AMERICAN ACADEMY OF OPHTHALMOLOGY®

Predicting Choroidal Nevus Transformation to Melanoma Using Machine Learning

Prashant D. Tailor, MD,[1] Piotr K. Kopinski, MD, PhD,[1] Haley S. D'Souza, MD,[1] David A. Leske, MS,[1] Timothy W. Olsen, MD,[1] Carol L. Shields, MD,[2] Jerry A. Shields, MD,[2] Lauren A. Dalvin, MD[1]

Purpose: To develop and validate machine learning (ML) models to predict choroidal nevus transformation to melanoma based on multimodal imaging at initial presentation.

Design: Retrospective multicenter study.

Participants: Patients diagnosed with choroidal nevus on the Ocular Oncology Service at Wills Eye Hospital (2007−2017) or Mayo Clinic Rochester (2015−2023).

Methods: Multimodal imaging was obtained, including fundus photography, fundus autofluorescence, spectral domain OCT, and B-scan ultrasonography. Machine learning models were created (XGBoost, LGBM, Random Forest, Extra Tree) and optimized for area under receiver operating characteristic curve (AUROC). The Wills Eye Hospital cohort was used for training and testing (80% training−20% testing) with fivefold cross validation. The Mayo Clinic cohort provided external validation. Model performance was characterized by AUROC and area under precision−recall curve (AUPRC). Models were interrogated using SHapley Additive exPlanations (SHAP) to identify the features most predictive of conversion from nevus to melanoma. Differences in AUROC and AUPRC between models were tested using 10 000 bootstrap samples with replacement and results.

Main Outcome Measures: Area under receiver operating curve and AUPRC for each ML model.

Results: There were 2870 nevi included in the study, with conversion to melanoma confirmed in 128 cases. Simple AI Nevus Transformation System (SAINTS; XGBoost) was the top-performing model in the test cohort [pooled AUROC 0.864 (95% confidence interval (CI): 0.864−0.865), pooled AUPRC 0.244 (95% CI: 0.243−0.246)] and in the external validation cohort [pooled AUROC 0.931 (95% CI: 0.930−0.931), pooled AUPRC 0.533 (95% CI: 0.531−0.535)]. Other models also had good discriminative performance: LGBM (test set pooled AUROC 0.831, validation set pooled AUROC 0.815), Random Forest (test set pooled AUROC 0.812, validation set pooled AUROC 0.866), and Extra Tree (test set pooled AUROC 0.826, validation set pooled AUROC 0.915). A model including only nevi with at least 5 years of follow-up demonstrated the best performance in AUPRC (test: pooled 0.592 (95% CI: 0.590−0.594); validation: pooled 0.656 [95% CI: 0.655−0.657]). The top 5 features in SAINTS by SHAP values were: tumor thickness, largest tumor basal diameter, tumor shape, distance to optic nerve, and subretinal fluid extent.

Conclusions: We demonstrate accuracy and generalizability of a ML model for predicting choroidal nevus transformation to melanoma based on multimodal imaging.

Financial Disclosures: Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2025;5:100584 © 2024 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Choroidal nevus, the most common benign intraocular tumor, presents a unique clinical challenge because of the potential for transformation into malignant melanoma.[1] The ability to accurately predict transformation holds significant implications for patient treatment strategy, ocular morbidity, prognosis, and mortality. Multimodal imaging techniques such as fundus photography, fundus autofluorescence (AF), spectral domain OCT, and B-scan ultrasonography have enhanced the characterization of choroidal nevi, facilitating nuanced understanding of their clinical behavior.[2] The combination of these techniques captures a diverse range of anatomical and functional characteristics of nevi and has led to the development of multivariate risk tools to help quantify the risk of transformation.[2] However, the information-rich output of these imaging modalities remains underutilized, primarily because of the complexity of the data generated and the consequent challenge in effective analysis.

In recent years, artificial intelligence (AI) has come to the forefront as a tool to analyze complex multivariate relationships. New machine learning (ML) algorithms offer a novel approach to analyzing complex classification problems and have demonstrated strong performance in diverse medical applications.[3,4] However, their potential in predicting the transformation of choroidal nevus to melanoma based on multimodal imaging data has remained largely unexplored.[5]

To address this need, we developed a ML model, Simple AI Nevus Transformation System (SAINTS), to predict choroidal nevus transformation risk to melanoma using

https://doi.org/10.1016/j.xops.2024.100584
ISSN 2666-9145/24

tabular multimodal imaging data in a multicenter study with both internal and external validation to enhance generalizability. Given the black box nature of many ML algorithms, we also sought to identify novel risk factors for transformation to melanoma identified by the ML models. Furthermore, we will investigate the efficacy and generalizability of SAINTS to predict choroidal nevus behavior for potential integration into clinical practice.

Methods

This project focuses on a supervised binary classification task aimed at predicting the risk of choroidal nevus conversion into melanoma based on initial presenting features. The project analyzed tabular multimodal imaging data, including fundus photography, fundus AF, spectral domain OCT, and B-scan ultrasonography. This study was approved by the Institutional Review Board/Ethics Committees of Mayo Clinic and Wills Eye Hospital and adhered to the tenets of the Declaration of Helsinki and Health Insurance Portability and Accountability Act. All patients provided informed consent.

Detailed methodology regarding the training and testing set from the Wills Ocular Oncology service has previously been described.[2] The external validation cohort was from a large tertiary referral center (Mayo Clinic) and included a total of 514 nevi. If an eye contained more than 1 nevus, only the largest nevus was included. Fundus photography was performed with Zeiss camera (Carl Zeiss Meditec Inc) at Wills Eye Hospital and the Topcon camera (Topcon Healthcare) at Mayo Clinic. Fundus AF was performed with special filters (580-nm excitation, 695-nm barrier filter) to avoid imaging the AF of the crystalline lens. Autofluorescence features included presence or absence of hyperautofluorescence (lipofuscin). The OCT used enhanced depth imaging technology and was performed through a dilated pupil (Heidelberg Spectralis HRAOCT; Heidelberg Engineering). The OCT findings quantified total subretinal and subfoveal fluid. Additional OCT features included retinal pigment epithelium (RPE) (trough), the presence of choroidal neovascularization and retinal invasion. Ultrasonography was performed with standard A-scan and B-scan imaging of the intraocular mass using Eye Cubed (Ellex) technology at Wills Eye Hospital or the ABSolu A/B/S/UBM Ultrasound Platform from Quantel Medical at Mayo Clinic. Ultrasound features included B-scan tumor configuration, acoustic quality, and standardized A-scan internal reflectivity. All melanocytic lesions were examined by an expert ocular oncologist (C.L.S., J.A.S., J.S.P., T.W.O., L.A.D.) with multimodal imaging and categorized as choroidal nevus or melanoma based on prior definitions.[6] For larger lesions categorized as choroidal nevi, longitudinal follow-up, absence of other high-risk features, and presence of features of chronicity such as drusen and/or fibrous metaplasia were often used to determine nevus vs. melanoma.

Data Abstraction, Processing, and Outcome Definition

Data were manually extracted from the electronic medical record for the training and test sets as previously described.[2] Data from the external validation set (Mayo Clinic) were extracted from a prospectively collected database (The Prospective Ocular Tumor Study). The ML algorithms (see below) utilized automatically handle missing data by incorporating all available data without ignoring records with missing information, so no imputation methods were used for missing data. The outcome, or label, was growth to melanoma as defined by an expert ocular oncologist (C.L.S., J.A.S., J.S.P., T.W.O., and L.A.D.) via documented growth (defined as $\geq$0.5 mm in $\leq$24 months) or cytopathologic or histopathologic confirmation.

