Marta Colmenar Herrera*, Şerife Seda Kucur, Pablo Márquez Neila, Raphael Sznitman, and Jan Unterlauft

Visual field-based machine learning model for predicting disease stage in glaucoma

https://doi.org/10.1515/cdbme-2025-0193

Abstract: Glaucoma is a leading cause of irreversible blindness, requiring early detection to prevent vision loss. This study evaluates non-linear machine learning (ML) models, such as AdaBoost, for forecasting glaucoma stage, comparing it to Logistic Regression (LR) using visual field data from three independent cohorts (Bern, Rotterdam, Washington). AdaBoost outperformed LR, achieving a kappa value of 0.85 (Bern) compared to 0.73 for LR. Notably, AdaBoost reached an AUC ROC of 0.93 (early glaucoma, Bern), surpassing LR's 0.47. Similar trends were observed across other datasets. These results demonstrate the ability of non-linear ML models to detect glaucoma progression, highlighting their potential to improve early intervention and treatment planning.

Keywords: Glaucoma, Forecasting, Machine Learning

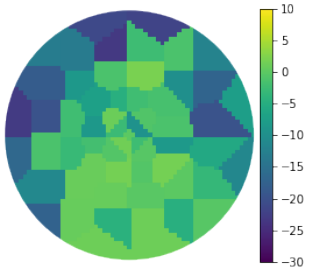

Fig. 1: A VF represents a subject's functional vision, measured at 59 test locations across the retina. Total deviation values, ranging from -30 to 10 dB, indicate the difference in visual sensitivity at each location compared to an age-matched healthy reference, with lower values suggesting greater visual loss.

1 Introduction

Glaucoma, a leading cause of irreversible blindness, is characterized by progressive optic nerve damage and peripheral vision loss [1]. Early progression detection is critical to prevent significant vision impairment, as the disease often progresses silently, with symptoms appearing only after substantial damage [2]. The standard for assessing functional changes in glaucoma is static automated perimetry, which maps a person's visual sensitivity across the visual field (VF) [3]. However, distinguishing true disease progression from normal fluctuations remains challenging, as no universally accepted definition of progression exists. In this study, progression is defined as a trend-based global measure, specifically the mean deviation (MD) over time, which reflects the average sensitivity across the VF, with lower values indicating worsening [4–6].

Clinically, glaucoma progression is often analyzed using linear regression models [7, 8], although alternative approaches such as robust regression, pointwise sensitivity analysis, and Bayesian models are increasingly being explored [9–11]. These methods aim to predict progression at individual VF locations. A related but simpler task—predicting the overall state of the glaucomatous eye—could still yield valuable insights into disease progression and help identify individuals experiencing rapid deterioration who may benefit from closer monitoring [12]. Traditionally, this task has been addressed using linear models such as linear regression. However, linear models are limited in their ability to capture the complex interactions between VF locations and disease progression.

In this study, we compare non-linear models with traditional linear approaches and show that non-linear methods provide more accurate forecasts of future glaucoma stages based on prior VF data. These findings highlight the potential of non-linear models as more effective tools for glaucoma monitoring.

2 Materials and Methods

Visual Field Data

We used three separate and independent cohorts, including:

– **Bern dataset**: 46,187 VFs from 9,869 subjects, collected at Inselspital Eye Clinic (Switzerland). VFs followed the G pattern (59 locations) using the OCTOPUS 900 Perimeter (Haag-Streit AG, Köniz, Switzerland).

***Corresponding author: Marta Colmenar Herrera, Pablo Márquez Neila, Raphael Sznitman,** ARTORG Center for Biomedical Engineering Research, University of Bern, Bern Switzerland, e-mail: marta.colmenar@unibe.ch
Şerife Seda Kucur, PeriVision SA, Epalinges, Switzerland
Jan Unterlauft, Department of Ophthalmology, Inselspital, Bern, Switzerland
Jan Unterlauft, Department of Ophthalmology, University of Leipzig, Leipzig, Germany.

