Laboratory Development Test Validation of H&E-Based Deep Learning Model to Predict Biomarkers in Endometrial Cancer

Qiyuan Hu[1], Kshitij Ingale[1], Sun Hae Hong[1], Rohan P. Joshi[1], Jacob W. H. Gordon[1], Yoni Muller[2], Ben Terdich[1], Jason Blue-Smith[1], Ryan Jones[1], Nike Beaubier[1], Chithra Sangli[1], Riccardo Miotto[1]
[1]Tempus AI, Inc., Chicago, IL, [2]Work done while at Tempus AI, Inc., Chicago, IL

Disclosure Statement: All authors are current or former employees of Tempus AI, Inc. All current employees hold stock in the company.

INTRODUCTION

Next-generation sequencing (NGS) can provide critical information for selecting appropriate therapies for cancer patients, improving outcomes. However, NGS is not always performed, especially when not indicated in clinical guidelines. We developed a digital pathology model to predict biomarkers in endometrial cancer, including MSI-High status and pathogenic or likely pathogenic alterations in CTNNB1, POLE, PTEN, and TP53 from digital hematoxylin and eosin (H&E) whole slide images (WSIs) and validated the model using a CAP/CLIA laboratory-developed test (LDT) based validation study.

METHODS

The model was trained and tuned using >4000 WSIs (4,809 patients) and validated on >300 WSIs (388 patients) of endometrial cancer. The ground truths for biomarker status was established using the Tempus xT DNA-seq assay. We trained a multi-task attention-based multiple instance learning model using H-optimus-0 foundation model embeddings, to predict biomarkers. During LDT validation, model was evaluated for analytical sensitivity, specificity, precision and accuracy with predefined acceptance criteria. Root mean squared error (RMSE), Lin's concordance correlation coefficient (CCC) and Pearson's correlation (r) were used to evaluate model robustness to tissue area, artifacts and scanners. Area under receiver operating curve (ROC-AUC) was used to evaluate prediction task