We used multiple tree-based ML algorithms (XGBoost, LGBM, Random Forest, and Extra Tree). Tree-based models have been shown to provide robust performance in the medical field.[4] SAINTS or Simple AI Nevus Transformation System used XGBoost, a ML algorithm based on the gradient boosting framework that both builds and iteratively refines an ensemble of weak prediction models in this case decision trees so that new models improve on a prior training model.[7] All imaging features available in both Wills Eye Hospital and Mayo Clinic cohorts were used. We used variance inflation factor (VIF) to test for multicollinearity. Features with VIF $>$10 were removed from the models. Features only available in 1 cohort were not used for model development. The Wills cohort was randomly segmented into an 80% training set—20% held-out test set. We used fivefold stratified cross-validation of the training set for hyperparameter tuning and set the optimization function as the area under receiver operating characteristic curve (AUROC). The Mayo cohort was utilized as an external validation set to assess the model's generalizability to independent external data. To identify the features most predictive of transformation, we explored the Shapley Additive exPlanations values for each model. Given our definition of growth to melanoma, a subanalysis was performed where only nevi with 5 or more years of follow-up were included on the top-performing algorithm (XGBoost). We tested a streamlined version of the top-performing algorithm, which utilized only 13 of the original 22 features (including age, race, sex, affected eye, tumor thickness, largest tumor diameter, tumor shape, distance to optic nerve, extent of subretinal fluid [SRF], distance to fovea, and internal reflectivity). This 'lite' model aimed to assess the algorithm's reliability and performance with fewer features. To establish threshold probabilities for both the full and lite models, we identified the points where Youden's J statistic reached its maximum for each cohort. In the full and long-term follow-up models, due to similar threshold probabilities and consistent performance, we chose the midpoint as the operational threshold. However, for the lite model, the threshold probabilities varied more significantly at the maximal Youden's J statistic. Consequently, we selected an intermediate value as the threshold to ensure a more balanced performance between both models.

The primary outcome was model performance as defined by its discriminative performance with AUROC and area under precision—recall curve (AUPRC). The AUPRC is a metric that evaluates the relationship between sensitivity and positive predictive value (PPV) across various thresholds and provides a measure of the model's ability to predict the positive class. To calculate 95% confidence intervals (CIs) for AUROC and AUPRC, we utilized 10 000 bootstrap samples with replacement. Microsoft's Fast and Lightweight AutoML Library was used for model hyperparameter optimization. Additionally, we used Python (version 3.7), pandas, scikit-learn, and seaborn for data analysis and visualization. Chi-squared and Mann—Whitney tests were used to assess for differences between the Wills and Mayo cohorts. Models were implemented and built for public use with Gradio and Hugging Face Spaces. Both models SAINTS and SAINTS Lite are available are the following links: Full Model (https://huggingface.co/spaces/ptailor3/SAINTS_Large) Lite Model: (https://huggingface.co/spaces/ptailor3/SAINTS_Lite).

Results

There were 2870 nevi included in the study with 128 nevi converted to melanoma. The rate of nevus to melanoma transformation was 3.8% in the Wills cohort (2356 nevi; 90 transformation to melanoma) and 7.4% in the Mayo cohort (514 nevi; 38 transformation to melanoma). The mean follow-up for the Wills cohort was 3 years (median: 3; range: <1−11 years).[2] The mean follow-up for the Mayo cohort was 5.5 years (median: 2.7 years; range: <1−11 years). The baseline characteristics for each cohort are summarized in Table 1. There were differences between the cohorts across multiple features as stated in Table 1, including: age, visual acuity at presentation, largest tumor basal diameter, largest tumor thickness, distance to optic nerve, SRF extent, SRF in the fovea, choroidal neovascular membrane, and internal reflectivity. Simple AI Nevus Transformation System and other models were trained on the following 22 features: age, race, sex, history of cutaneous melanoma, welder occupation, known germline BRCA1-associated protein 1 mutation, affected eye, melanocytosis, presenting visual acuity, tumor thickness, largest tumor diameter, tumor shape, distance to optic nerve, SRF extent, distance to fovea, RPE invasion, color, SRF in fovea, orange pigment, choroidal neovascular membrane, RPE trough, and internal reflectivity. One feature (anteroposterior location) was removed based on VIF.

Simple AI Nevus Transformation System outperformed the other algorithms across all cohorts (Table 2, Figs 1−4) ($P < 0.001$; $P < 0.001$; $P < 0.001$). On the Wills held-out test set, SAINTS achieved an AUROC of 0.864 (95% CI: 0.864−0.865) and an AUPRC of 0.244 (95% CI: 0.243−0.246). On the Mayo cohort (external validation), the model achieved an AUROC of 0.931 (95% CI: 0.930−0.931) and AUPRC of 0.533 (95% CI: 0.531−0.535) (Fig 1). Detailed model metrics are found in Table 3. Other models also displayed acceptable performance. On the Wills held-out test cohort, LGBM had AUROC value of 0.831 (95% CI: 0.831−0.832) and AUPRC value of 0.171 (95% CI: 0.169−0.172), whereas, on the external Mayo cohort, the AUROC was 0.815 (95% CI: 0.814−0.815) and the AUPRC was 0.277 (95% CI: 0.276−0.279) (Fig 2). On the Wills held-out test set, Random Forest achieved AUROC values of 0.812 (95% CI: 0.811−0.813) and AUPRC value of 0.122 (95% CI: 0.121−0.123); on the external Mayo cohort, the model achieved AUROC of 0.866 (95% CI: 0.866−0.867) and AUPRC of 0.418 (95% CI: 0.417−0.420) (Fig 3). Finally, on the Wills cohort, Extra Tree had AUROC value of 0.826 (95% CI: 0.826−0.827) and AUPRC value of 0.119 (95% CI: 0.118−0.119); on the Mayo cohort, the AUROC value was 0.915 (95% CI: 0.915−0.916), and the AUPRC value was 0.511 (95% CI: 0.509−0.513) (Fig 4). Simple AI Nevus Transformation System lite demonstrated reliable but worse performance. On the Wills held-out test cohort, SAINTS lite had AUROC value of 0.866 (95% CI: 0.866−0.867) and AUPRC value of 0.216 (95% CI: 0.215−0.217), whereas on the external Mayo cohort, the AUROC was 0.898 (95% CI: 0.898−0.899), and the AUPRC was 0.478 (95% CI: 0.476−0.480).

Further evaluation of the top-performing SAINTS model and its simplified variant, SAINTS Lite, both utilizing XGBoost, was conducted across both cohorts, focusing on optimal threshold probabilities and a range of key performance metrics (Table 3). The SAINTS model, set at a threshold probability (TP) of 0.38, showcased commendable accuracy in both the Wills Test Held-out cohort (0.910) and the Mayo External Validation cohort (0.889). A marked contrast was observed in its sensitivity, which was significantly higher in the Mayo cohort (0.869) compared with that of the Wills cohort (0.635), while maintaining robust specificity in both groups. Simple AI Nevus Transformation System Lite exhibited varying optimal TPs tailored to each cohort—0.545 for Wills and 0.383 for Mayo. At a TP of 0.383, the Wills cohort showed notably low accuracy and PPV but high sensitivity (1.0) and specificity (0.922). Contrastingly, in the Mayo cohort, SAINTS Lite demonstrated a substantial improvement with an accuracy of 0.897 and a PPV of 0.397. With a TP of 0.545, the model's accuracy and PPV were more favorable in the Mayo cohort than in the Wills cohort. Operating at a balanced TP of 0.5, SAINTS Lite achieved comparable accuracies of approximately 0.9 in both cohorts, albeit with reduced PPV and sensitivity. Notably, specificity remained consistently high across all scenarios and cohorts.

SHapley Additive exPlanations values were calculated to identify the most important features for prediction of nevus transformation to melanoma (Fig 5). The top 5 most important features in SAINTS were: tumor thickness, largest tumor basal diameter, tumor shape, distance to optic nerve, and SRF extent. The top 5 most important features included in the top 5 features of all 4 models were: tumor thickness (4/4), largest tumor diameter (4/4), patient age at presentation (2/4), tumor shape (2/4), tumor distance to optic nerve (2/4), SRF extent (2/4), distance to fovea (2/4), presenting visual acuity (1/4), and tumor internal reflectivity (1/4).

Given the definition of transformation in this study is based on growth, an additional model utilizing XGBoost was constructed solely using patients with at least 5 years of follow-up in both cohorts (Wills n = 573, Mayo n = 289). This model achieved a slightly worse AUROC in both the Wills (0.824 [95% CI: 0.823−0.824]) and Mayo cohorts (0.864 [95% CI: 0.864−0.865]); however, this model did achieve the best AUPRC in both the Wills (0.591 [95% CI: 0.591−0.592]) and Mayo cohorts (0.651 [95% CI: 0.651−0.652]) (Fig 6). The top 5 features of this model in order of SHapley Additive exPlanations value were: tumor thickness, largest tumor diameter, tumor shape, SRF extent, and tumor distance to optic nerve. Detailed evaluation of this model based on an optimal TP was also performed and demonstrated this model had the highest F1-score and PPV (Table 4).