Dataset	Device	No. Subjects	No. Visual Fields	Age (Mean± SD [Min, Max])	MD (Mean± SD [Min, Max])	Glaucoma Staging Distribution (%)					
						Earliest	Early	Moderate	Advanced	Severe	End-stage
Bern	OCTOPUS 900	8,463	38,832	57.49 ± 17.83 $[7.22, 95.57]$	4.45 ± 5.61 $[-9.20, 28.90]$	8.09	57.92	17.95	9.76	5.6	0.68
Rotterdam	Humphrey Field Analyzer II	139	4,572	64.24 ± 10.69 $[25.75, 87.15]$	-9.35 ± 7.63 $[-31.26, -0.01]$	5.98	—	35.35	29.04	29.63	—
Washington	Humphrey Field Analyzer II	3,867	28,553	64.50 ± 14.46 $[10.06, 90.00]$	-6.94 ± 6.18 $[-32.95, 0.00]$	—	52.56	30.28	11.61	5.55	—

Tab. 1: Summary of demographic of the datasets used in the study, including device type, number of subjects and VFs, age range, mean deviation (MD) values, and distribution of instances per stage.

– **Rotterdam dataset**: 4,863 VFs from 139 subjects, collected at Rotterdam Eye Institute (Netherlands) [14, 15]. VFs were measured with the 24-2 pattern (54 locations) using the Humphrey Field Analyzer II (HFA, Carl Zeiss Meditec AG, Jena, Germany).
– **Washington dataset**: 28,943 VFs from 3,871 subjects, acquired by the University of Washington [16]. VFs were measured with the 24-2 pattern (54 locations) using the HFA (Carl Zeiss Meditec AG, Inc. Dublin, CA, USA).

The Rotterdam and Washington datasets are publicly available, while the Bern dataset was curated to exclude VFs with false positive/negative rates over 30% or those not aligning with glaucoma patterns (e.g., stroke). In the Bern dataset for the OCTOPUS perimeter, MD values are inverted, with higher positive numbers indicating greater visual impairment.

Glaucoma Staging

We applied the MD-based stage classification system by Richard et al. [17], adapting it for HFA and OCTOPUS devices (Table 2). It classifies severity from ocular hypertension (earliest glaucoma) to end-stage glaucoma. We excluded stages from cohorts with fewer than 400 VFs to mitigate biases from low-participant numbers: end-stage glaucoma in Bern and earliest-stage cases in Rotterdam. A summary of the datasets is provided in Table 1.

Stage	Description	HFA MD Score (dB)	OCTOPUS MD Score (dB)
0	Earliest	> 0.00	≤ -0.80
1	Early	-0.01 to -5.00	-0.70 to $+4.40$
2	Moderate	-5.01 to -12.00	$+4.50$ to $+9.40$
3	Advanced	-12.01 to -20.00	$+9.50$ to $+15.30$
4	Severe	< -20.01	$+15.40$ to $+23.10$
5	End-stage	—	$\geq +23.20$

Tab. 2: Glaucoma severity classification based on MD for HFA and OCTOPUS perimeters.

Machine Learning Model and Baseline

This study focused in glaucoma progression by forecasting the glaucoma stage, framing it as a multi-class classification problem (stages 0–4). We construct triplets from subjects with at least three visits, sequentially forming triplets for each dataset: 15,725 (Bern), 4,021 (Rotterdam), and 13,848 (Washington). The average time between VFs was 0.91 ± 1.71 (Bern), 0.56 ± 0.20 (Rotterdam), and 1.30 ± 1.05 years (Washington), with triplet windows of 1.78 ± 2.14, 1.12 ± 0.28, and 2.67 ± 1.73 years, respectively.

Our goal was to predict the stage of the third VF test using data from the first two. To this end, we trained an AdaBoost model [18] using total deviations (Figure 1) from the first two VFs, along with the subject's age as input features. Each dataset was split into five subject-exclusive folds (80% training, 20% testing). Logistic Regression (LR) served as a linear benchmark for comparison.

