השוואה בין מודל למידת מכונה לקריטריוני G-ROP בסקר לרטינופתיה של פגות מסוג 1: ולידציה מקומית והפחתת עומס בדיקות

Hadas Ben-Eli*[1,2], Ayelet Goldstein*[3], Edna Granit[1], Smadar Eventov-Friedman[4], Noa Ofek Shlomai[4], Sinan Abu Leil[4], Milka Matanis-Suidan[1], Hadas Mechoulam

[1]Department of Ophthalmology, Hadassah-Hebrew University Medical Center;
[2]Department of Optometry and Vision Science, Jerusalem Multidisciplinary College;
[3]Department of Computer Science, Jerusalem Multidisciplinary College;
[4]Department of Neonatology, Hadassah Medical Center, Faculty of Medicine, Hebrew University of Jerusalem

Introduction:
Retinopathy of prematurity (ROP) is a major cause of preventable childhood blindness. While current screening guidelines such as the North-American - American Academy of Ophthalmology (NA-AAO) and Growth and ROP (G-ROP) criteria[1] offer high sensitivity, they also lead to numerous unnecessary examinations[2]. Unnecessary examinations are resource-intensive, costly, stressful for infants and parents, and potentially harmful. American

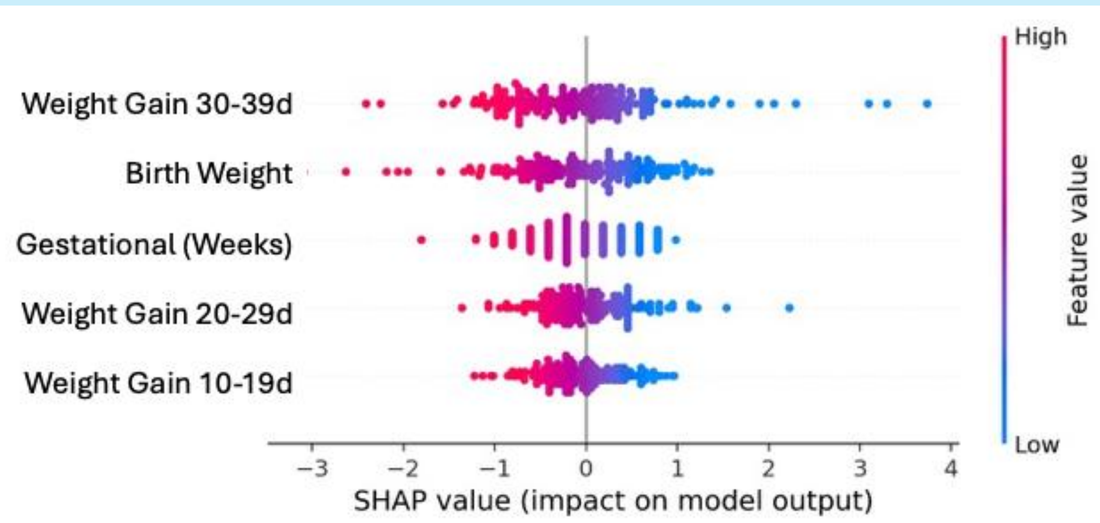
Objective:
To evaluate whether a locally developed, interpretable machine learning (ML) model could substantially reduce the ROP screening burden while maintaining 100% sensitivity for detecting Type 1 ROP in a contemporary local cohort

Methods:
We retrospectively evaluated 2,159 infants admitted to Hadassah Medical Center (Jan 2023–Feb 2025); 522 met inclusion criteria (gestational age ≤33 weeks or birth weight ≤1500 g), and 193 had complete data for analysis. We compared NA-AAO and G-ROP criteria with a logistic regression–based ML model using gestational age, birth weight, and postnatal weight gain. Performance was assessed via leave-one-out cross-validation; interpretability was evaluated using SHapley Additive exPlanations (SHAP)