Table 1. Predicting Choroidal Nevus Transformation to Melanoma with Machine Learning: Summary Statistics for Wills Eye Hospital Cohort and Mayo Clinic Cohort

Nevi Demographics	Wills Cohort (n = 2356)	Mayo Cohort (n = 514)	*P* Values
Age (yrs) mean ± std (median, range)	71.1 ± 16.0 (73, 11–107)	65.81 ± 16.7 (69, 8–98)	<0.001
Race (%)			
White	95.63 (2253)	98.83 (508)	>0.99
African American	0.76 (18)	0 (0)	
Hispanic	0.85 (20)	0.39 (2)	
Asian	0 (1)	0.19 (1)	
Indian	0.25 (6)	0.19 (1)	
Others	0 (2)	0 (0)	
Unknown	2.38 (56)	0 (0)	
Middle Eastern	0 (0)	0.19 (1)	
Sex (%)			>0.99
Male	37.01 (872)	38.33 (197)	
Female	62.82 (1480)	61.67 (317)	
History of cutaneous melanoma (%)			>0.99
Yes	4.84 (114)	0 (0)	
No	95.16 (2242)	100 (514)	
Welder occupation (%)			>0.99
Yes	0.34 (8)	0 (0)	
No	99.66 (2348)	100 (514)	
Germline BAP1 (%)			>0.99
No	100 (2356)	100 (514)	
Affected eye (%)			>0.99
Right	46.01 (1084)	44.75 (230)	
Left	40.49 (954)	46.3 (238)	
Both	13.5 (318)	8.95 (46)	
Melanocytosis (%)			>0.99
No	100 (2356)	100 (514)	
Visual acuity at presentation (logMAR) mean ± std (median, range)	0.16 ± 0.29 (0.1, 0–4)	0.13 ± 0.33 (0, 0–4)	<0.001
Largest tumor quadrant (%)			>0.99
Macula	26.95 (635)	12.84 (66)	
Inferior	21.05 (496)	23.54 (121)	
Temporal	20.29 (478)	26.26 (135)	
Superior	17.23 (406)	23.15 (119)	
Nasal	14.47 (341)	14.2 (73)	
Anteroposterior location of epicenter (%)			>0.99
Macula	27.72 (1428)	30.35 (156)	
Macula to equator	60.61 (653)	63.42 (326)	
Equator to ora	11.63 (274)	6.23 (32)	
Largest tumor basal diameter (mm) mean ± std (median, range)	4.74 ± 3.19 (4, 0.1–20)	6.49 ± 3.42 (6, 1–25)	<0.001
Largest tumor thickness (mm) mean ± std (median, range)	1.48 ± 0.70 (1.5, 0.1–6.7)	1.10 ± 0.80 (1, 0–5.9)	<0.001
Distance to optic nerve (mm) mean ± std (median, range)	5.17 ± 3.78 (5, 0–23)	5.94 ± 4.61 (5, 0–24)	<0.001
Distance to fovea (mm) mean ± std (median, range)	4.99 ± 3.89 (4, 0–20)	5.63 ± 4.87 (4, 0–24)	0.081
Color (%)			0.112
Pigmented	83.4 (1965)	74.12 (381)	
Mixed	9.8 (231)	14.98 (77)	
Nonpigmented	6.79 (160)	10.89 (56)	
Orange pigment (%)			0.672
Yes	4.29 (101)	15.37 (79)	
No	95.71 (2255)	84.44 (434)	
Subretinal fluid extent (%)			<0.001
None	93.38 (2200)	76.26 (392)	
Cap over nevus	3.74 (88)	19.46 (100)	
Greater than Cap	2.89 (68)	4.28 (22)	
Subretinal fluid in fovea (%)			0.011
Yes	1.23 (29)	3.5 (18)	
No	91.38 (2153)	96.5 (496)	
RPE trough (%)			0.612
Yes	1.4 (33)	5.06 (26)	
No	98.6 (2323)	94.94 (488)	
CNVM (%)			<0.001
Yes	0.55 (13)	0.97 (5)	

Table 1. (Continued.)

Nevi Demographics	Wills Cohort (n = 2356)	Mayo Cohort (n = 514)	*P* Values
No	99.45 (2343)	99.03 (509)	
RPE invasion (%)			>0.99
Yes	0.3 (7)	0 (0)	
No	99.7 (2349)	100 (514)	
Tumor shape (%)			0.073
Flat	60.14 (1417)	26.65 (137)	
Dome	32.64 (769)	71.6 (368)	
Internal reflectivity (%)			<0.001
Low	1.15 (27)	21.21 (109)	
Medium	36.88 (869)	15.37 (79)	
High	53.61 (1263)	63.42 (326)	

BAP1 = BRCA1-associated protein 1; CNVM = choroidal neovascular membrane; logMAR = logarithm of the minimum angle of resolution; RPE = retinal pigment epithelium; std = standard deviation.
Statistical testing used were chi-squared and Mann–Whitney *U* Tests.

Discussion

In this multicenter retrospective study, we developed, externally validated and released a ML algorithm (SAINTS; https://huggingface.co/spaces/ptailor3/SAINTS_Large; SAINTS Lite; https://huggingface.co/spaces/ptailor3/SAINTS_Lite) to predict choroidal nevus to melanoma transformation based on initial presenting features. Simple AI Nevus Transformation System, the best performing model performed well with tabular data due to the: (1) ability to handle diverse features, (2) robustness in handling outliers and missing values, and (3) sophisticated regularization parameters that helped prevent overfitting while maintaining high predictive accuracy.[7] The model performed the best of the 4 ML algorithms on both the Wills held-out test set and the Mayo external validation set, demonstrating that the model's performance is generalizable to other independent data. In fact, the model had better AUROC and AUPRC on the external validation set than the test set. We postulate the differences in performance are related to the differences in practice settings between the cohorts and the transformation rates. The Wills Eye Hospital Ocular Oncology Service serves not only as a tertiary or quaternary referral center but also serves a large community-based population. This differs from the Mayo Clinic cohort because the Mayo Clinic practice setting is primarily a tertiary care referral center for high-risk choroidal nevi that have been referred by outside providers. Lower risk nevi from the community are often seen at Mayo Clinic by providers outside of the Ocular Oncology Service and, therefore, may not have been included in the dataset. We only used Mayo Clinic data that included nevi evaluated with a full panel of multimodal imaging. These differences likely also contributed to the difference in conversion to choroidal melanoma rates (3.8% vs. 7.4%). We postulate that the model performs better on the Mayo cohort, due to both a higher transformation rate and PPV for transformation. This is seen in the AUPRC curve between the 2 models. Despite differing patient populations and practice settings, robust discriminative model performance was observed in both cohorts.

In an appropriate clinical context, the algorithms developed in this study will assist the clinician to stratify risk for patients with choroidal nevi and help determine cases that warrant referral to ocular oncology or inform monitoring frequency. Early referral of high-risk nevi to an expert ocular oncologist is essential. Early treatment of choroidal melanoma, when small, has a better long-term prognosis.[7] Of course, this model relies on accurate input of demographic and imaging data that may be challenging obtain in a general ophthalmology practice. This limitation necessitates future research to train models to recognize many differing multimodal imaging features. At the subspecialty care level, an ocular oncologist could leverage information from ML prediction models to tailor follow-up and management.

Table 2. Predicting Choroidal Nevus Transformation to Melanoma with Machine Learning: Discriminative Performance of all Models Based on AUROC and AUPRC

	XGBoost	LGBM	Random Forest	Extra Tree
AUROC				
Wills test held-out cohort (95% CI)	0.864 (0.864–0.865)	0.831 (0.831–0.832)	0.812 (0.811–0.813)	0.826 (0.826–0.827)
Mayo external validation cohort (95% CI)	0.931 (0.930–0.931)	0.815 (0.814–0.815)	0.866 (0.866–0.867)	0.915 (0.915–0.916)
AUPRC				
Wills test held-out cohort (95% CI)	0.244 (0.243–0.246)	0.171 (0.169–0.172)	0.122 (0.121–0.123)	0.119 (0.118–0.119)
Mayo external validation cohort (95% CI)	0.533 (0.531–0.535)	0.277 (0.276–0.279)	0.418 (0.417–0.420)	0.511 (0.509–0.513)

AUROC = area under receiver operating characteristics curve; AUPRC = area under precision–recall curve; CI = confidence interval.
95% CIs were generated by bootstrapping 10 000 samples with replacement.