Given that a majority of cases remain stable over time, we also included a *no-change classifier* in our evaluation. This baseline model assumes no change in disease stage and simply assigns the future stage to match that of the last available VF. Comparing AdaBoost and LR against this baseline provides a clinically meaningful reference point for assessing model performance [13].

Model performance was evaluated using Cohen's kappa, which accounts for class imbalance and ranges from -1 (complete disagreement) to 1 (perfect agreement) [19]. Additionally, we report the Area Under the Curve (AUC) for both the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves.

3 Results

Kappa values from 5-fold cross-validation are used to compare the performance of AdaBoost, LR, and the no-change classifier across all datasets (Table 3). Results are presented for all cases and separately for progressed, stable, and improved cases, with progression defined as a higher VF stage, stability

as no change, and improvement as a lower stage compared to the last available VF.

Dataset	Category	AdaBoost	Logistic Regression	No-change classifier
Bern	All	0.85 ± 0.02	0.73 ± 0.05	0.87 ± 0.02
	Progressed	0.63 ± 0.03	0.55 ± 0.01	0.52 ± 0.02
	Stable	0.94 ± 0.02	0.81 ± 0.01	1.00 ± 0.00
	Improved	0.63 ± 0.02	0.58 ± 0.02	0.59 ± 0.03
Rotterdam	All	0.86 ± 0.00	0.85 ± 0.05	0.89 ± 0.01
	Progressed	0.51 ± 0.07	0.34 ± 0.04	0.30 ± 0.01
	Stable	0.94 ± 0.01	0.88 ± 0.03	1.00 ± 0.00
	Improved	0.52 ± 0.02	0.39 ± 0.06	0.33 ± 0.02
Washington	All	0.85 ± 0.01	0.76 ± 0.01	0.84 ± 0.00
	Progressed	0.49 ± 0.02	0.43 ± 0.02	0.39 ± 0.01
	Stable	0.95 ± 0.00	0.87 ± 0.01	1.00 ± 0.00
	Improved	0.57 ± 0.03	0.49 ± 0.04	0.42 ± 0.02

Tab. 3: Mean $\pm$ SD of Cohen's kappa metric across 5 folds for AdaBoost, Logistic Regression, and the no-change baseline, stratified by dataset and progression category.

The AUCs of the models across stages and datasets, using 5-fold cross-validation with a one-vs-rest approach, are also compared (Table 4).

Using the AUROC values, we analyzed the performance of AdaBoost and LR models at each glaucoma stage for a unique fold, focusing on instances of progression (Table 5). Cases with stage transitions of more than one stage were excluded from the analysis due to their low occurence in the dataset.

		AUC ROC		AUC PR	
	Stage	AdaBoost	Logistic Regression	AdaBoost	Logistic Regression
Bern	0	0.930 ± 0.006	0.467 ± 0.014	0.611 ± 0.025	0.096 ± 0.006
	1	0.907 ± 0.004	0.773 ± 0.007	0.891 ± 0.007	0.700 ± 0.010
	2	0.879 ± 0.006	0.826 ± 0.007	0.612 ± 0.017	0.485 ± 0.016
	3	0.935 ± 0.005	0.883 ± 0.007	0.647 ± 0.020	0.384 ± 0.019
	4	0.977 ± 0.004	0.948 ± 0.004	0.817 ± 0.0018	0.471 ± 0.030
Rotterdam	3	0.960 ± 0.005	0.959 ± 0.005	0.948 ± 0.007	0.880 ± 0.021
	4	0.852 ± 0.011	0.913 ± 0.009	0.647 ± 0.027	0.851 ± 0.019
	5	0.938 ± 0.008	0.973 ± 0.006	0.903 ± 0.012	0.964 ± 0.006
Washington	1	0.945 ± 0.003	0.899 ± 0.005	0.928 ± 0.006	0.855 ± 0.010
	2	0.875 ± 0.006	0.797 ± 0.007	0.754 ± 0.013	0.617 ± 0.014
	3	0.929 ± 0.006	0.880 ± 0.006	0.702 ± 0.022	0.379 ± 0.016
	4	0.970 ± 0.005	0.946 ± 0.005	0.835 ± 0.020	0.400 ± 0.027

Tab. 4: Mean $\pm$ SD of AUC for ROC and PR curves over 5 folds, reported per stage using a one-vs-rest setup.