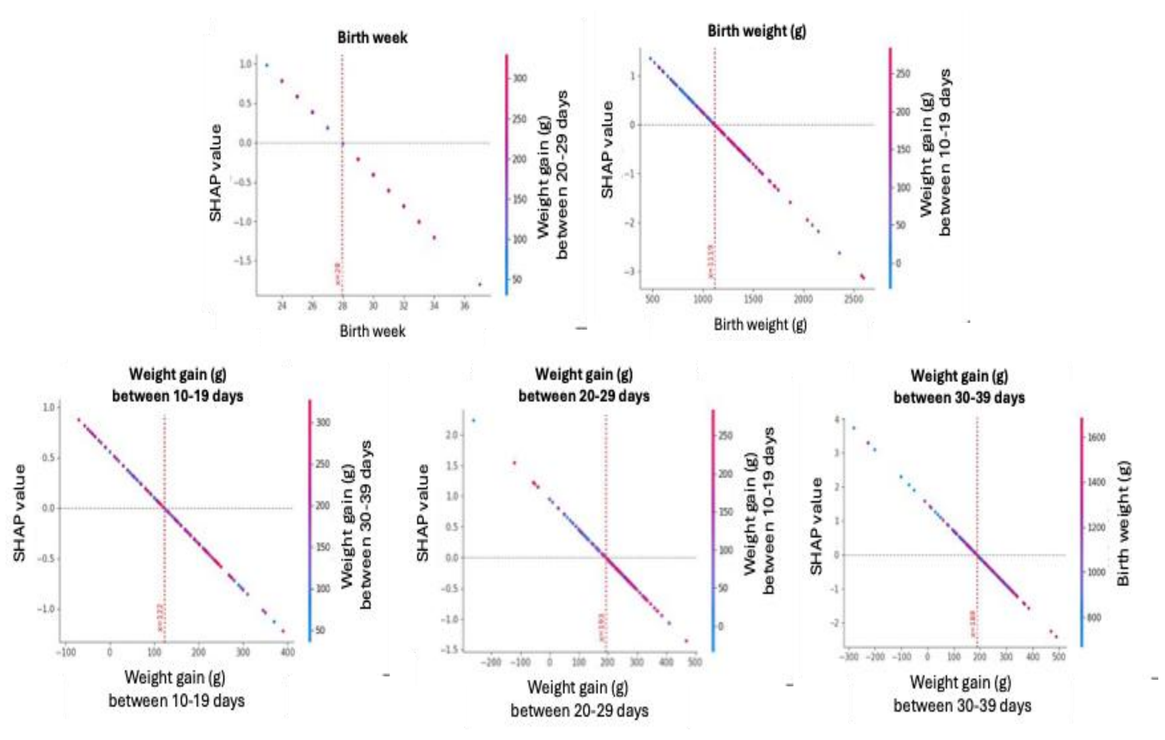
Results
Six infants developed type 1 ROP. Both NA-AAO and G-ROP criteria achieved 100% sensitivity. G-ROP improved specificity to 28% and reduced screening burden by 26% relative to local practice (97.9% to 72.5%). The ML model also preserved 100% sensitivity but improved specificity to 82.9%. It reduced the number of infants flagged for screening from 179 to 38 vs. NA-AAO (−78.8%), 140 to 38 vs. G-ROP (−72.9%), and 189 to 38 vs. local practice (−79.9%). SHAP confirmed reliance on clinical features and improved risk integration.

Model	Flagged (n)	% of 193	% Change vs. Local (n=189)	% Change vs. All (n=193)
Current Criteria (GA<=32 weeks OR BW<1501 g)	189	97.9%	-	-2.1%
NA-AAO Criteria (GA<=30 weeks OR BW <1501 g)	179	92.7%	+5.2%	-7.3%
G-ROP	140	72.5%	-25.4%	-27.5%
Suggested ML Model (LR)	38	19.7%	-78.2%	-80.3%

Number of infants flagged for ROP screening and reduction in screening by each method (N=193). GA=gestational age. BW=birth weight. LR=Logistic Regression.



SHAP summary plot showing global feature importance for type 1 ROP prediction. Dots represent individual infants, with color indicating the feature value (red = high, blue = low).



SHAP dependence plots for (A) birth week, (B) birth weight, (C) weight gain 10–19 days, (D) weight gain 20–29 days, and (E) weight gain 30–39 days. Each dot represents a patient. SHAP value (y-axis) shows the feature's contribution to risk prediction; vertical lines indicate thresholds identified by the model.

	ML Model	G-ROP
Gestational age	<28 weeks	<28 weeks
Birth weight	<1119	<1051
Weight gain between PNA 10-19d	<122	<120
Weight gain between PNA 20-29d	<193	<180
Weight gain between PNA 30-39d	<188	<170

Threshold comparison between G-ROP criteria and SHAP-derived cutoffs from the ML model. Thresholds are shown for gestational age, birth weight, and postnatal weight gain across specified PNA intervals.

Conclusions / Relevance
A locally validated, interpretable ML model maintained perfect sensitivity with significantly fewer screenings. These results support ML-based, population-specific screening to optimize ROP care.

Contact details: Dr. Hadas Ben-Eli
hadasben@jmc.ac.il; hadasben@hadassah.org.il
Disclosure: None