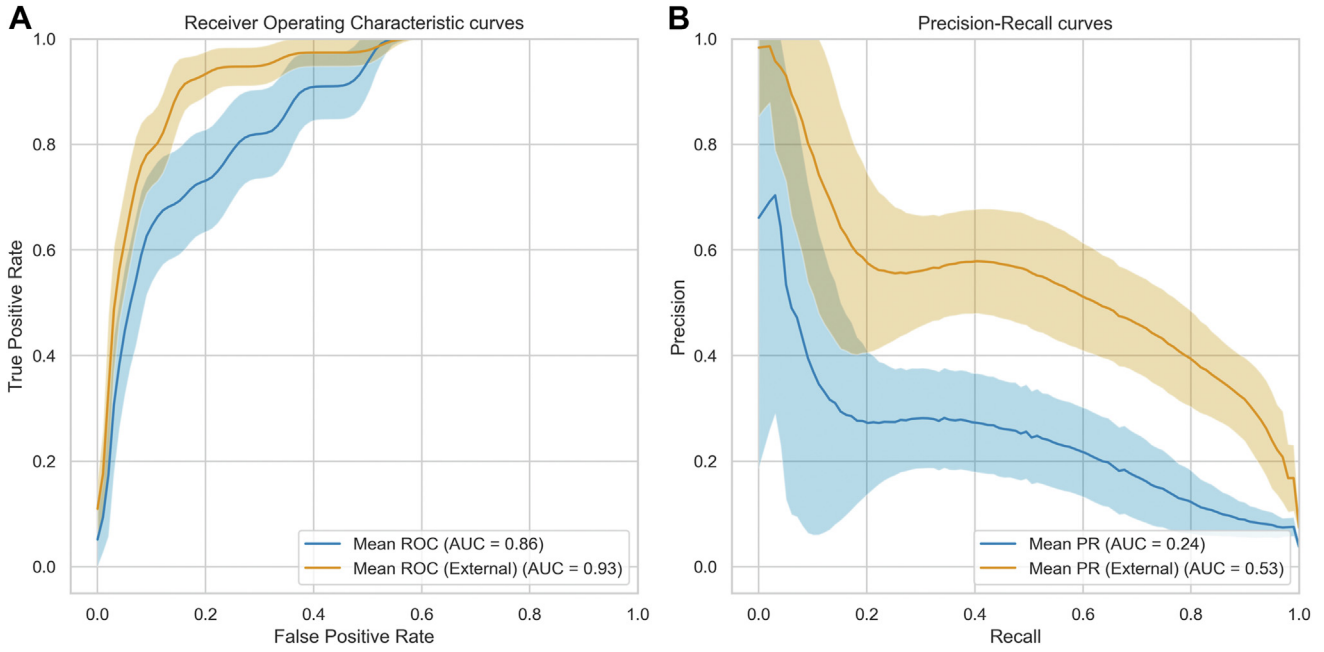

Figure 1. Receiver operating characteristics (ROC) curves and precision–recall (PR) curves for SAINTS (XGBoost). Receiver operating characteristics (**A**) and PR (**B**) curves for SAINTS (XGBoost) plot (**A**) true positive rate (sensitivity) vs. false positive rate (1-specificity) and (**B**) precision (positive predictive value) vs. recall (sensitivity) for a machine learning algorithm based on XGBoost. The 95% confidence intervals were generated using 10 000 bootstrapped samples with replacement. Pooled area under the curve values are given for the Wills held-out test set (blue) and Mayo external validation set (yellow) for ROC (**A**) and PR (**B**) curves. SAINTS demonstrates good discriminative performance on ROC curves (Wills: 0.86; Mayo: 0.93); however, has worse performance on PR curve (Wills: 0.24; Mayo: 0.53). AUC = area under the curve; SAINTS = Simple AI Nevus Transformation System.

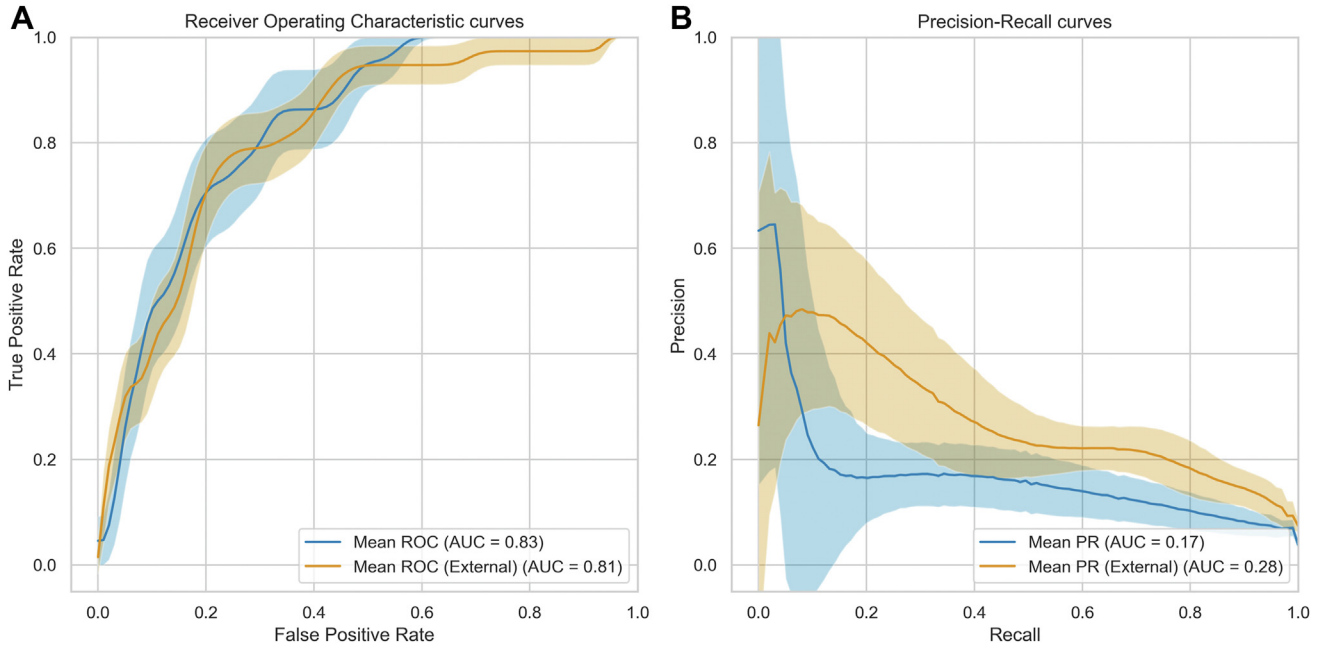

Figure 2. Receiver operating characteristics (ROC) curves and precision–recall (PR) curves for LBGM. Receiver operating characteristics (**A**) and PR (**B**) curves for LGBM plot (**A**) true positive rate (sensitivity) vs. false positive rate (1-specificity) and (**B**) precision (positive predictive value) vs. recall (sensitivity) for the LGBM model. The 95% confidence intervals were generated using 10 000 bootstrapped samples with replacement. Pooled area under the curve values are given for the Wills held-out test set (blue) and Mayo external validation set (yellow) for ROC (**A**) and PR (**B**) curves. LGBM demonstrates good discriminative performance on ROC curves (Wills: 0.83; Mayo: 0.81); however, has worse performance on PR curve (Wills: 0.17; Mayo: 0.28). AUC = area under the curve.