	AdaBoost					Logistic Regression					
Bern	0.65	**0.64**				0.50	**0.51**				0
	0.84	0.72	**0.83**			0.44	0.56	**0.73**			1
		0.78	0.60	**0.76**			0.64	0.62	**0.68**		2
			0.70	0.61	**0.71**			0.67	0.59	**0.64**	3
				0.78	0.76				0.41	0.52	4
Rotterdam			0.86	**0.86**				0.85	**0.85**		2
			0.84	0.69	**0.79**			0.74	0.62	**0.76**	3
				0.85	0.89				0.75	0.75	4
Washington		0.78	**0.78**				0.68	**0.69**			1
		0.82	0.69	**0.82**			0.73	0.58	**0.74**		2
			0.77	0.67	**0.70**			0.67	0.56	**0.56**	3
				0.78	0.79				0.33	0.57	4
	0	1	2	3	4	0	1	2	3	4	

Starting Stage

Ending Stage

Tab. 5: Stage-wise comparison of AUROC for glaucoma stage prediction using AdaBoost and Logistic Regression. Each cell represents the AUROC for predicting a transition from a given starting stage (based on the second VF) to a future stage (the third VF). Rows correspond to starting stages, columns to predicted ending stages. Clinically relevant progression cases are highlighted in **bold**. Empty cells indicate transitions with insufficient samples to compute reliable metrics.

4 Discussion

This study evaluates the potential of non-linear models in forecasting glaucoma stage. As shown in Table 3, AdaBoost consistently outperforms LR across all datasets, achieving higher kappa values. In both the Bern and Washington datasets, AdaBoost shows a significant improvement over LR. More importantly, in the progressed cases, AdaBoost demonstrates a clear advantage, outperforming LR by a clear marging.

AUROC and AUC PR results (Table 4) further confirm AdaBoost's performance, maintaining AUROC values above 0.85 across all datasets—exceeding the threshold for clinically meaningful classification [20]. In particular, for early-stage glaucoma in Bern, AdaBoost shows an improvement of over 0.4 points in AUROC compared to LR, while for advanced glaucoma in Washington, AdaBoost achieves an improvement of approximately 0.3 points in AUC PR over LR.

Additionally, Table 5 highlights AdaBoost's superior performance in predicting glaucoma stage transitions, particularly from stage 1 to 2 in the Bern dataset and from stage 3 to 4 in the Washington dataset. In contrast, in the Rotterdam dataset, where the data is more structured and short-term, both models perform similarly, suggesting that LR may still be a reliable choice in controlled environments. These results emphasize AdaBoost's potential for more accurately predicting glaucoma progression in real-world clinical settings.

The no-change classifier served as a baseline to evaluate model performance in progressed cases. As shown in Table 3, non-linear models achieve higher kappa values in progressed cases, where accurate detection is most critical. This baseline demonstrates that non-linear models fit better progression trends as compared with linear models.

We acknowledge the limitations in our study. The lack of longitudinal subject history prevents evaluating the effect of additional visits on predictions. The Bern dataset is missing exact diagnostic information, possibly including non-glaucoma subjects, and none of the datasets contain treatment details, which could influence disease progression. Additionally, the rigid staging system may exaggerate minor changes, occasionally suggesting improvements. However, this simplification helps evaluate non-linear models in detecting progression. While glaucoma staging is ordinal, our approach assumes equal stage intervals, which may not fully align with clinical reality. Still, the results highlight the potential of non-linear models in capturing disease progression.

In conclusion, non-linear models show promise in detecting glaucoma progression more effectively than linear models, potentially aiding earlier intervention. This could help optimize follow-up schedules and prioritize high-risk subjects. Future work should address current limitations by incorporating more diagnostic data, and treatment information to enhance predictive accuracy.

