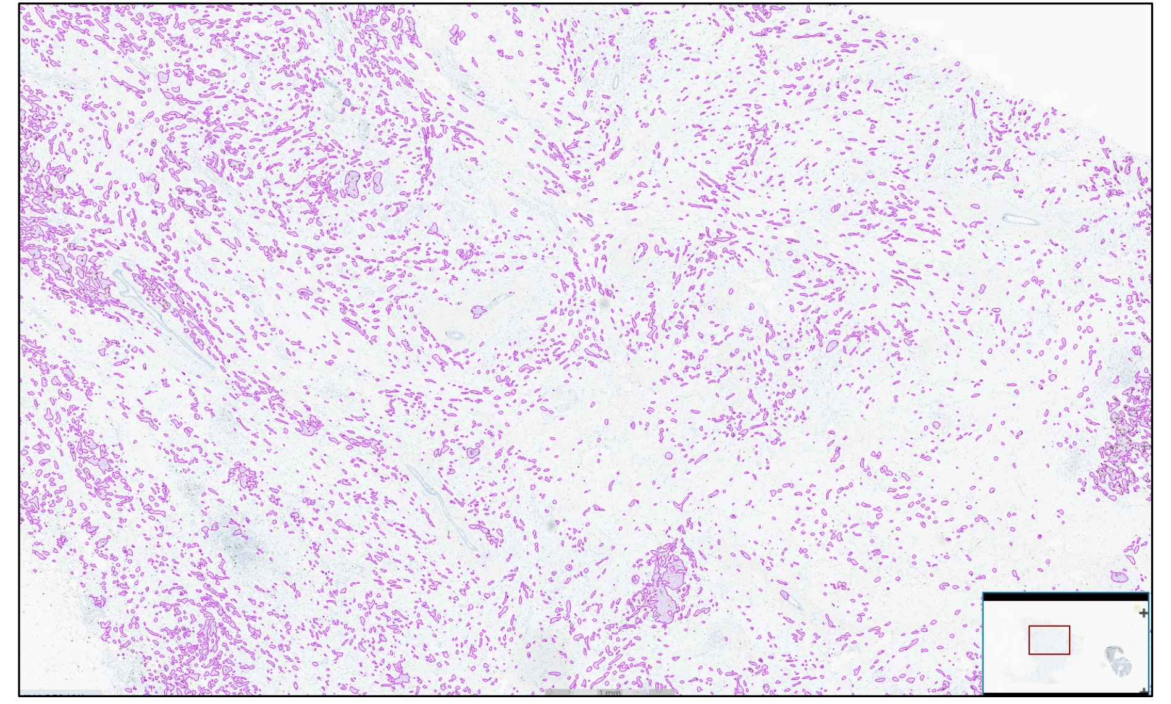
948 Ki67

CONFIRM REPORT

KI67

Ki67 nuclear positivity 4.8 %

Diffuse lobular tumor cells would make manually annotating difficult without tumor detection. Pink areas classified as tumor by AI model

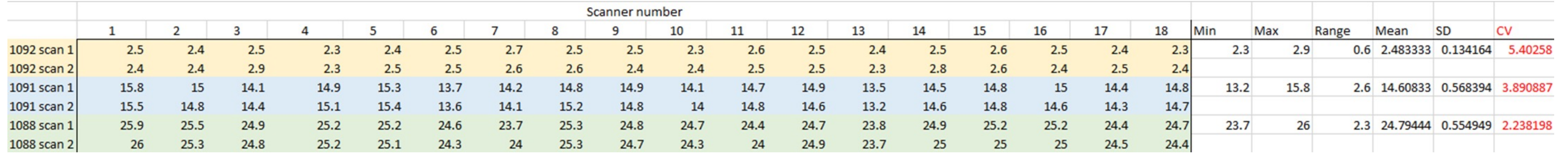


PRECISION

Table 2. Summary of Precision Studies

Study Description	r^2	Pearson	CV
Precision of AI model itself	1	1	-
Interobserver precision	0.98	0.99	-
Intraobserver precision	0.99	0.99	-
Multiple scanner precision (table 3)	-	-	2.2-5.4
Same scanner precision	-	-	0.5-3.9

Table 3. Percent positive Ki67 staining from images scanned on 18 different scanners twice of 3 glass slides. These results were also used to calculate acceptable ranges (3 standard deviations) for ongoing maintenance/evaluations of scanners.



	Scanner number																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Min	Max	Range	Mean	SD	CV
1092 scan 1	2.5	2.4	2.5	2.3	2.4	2.5	2.7	2.5	2.5	2.3	2.6	2.5	2.4	2.5	2.6	2.5	2.4	2.3	2.3	2.9	0.6	2.483333	0.134164	5.40258
1092 scan 2	2.4	2.4	2.9	2.3	2.5	2.5	2.6	2.6	2.4	2.4	2.5	2.5	2.3	2.8	2.6	2.4	2.5	2.4						
1091 scan 1	15.8	15	14.1	14.9	15.3	13.7	14.2	14.8	14.9	14.1	14.7	14.9	13.5	14.5	14.8	15	14.4	14.8	13.2	15.8	2.6	14.60833	0.568394	3.890887
1091 scan 2	15.5	14.8	14.4	15.1	15.4	13.6	14.1	15.2	14.8	14	14.8	14.6	13.2	14.6	14.8	14.6	14.3	14.7						
1088 scan 1	25.9	25.5	24.9	25.2	25.2	24.6	23.7	25.3	24.8	24.7	24.4	24.7	23.8	24.9	25.2	25.2	24.4	24.7	23.7	26	2.3	24.79444	0.554949	2.238198
1088 scan 2	26	25.3	24.8	25.2	25.1	24.3	24	25.3	24.7	24.3	24	24.9	23.7	25	25	25	24.5	24.4						

DISCUSSION

- Software tools were developed to fit our complex workflow at Mayo Clinic including the ability to make manual adjustments when necessary
- Acceptance criteria for manual vs AI tumor detection may benefit from further investigation
  - ➢ Feasibility of manually annotating tumor cells vs the AI model can be problematic
  - ➢ Question whether an F1 score based on area is meaningful vs. a cell count
- Variables to consider that may affect the AI model:
  - ✓ Scanners used
  - ✓ Compression rates/image quality
  - ✓ Pixel size of images
  - ✓ Variety/number of images used in training and percentage of them with either true or artificial augmentation performed

CONCLUSIONS

- In depth studies were successful in validating an AI model for clinical use
- These validation studies serves the purpose as a guide for future biomarker AI models
- Careful planning by the laboratory and close collaboration with the AI provider is required
- Ground truth refinement and software adaption is crucial
- Final verification in an environment identical to clinical production is recommended

aiforia
AI for image analysis