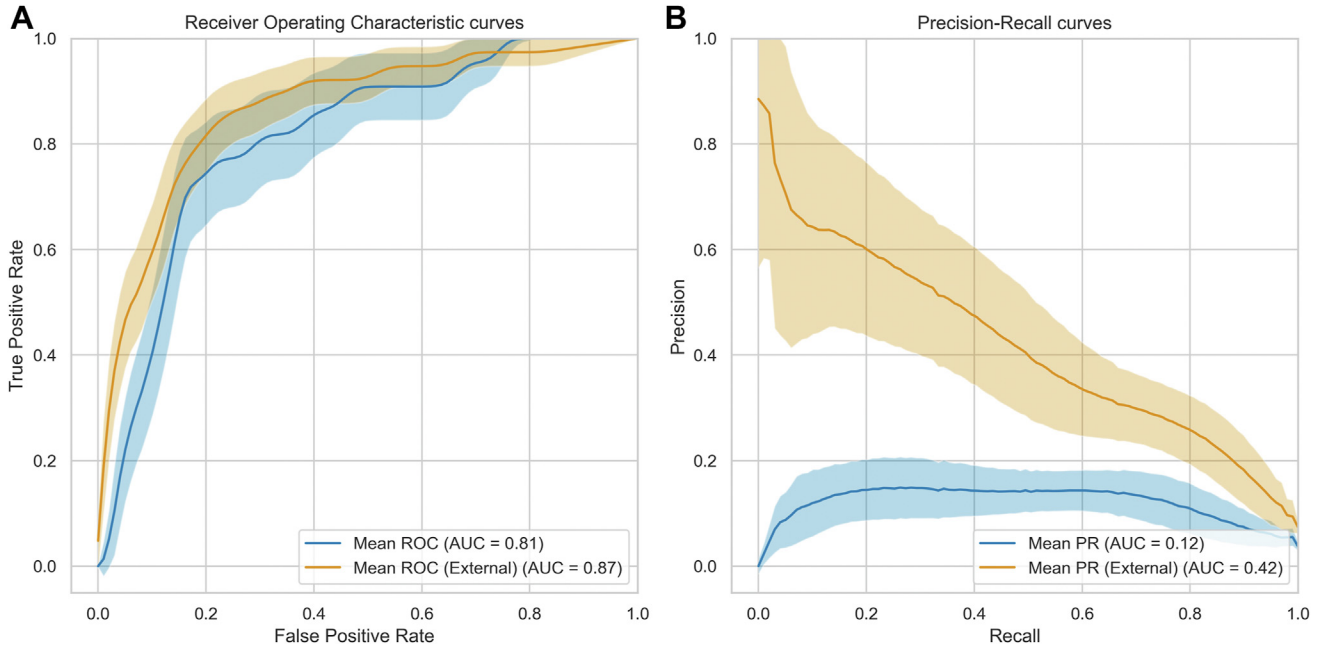

Figure 3. Receiver operating characteristics (ROC) curves and precision–recall (PR) curves for Random Forest. Receiver operating characteristics (**A**) and PR (**B**) curves for Random Forest plot (**A**) true positive rate (sensitivity) vs. false positive rate (1-specificity) and (**B**) precision (positive predictive value) vs. recall (sensitivity) for the Random Forest model. The 95% confidence intervals were generated using 10 000 bootstrapped samples with replacement. Pooled area under the curve values are given for the Wills held-out test set (blue) and Mayo external validation set (yellow) for ROC (**A**) and PR (**B**) curves. Random Forest demonstrates good discriminative performance on ROC curves (Wills: 0.81; Mayo: 0.87); however, has worse performance on PR curve (Wills: 0.12; Mayo: 0.42). AUC = area under the curve.

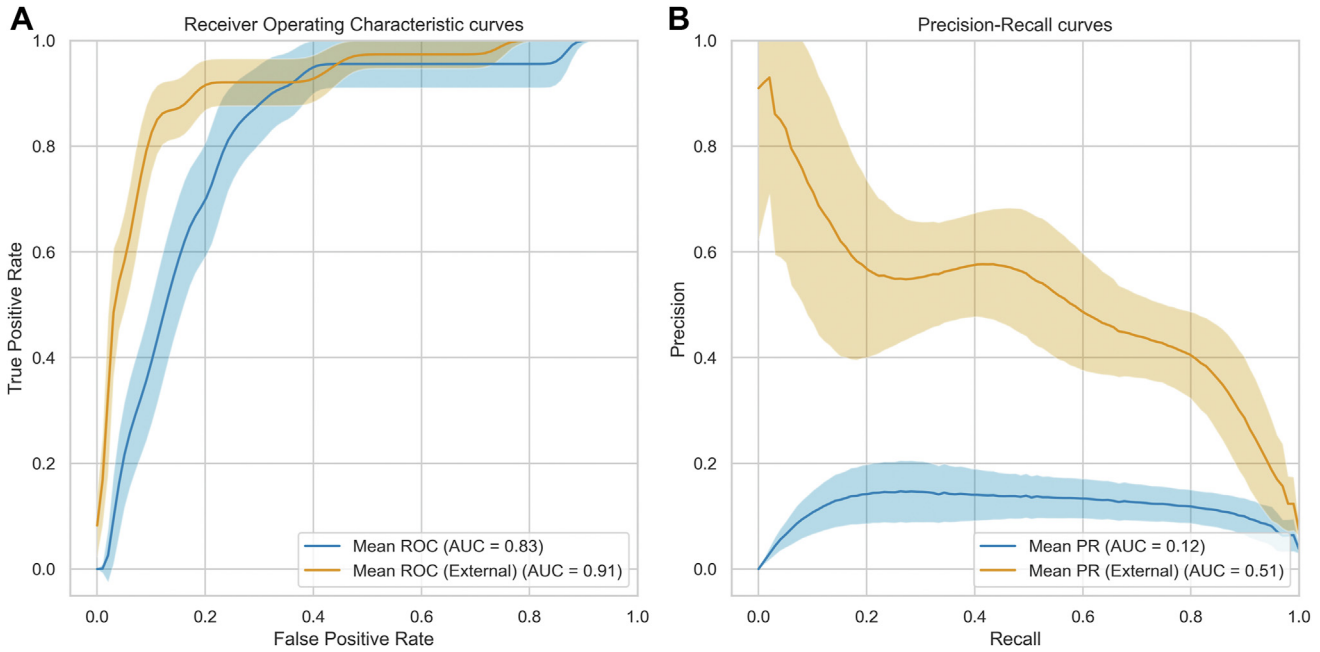

Figure 4. Receiver operating characteristics (ROC) curves and precision–recall (PR) Curves for Extra Tree. Receiver operating characteristics (**A**) and PR (**B**) curves for Extra Tree plot (**A**) true positive rate (sensitivity) vs. false positive rate (1-specificity) and (**B**) precision (positive predictive value) vs. recall (sensitivity) for the Extra Tree model. The 95% confidence intervals were generated using 10 000 bootstrapped samples with replacement. Pooled area under the curve values are given for the Wills held-out test set (blue) and Mayo external validation set (yellow) for ROC (**A**) and PR (**B**) curves. Extra Tree demonstrates good discriminative performance on ROC curves (Wills: 0.83; Mayo: 0.91); however, has worse performance on PR curve (Wills: 0.12; Mayo: 0.51). AUC = area under the curve.