Author Statement

Research funding: This work was partially supported by the Innosuisse Grant #100.767 IP-LS. Conflict of interest: SSK (O), RS (O). Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

[1] Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. JAMA 2014;311:1901-1911.

[2] Chauhan BC, Garway-Heath DF, Goñi FJ, et al. Practical recommendations for measuring rates of visual field change in glaucoma. The British journal of ophthalmology. 2008;92(4):569-73.

[3] Racette L, Fischer M, Bebie H, Holló G, Johnson CA, Matsumoto C. Automated Perimetry Visual Field Digest. 6th ed. Switzerland: Haag-Streit AG; 2016.

[4] Vianna JR, Chauhan BC. How to detect progression in glaucoma. Progress in Brain Research 2015;221:135-58.

[5] Hu R, Racette L, Chen KS, Johnson CA. Functional assessment of glaucoma: Uncovering progression. Survey Ophthalmology 2020;65(6):639-661.

[6] Leung CK, Cheung CY, Weinreb RN, Qiu K, Liu S, Li H, et al. Evaluation of retinal nerve fiber layer progression in glaucoma: a study on optical coherence tomography guided progression analysis. Investigative ophthalmology & visual science 2010;51(1):217-22.

[7] Bengtsson B, Patella VM, Heijl A. Prediction of glaucomatous visual field loss by extrapolation of linear trends. Archives of ophthalmology. 2009;127(12):1610-5.

[8] Russell RA, Crabb DP, Malik R, Garway-Heath DF. The relationship between variability and sensitivity in large-scale longitudinal visual field data. Investigative ophthalmology & visual science 2012;53(10):5985-90

[9] Hu R, Marín-Franch I, Racette L. Prediction accuracy of a novel dynamic structure-function model for glaucoma progression. Investigative ophthalmology & visual science 2014;55(12):8086-94.

[10] Medeiros FA, Zangwill LM, Weinreb RN. Improved prediction of rates of visual field loss in glaucoma using empirical Bayes estimates of slopes of change. Journal of Glaucoma 2012;21(3):147-54.

[11] Murata H, Araie M, Asaoka R. A new approach to measure visual field progression in glaucoma patients using variational bayes linear regression. Investigative ophthalmology & visual science 2014 Nov 20;55(12):8386-92.

[12] Saunders LJ, Medeiros FA, Weinreb RN, Zangwill LM. What rates of glaucoma progression are clinically significant?. Expert review of ophthalmology 2016;11(3):227–234.

[13] Eslami M, Kim JA, Zhang M, Boland MV, Wang M, Chang DS, et al. Visual Field Prediction: Evaluating the Clinical Relevance of Deep Learning Models. Ophthalmology science 2022;3(1):100222.

[14] Bryan SR, Vermeer KA, Eilers PH, Lemij HG, Lesaffre EM. Robust and censored modeling and prediction of progression in glaucomatous visual fields. Investigative ophthalmology & visual science 2013;54(10):6694-700.

[15] Erler NS, Bryan SR, Eilers PH, Lesaffre EM, Lemij HG, Vermeer KA. Optimizing structure-function relationship by maximizing correspondence between glaucomatous visual fields and mathematical retinal nerve fiber models. Investigative ophthalmology & visual science 2014;55(4):2350-7.

[16] Montesano G, Chen A, Lu R, Lee CS, Lee AY. UWHVF: A Real-World, Open Source Dataset of Perimetry Tests From the Humphrey Field Analyzer at the University of Washington. Translational vision science & technology 2022;11(1):2.

[17] Mills RP, Budenz DL, Lee PP, Noecker RJ, Walt JG, Siegartel LR, Evans SJ, Doyle JJ. Categorizing the stage of glaucoma from pre-diagnosis to end-stage disease. American journal of ophthalmology 2006;141(1):24-30.

[18] Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences. 1997;55(1):119-139.

[19] McHugh L. Interrater reliability: the kappa statistic. Biochemia medica. 2012;22(3):276-82.

[20] Çorbacıoğlu ŞK, Aksel G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. Turkish journal of emergency medicine. 2023; 23(4): 195-198.