Figure 1. Deep learning model schematic

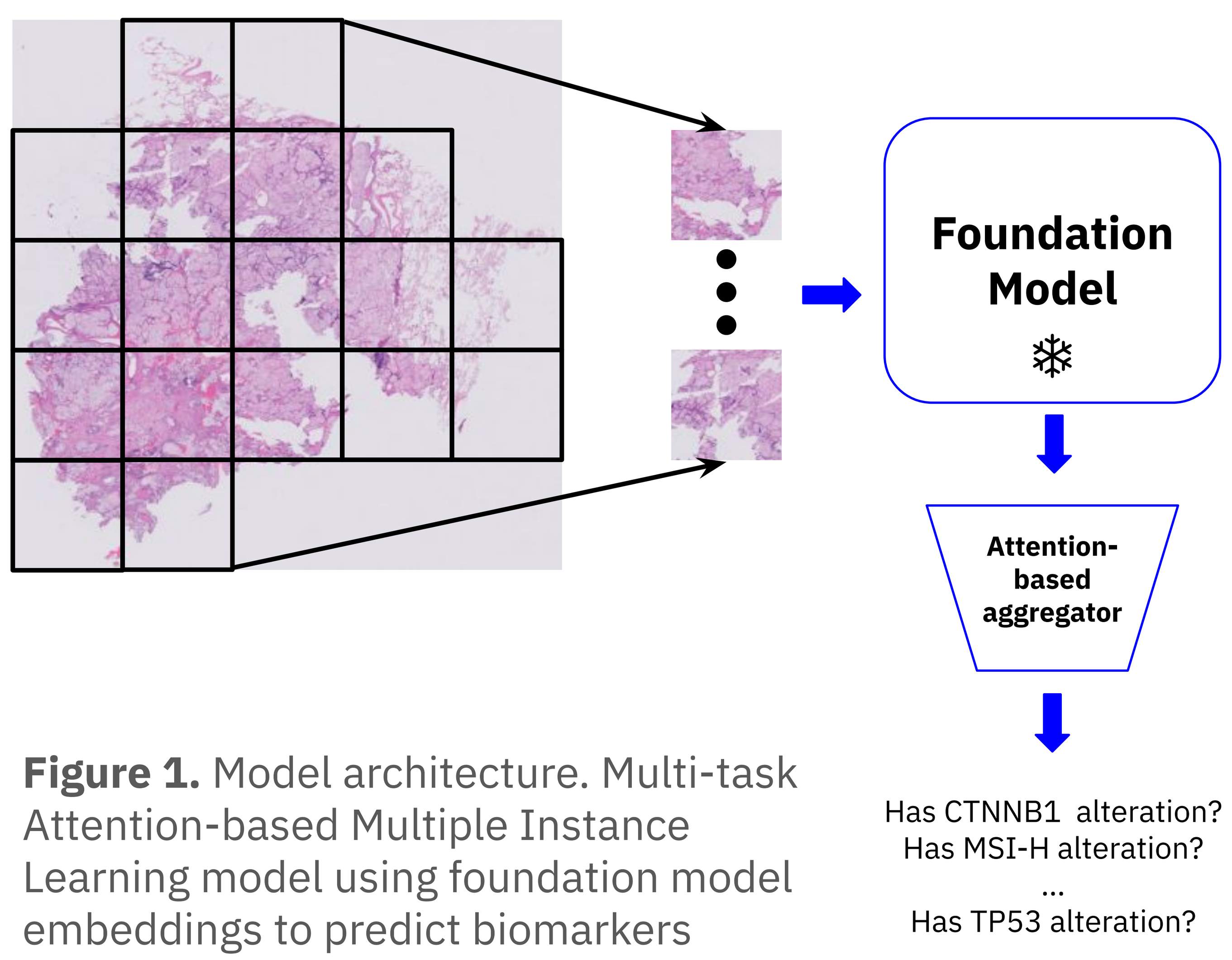

Figure 1. Model architecture. Multi-task Attention-based Multiple Instance Learning model using foundation model embeddings to predict biomarkers

RESULTS

Table 1. LDT validation experiments

Study	Acceptance Criteria	Results
Analytical sensitivity	RMSE < 0.1 between predictions on subsampled regions vs. entire slides	LoD = 1.08 mm^2, RMSE=0.67~0.90
Analytical specificity	RMSE < 0.1 between predictions on perturbed vs. original slides	Maximum artifact = 20%, RMSE=0.17~0.63
Analytical precision	RMSE < 0.1 between predictions on the intra/inter-scanner rescans vs. original scans	Intra-scanner: RMSE = 0.035~0.059
		Inter-scanner: RMSE = 0.085~0.095
Analytical accuracy	AUC significantly predictive	AUC = 0.82 ~ 0.93 (Figure 2)

Table 1. Studies evaluating model in the validation along with their corresponding acceptance criteria.

- **Analytical sensitivity:** This study identifies a robust Limit of Detection (LoD) for the model by simulating smaller tissue regions fed to the model. For this study, we chose 1, 2, 5, 10, 20, 50, 100 tiles as input and compared the model score against whole slide inference score
- **Analytical specificity:** This study evaluates model robustness to artifacts by simulating blur and color perturbation on the H&E image. For this study, we perturbed 1%, 2%, 5%, 10%, 20% area with color or blur artifacts and compared the model score against unperturbed image score.
- **Analytical precision:** This study explored model reproducibility and repeatability by comparing model scores on scans generated using different scanners (inter-scanner) and scans using same scanner at different time points (intra-scanner) respectively
- **Analytical accuracy:** This study aims to quantify model performance in terms of ability to predict target based on ROC-AUC

Table 2. Analytical accuracy results based on operating point

Biomarker	Cohort Prevalence (N=388)*	Operating point	Sensitivity	Specificity
CTNNB1	27% (17%)	Specificity 85%	84%	84%
MSI-High	20% (22%)	Specificity 90%	40%	91%
POLE	6% (2%)	Sensitivity 85%	78%	72%
PTEN	38% (47%)	Specificity 85%	75%	80%
TP53	55% (50%)	Specificity 85%	89%	85%

Table 2. Operating point based metrics for all biomarkers. Predictions were obtained from model scores using thresholds determined with a tuning set. Model achieves higher PPV relative to real-world prevalence demonstrating potential for patient stratification
*The validation test set was enriched for positive cases for rare biomarkers

SUMMARY

- Our model successfully predicts actionable biomarkers in Endometrial cancer using H&E images
- All of the validation criteria evaluating robustness to various deployment scenarios are satisfied by model
- Model performance also demonstrates potential for effective triaging to encourage confirmatory testing for patients likely to exhibit genomic alterations