Table 3. Predicting Choroidal Nevus Transformation to Melanoma with Machine Learning: Detailed Model Performance Metrics for SAINTS and SAINTS Lite

	Accuracy	PPV	Sensitivity	F1-Score	Specificity
SAINTS (TP = 0.38)					
Wills test held-out cohort (95% CI)	0.910 (0.910–0.910)	0.238 (0.237–0.239)	0.635 (0.633–0.638)	0.344 (0.342–0.345)	0.981 (0.978–0.982)
Mayo external validation cohort (95% CI)	0.889 (0.889–0.890)	0.389 (0.388–0.390)	0.869 (0.868–0.870)	0.535 (0.534–0.536)	0.984 (0.982–0.986)
SAINTS Lite					
TP (0.383)					
Wills test held-out cohort (95% CI)	0.037 (0.037–0.037)	0.037 (0.037–0.037)	1.0 (1.0–1.0)	0.072 (0.071–0.072)	0.922 (0.922–0.923)
Mayo external validation cohort (95% CI)	0.897 (0.897–0.897)	0.397 (0.396–0.399)	0.763 (0.762–0.765)	0.520 (0.519–0.522)	0.911 (0.911–0.911)
TP (0.545)					
Wills test held-out cohort (95% CI)	0.896 (0.896–0.897)	0.217 (0.216–0.218)	0.681 (0.679–0.683)	0.327 (0.326–0.328)	0.967 (0.965–0.968)
Mayo external validation cohort (95% CI)	0.907 (0.906–0.907)	0.400 (0.399–0.401)	0.525 (0.523–0.527)	0.451 (0.450–0.453)	0.977 (0.976–0.979)
Balanced TP (0.5)					
Wills test held-out cohort (95% CI)	0.808 (0.808–0.809)	0.136 (0.135–0.136)	0.772 (0.770–0.774)	0.230 (0.229–0.231)	0.953 (0.952–0.952)
Mayo external validation cohort (95% CI)	0.906 (0.906–0.907)	0.400 (0.398–0.401)	0.526 (0.524–0.528)	0.451 (0.450–0.453)	0.948 (0.948–0.948)

CI = confidence interval; PPV = positive predictive value; SAINTS = Simple AI Nevus Transformation System; TP = threshold probability.
95% CIs were generated by bootstrapping 10 000 samples with replacement. Threshold probabilities for both models determined by Youden's J statistic. The threshold probabilities for SAINTS were 0.36 and 0.40 with similar performance so we show the performance for the midpoint.

The implications of false positive and false negatives results from ML models are also important to consider. A false negative in the model would predict that a nevus does not convert to melanoma, when it truly converts to a melanoma. A false positive would involve the model predicting nevus conversion to melanoma when the lesion remains a benign nevus. Given the risk to life if a melanoma is missed, we advocate for models with a low false negative rate. Clearly, we also understand the challenges associated with potential over treatment from false positive results. Because monitoring low risk choroidal nevi is not an insignificant cost in the health care system, we believe AI-based algorithms may provide a cost-effective solution to improve patient outcomes and ensure timely follow-up.[8] An ideal model should minimize the risk of both false negative and false positive errors. The AUROC values in this study were high, indicating that the model has a high likelihood of effectively differentiating benign nevi that convert to melanoma. In our study, we optimized the SAINTS model by setting its probability threshold at 0.38. This decision was guided by the maximization of Youden's J statistic, striking a balance between sensitivity and specificity across the 2 cohorts. At this calibrated threshold, the SAINTS model exhibited a sensitivity of 0.635 in the Wills cohort and a significantly higher sensitivity of 0.869 in the Mayo cohort. Conversely, for SAINTS Lite, we selected a balanced TP of 0.5, aiming for a more equilibrated model performance. This adjustment yielded sensitivities of 0.772 in the Wills cohort and 0.526 in the Mayo cohort. Notably, both models maintained high specificity levels, with SAINTS Lite achieving 0.953 in the Wills cohort and 0.948 in the Mayo cohort, underscoring the robustness of the models in accurately identifying true negatives.

The AUPRC curve provides visualization for ML performance of the positive class (i.e., the subset of nevi with transformation to melanoma) while also showing the rate of false negatives. The AUPRC values were notably lower than the corresponding AUROC values, indicating that despite good overall discriminative abilities, the tradeoff between precision (PPV) and recall (sensitivity) regarding conversion to melanoma is suboptimal. Subanalysis demonstrated significantly improved performance when only evaluating nevi with long-term follow-up (>5 years), which indicates that excluding cases with shorter follow-up periods reduces noise and enhances the model's ability to correctly identify true positives. This reduction in noise leads to better sensitivity and PPV, as the model is trained on more representative and comprehensive data. Additionally, longer follow-up periods ensure more accurate CIs and higher predictive value, reflecting the true long-term risk of transformation. Although the features used demonstrate predictive value, additional features in the raw images or novel biomarkers that remain undiscovered could enhance model performance. The use of tabular data suffers from being reductive as it simplifies complex multimodal imaging data and is reliant on user input. Future research leveraging direct deep learning analysis of the multimodal images would help address this shortcoming.

The aggregate features identified by the models are supported by prior studies and are part of already existing risk assessment tools.[2] Tumor thickness is a well-established risk factor with multiple studies correlating increased thickness with

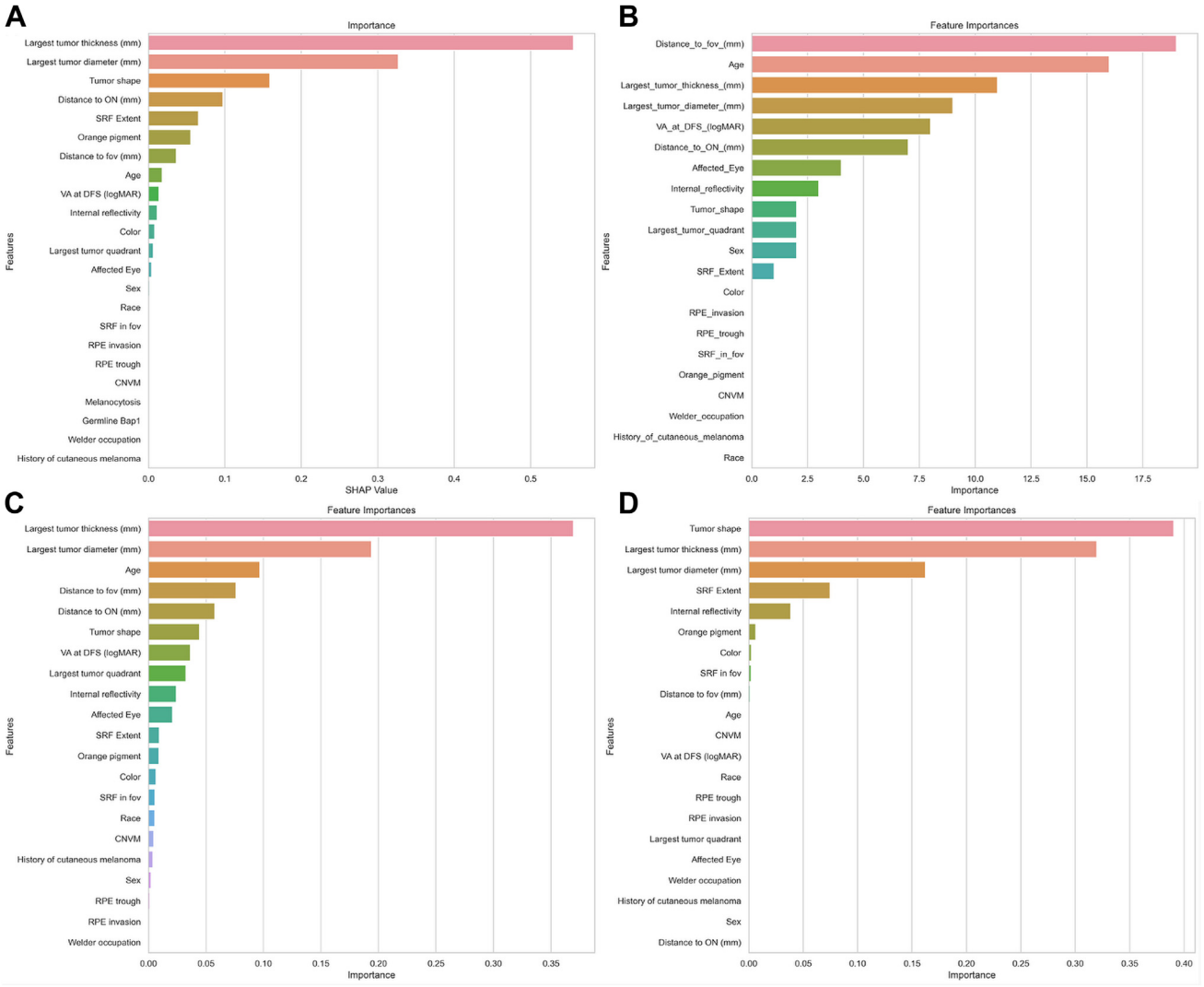

Figure 5. SHapley Additive exPlanations (SHAP) for 4 machine learning models for prediction of choroidal nevus transformation to melanoma (**A,** SAINTS; **B,** LGBM; **C,** Random Forest; **D,** Extra Tree). The graphs show the SHAP values for features that contribute to the prediction of each model. The SHAP values measure the impact of each feature on the model output. The features are ordered by their average absolute SHAP value across all samples. The color represents the feature value (red high, blue low). The top 5 features for each model (in order of SHAP value) are (**A**) SAINTS: tumor thickness, largest tumor basal diameter, tumor shape, distance to ON, and subretinal fluid extent; (**B**) LGBM: Tumor distance to fovea, patient age, tumor thickness, largest tumor basal diameter, and VA at presentation; (**C**) Random Forest: tumor thickness, largest tumor basal diameter, patient age, tumor distance to fovea, and distance to ON; (**D**) tumor shape, tumor thickness, largest tumor basal diameter, subretinal fluid extent, and internal reflectivity. CNVM = choroidal neovascular membrane; DFS = days first seen; logMAR = logarithm of the minimum angle of resolution; ON = optic nerve; RPE = included retinal pigment epithelium; SAINTS = Simple AI Nevus Transformation System; SHAP = SHapley Additive exPlanations; SRF = subretinal fluid; VA = visual acuity.

increasing melanoma risk.[2,9] Furthermore, increasing thickness of uveal melanoma is associated with increasing risk for metastasis.[10,11] Larger tumor diameter has also associated been associated with higher risk.[2,11] Tumor shape, specifically dome-shaped configuration, has been linked to transformation.[2] Greater SRF can be a sign of early transformation.[2] Interestingly, one of the most predictive factors identified in this study was tumor shape, not a top discriminating feature in other studies (e.g., Shields et al. 2019).[2] This highlights the complexity of predicting choroidal nevus transformation and suggests that ML approaches could identify novel predictive combinations of variables not captured by traditional statistical methods. The gradient boosting framework used by SAINTS aggregates the effects of multiple weak decision tree models into a more robust predictor. This flexible nonlinear approach appears better suited to unraveling the intricacies of factors underlying nevus progression compared to prior logistic regression-based risk tools. The expanded set of

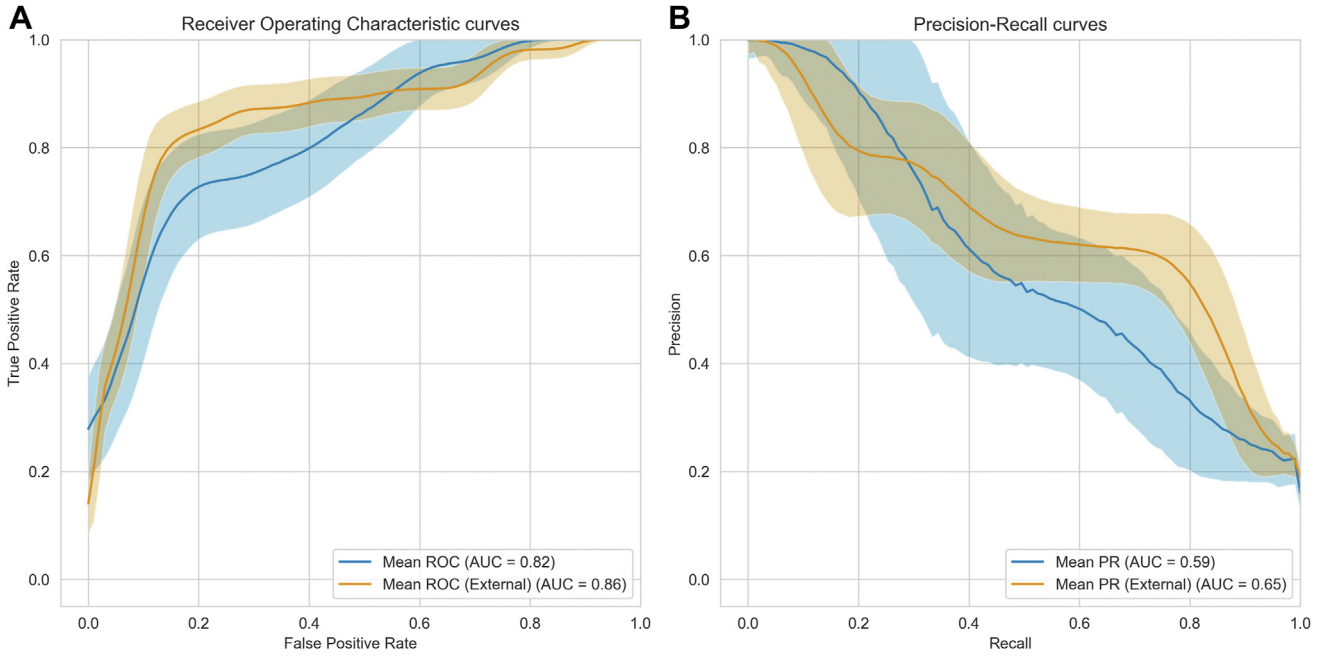

Figure 6. Receiver operating characteristics (ROC) curves and precision–recall (PR) curves for XGBoost for nevi with long-term follow-up (>5 years). Receiver operating characteristics (**A**) and PR (**B**) curves for XGBoost plot (**A**) true positive rate (sensitivity) vs. false positive rate (1-specificity) and (**B**) precision (positive predictive value) vs. recall (sensitivity) for the XGBoost model. The 95% confidence intervals were generated using 10 000 bootstrapped samples with replacement. Pooled area under the curve values are given for the Wills held-out test set (blue) and Mayo external validation set (yellow) for ROC (**A**) and PR (**B**) curves. XGBoost demonstrates good discriminative performance on ROC curves (Wills: 0.82; Mayo: 0.86) and on PR curves (Wills: 0.59; Mayo: 0.65).Nevus to melanoma transformation with ML/tailor/. AUC = area under the curve; ML = machine learning.

variables driving the model underscores the multifaceted clinical nature of this problem.

There are few previous studies applying ML for choroidal nevus transformation. One related study by Zabor et al developed a logistic regression model to diagnose small choroidal melanoma, achieving similar AUROC values to SAINTS.[5] However, there are notable differences between studies. Simple AI Nevus Transformation System was developed using a larger cohort for both model training (2870 vs. 123 patients) and external validation (514 vs. 240 patients).[5] This was true even in the long-term nevi cohort (>5 years of follow-up). Additionally, the proportion of nevi transforming to melanoma was lower in both the Wills Eye and Mayo Clinic cohorts (3.8% and 7.4%) compared with the training (49.6%) and validation (15.8%) sets in the Zabor et al study.[5] Larger sample sizes that better reflect the true clinical prevalence of transformation are critical for achieving generalizable models. Despite these differences, Zabor et al also identified tumor thickness, optic nerve proximity, and SRF as predictive factors, providing further evidence that such factors have predictive value.[5] Both studies demonstrate that ML has potential for identifying the risk of transformation from nevi to choroidal melanoma.

The strengths of this study include its large multicenter cohort of 2870 choroidal nevi from both community and tertiary settings, use of multiple imaging modalities to comprehensively phenotype nevi, and external validation demonstrating model robustness. However, limitations include the retrospective design relying on existing records, lack of tissue confirmation of disease in most cases, a primary White population in both datasets, use of clinician annotations

Table 4. Predicting Choroidal Nevus Transformation to Melanoma with Machine Learning: Detailed Model Performance Metrics for Long-Term (>5 years) Follow-Up Nevi

	Accuracy	PPV	Sensitivity	F1-Score	Specificity
Model (TP = 0.76)					
Wills test held-out cohort (95% CI)	0.820 (0.818–0.820)	0.458 (0.456–0.460)	0.695 (0.694–0.697)	0.548 (0.546–0.549)	0.837 (0.836–0.837)
Mayo external validation cohort (95% CI)	0.831 (0.830–0.831)	0.536 (0.535–0.537)	0.819 (0.817–0.820)	0.646 (0.645–0.647)	0.833 (0.832–0.833)

CI = confidence interval; PPV = positive predictive value; TP = threshold probability.
95% CIs were generated by bootstrapping 10 000 samples with replacement. Threshold probabilities for both models determined by Youden's J statistic.

and tabular data rather than raw imaging feature extraction, the imbalanced nature of the classification problem with relatively uncommon transformation events, use of only the largest nevus in each eye when multiple were present, and the risk of overstating utility without further validation guiding clinical integration. In our study, the definition of choroidal nevus is empirical and widely used.[6] However, we recognize its limitation, as morphologic features of large nevi and small melanomas have been shown to overlap and cause misclassifications[12,13] While clinical criteria are useful in everyday practice to separate high risk from low-risk lesions efficiently, they can have limited utility in classifying borderline cases or predicting metastatic potential. Fine needle aspiration of tissue with subsequent genetic analysis allows confirmation of malignancy and prediction of metastatic risk[14,15] Indeed, the very definition of nevus growth vs malignant transformation will benefit from future molecular studies, since nevus growth does not necessarily mean the lesion is malignant, and a growing benign tumor does not necessarily have the capacity to metastasize. Overall, while this study provides evidence that ML can risk-stratify nevi, more research is needed for model optimization and validation using even larger cohorts.