Figure 2. Analytical accuracy study

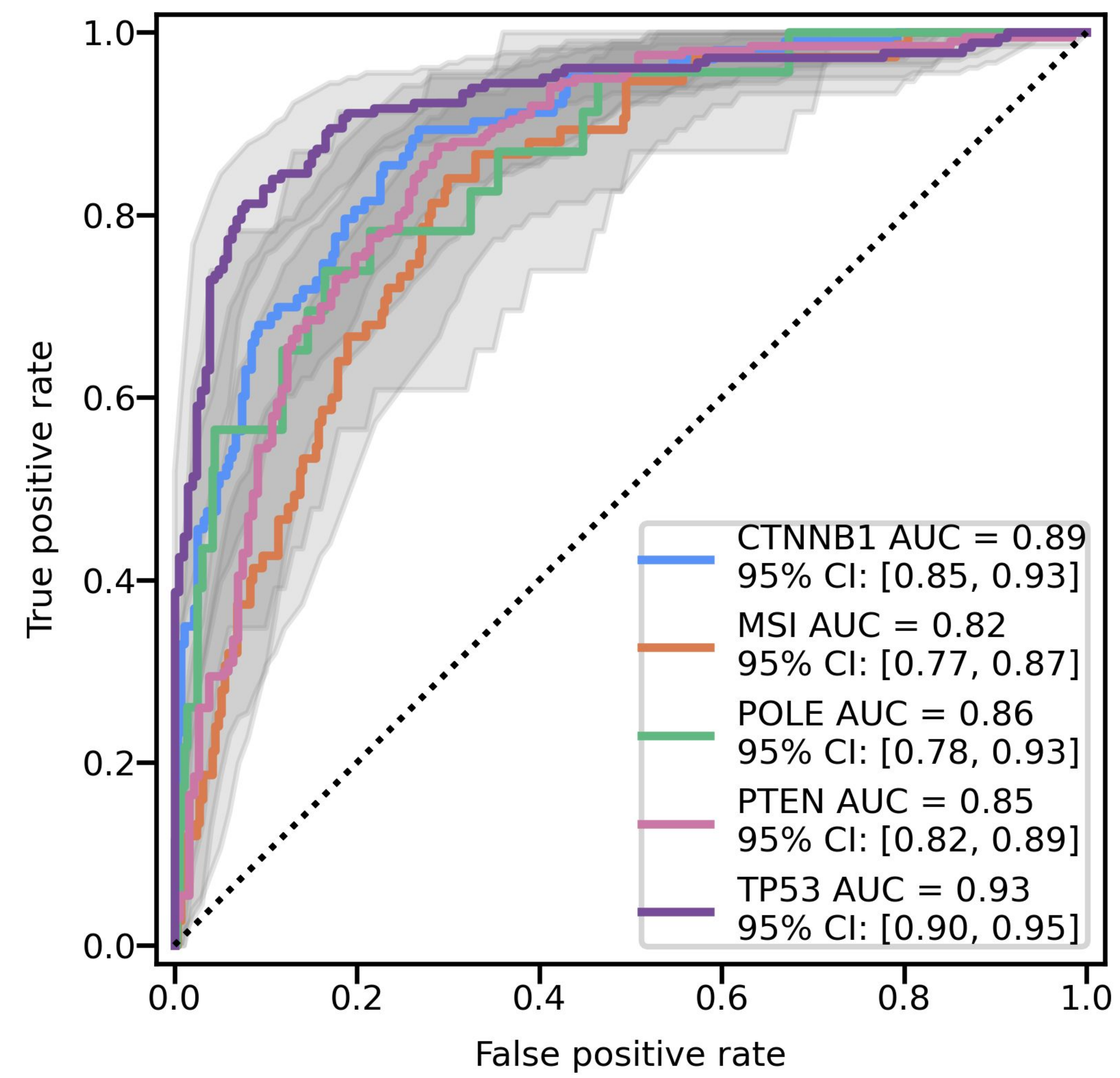

Figure 2. Receiver Operating characteristic curve for all biomarkers. Model is significantly predictive of biomarker status with AUCs around 0.82 to 0.93

ACKNOWLEDGMENTS
We thank Dana DeSantis from the Tempus Science Communications team for poster development.

Correspondence: isabelle.hu@tempus.com