In conclusion, we have developed and validated SAINTS, a ML model to predict choroidal nevus transformation into melanoma. Simple AI Nevus Transformation System demonstrated strong discriminative ability on both the internal and external validation sets, supporting its potential for generalization to new clinical settings. The model identified tumor thickness, largest basal diameter, shape, distance to optic nerve, and SRF extent are the most robust predictive factors for nevus transformation to melanoma. While promising, there is room for improvement of the model by bypassing tabular data and utilizing deep learning to predict transformation based on raw images. This study provides proof-of-concept for the potential of ML to identify high-risk choroidal melanocytic lesions. Further prospective studies are warranted to further refine such models and test clinical integration. Simple AI Nevus Transformation System represents an important step towards leveraging AI for personalized data-driven management of choroidal nevi.

Acknowledgments

The authors thank Dr Jose S. Pulido for allowing us to use some of his former patients in this study.

Footnotes and Disclosures

Originally received: January 22, 2024.
Final revision: June 7, 2024.
Accepted: July 16, 2024.
Available online: ■■■. Manuscript no. XOPS-D-23-00328.

[1] Department of Ophthalmology, Mayo Clinic, Rochester, Minnesota, 55905.

[2] Ocular Oncology Service, Wills Eye Hospital, Thomas Jefferson University, Philadelphia, Pennsylvania, 19107.

This work was presented at AAO 2023.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors have made the following disclosures:

T.W.O.: Equity − iMacular Regeneration, LLC; Leadership − Owner IMacula Society (No role or relevance to this study).

L.A.D.: Honoraria − University of Texas, Wisconsin Academy of Ophthalmology.

Leonard and Mary Lou Hoeft Career Development Award Fund in Ophthalmology Research, Grant Number P30 CA015083 from the National Cancer Institute, and CTSA Grant Number KL2 TR002379 from the National Center for Advancing Translational Science (NCATS). The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

HUMAN SUBJECTS: Human subjects were included in this study. This study was approved by the institutional review board/ethics committees of Mayo Clinic and Wills Eye Hospital and adhered to the tenets of the Declaration of Helsinki and Health Insurance Portability and Accountability Act. All patients provided informed consent.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Tailor, Kopinski, D'Souza, Leske, C.L. Shields, J.A. Shields, Dalvin

Data collection: Tailor, Kopinski, D'Souza, Olsen, C.L. Shields, J.A. Shields, Dalvin

Analysis and interpretation: Tailor, Kopinski, D'Souza, Leske, Olsen, C.L. Shields, J.A. Shields, Dalvin

Obtained funding: N/A

Overall responsibility: Tailor, Kopinski, D'Souza, Leske, Olsen, C.L. Shields, J.A. Shields, Dalvin

Abbreviations and Acronyms:
AF = autofluorescence; **AI** = artificial intelligence; **AUROC** = area under receiver operating characteristic curve; **AUPRC** = area under precision−recall curve; **CI** = confidence interval; **ML** = machine learning; **PPV** = positive predictive value; **RPE** = retinal pigment epithelium; **SAINTS** = Simple AI Nevus Transformation System; **SHAP** = SHapley Additive exPlanations; **SRF** = subretinal fluid; **TP** = threshold probability; **VIF** = variance inflation factor.

Keywords:
Artificial Intelligence, Choroidal melanoma, Choroidal nevus, Machine learning, Ocular oncology.

Correspondence:
Lauren A. Dalvin, MD, Department of Ophthalmology, Mayo Clinic, 200 1st St SW, Rochester, MN 55905. E-mail: dalvin.lauren@mayo.edu.

References

1. Shields JA, Shields CL. *Intraocular Tumors: An Atlas and Textbook*. 3 ed. Philadelphia, PA: Wolters Kluwer Health; 2016.
2. Shields CL, Dalvin LA, Ancona-Lezama D, et al. Choroidal nevus imaging features in 3,806 cases and risk factors for transformation into melanoma in 2,355 cases: the 2020 Taylor R. Smith and Victor T. Curtin Lecture. *Retina*. 2019;39: 1840−1851.
3. Pfaff ER, Girvin AT, Bennett TD, et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digit Health*. 2022;4: e532−e541.
4. Kwong JCC, Khondker A, Meng E, et al. Development, multi-institutional external validation, and algorithmic audit of an artificial intelligence-based Side-specific Extra-Prostatic Extension Risk Assessment tool (SEPERA) for patients undergoing radical prostatectomy: a retrospective cohort study. *Lancet Digit Health*. 2023;5:e435−e445.
5. Zabor EC, Raval V, Luo S, et al. A prediction model to discriminate small choroidal melanoma from choroidal nevus. *Ocul Oncol Pathol*. 2022;8:71−78.
6. Augsburger JJ, Schroeder RP, Territo C, et al. Clinical parameters predictive of enlargement of melanocytic choroidal lesions. *Br J Ophthalmol*. 1989;73:911−917.
7. Chen T, Guestrin C. XGBoost: a scalable tree boosting System. arXiv:1603.02754; 2016. https://ui.adsabs.harvard.edu/abs/2016arXiv160302754C. Accessed August 11, 2024
8. Barsam AS, Gibbons A, McClellan AJ, et al. Follow the nevus: the cost-utility of monitoring for growth of choroidal nevi. *Int J Ophthalmol*. 2019;12:1456−1464.
9. Shields CL, Dalvin LA, Yu MD, et al. Choroidal nevus transformation into melanoma per millimeter increment in thickness using multimodal imaging in 2355 cases: the 2019 Wendell L. Hughes Lecture. *Retina*. 2019;39: 1852−1860.
10. Shields CL, Furuta M, Thangappan A, et al. Metastasis of uveal melanoma millimeter-by-millimeter in 8033 consecutive eyes. *Arch Ophthalmol*. 2009;127:989−998.
11. Roelofs KA, O'Day R, Harby LA, et al. The MOLES system for planning management of melanocytic choroidal tumors: is it safe? *Cancers*. 2020;12:1311.
12. Augsburger JJ, Correa ZM, Trichopoulos N, Shaikh A. Size overlap between benign melanocytic choroidal nevi and choroidal malignant melanomas. *Invest Ophthalmol Vis Sci*. 2008;49: 2823−2828.
13. Harbour JW, Paez-Escamilla M, Cai L, et al. Are risk factors for growth of choroidal nevi associated with malignant transformation? Assessment with a validated genomic biomarker. *Am J Ophthalmol*. 2019;197:168−179.
14. Damato B, Duke C, Coupland SE, et al. Cytogenetics of uveal melanoma: a 7-year clinical experience. *Ophthalmology*. 2007;114:1925−1931.
15. Onken MD, Worley LA, Ehlers JP, Harbour JW. Gene expression profiling in uveal melanoma reveals two molecular classes and predicts metastatic death. *Cancer Res*. 2004;64:7205−7209.

Predicting Choroidal Nevus Transformation to Melanoma Using Machine Learning

Prashant D. Tailor, MD, Piotr K. Kopinski, MD, PhD, Haley S. D'Souza, MD, David A. Leske, MS, Timothy W. Olsen, MD, Carol L. Shields, MD, Jerry A. Shields, MD, Lauren A. Dalvin, MD

A machine learning model, Simple AI Nevus Transformation System, was created and validated to have good discriminative performance on test (area under receiver operating characteristic curve [AUROC] 0.864) and external (AUROC 0.931) validation cohorts for predicting choroidal nevus transformation to melanoma.