Subgroup Performance Analysis of an Interpretable Multimodal Framework for Glaucoma Screening

Amira SOLTANI
Esprit School Of Business
National Higher Engineering School of Tunis,
LR: LATICE
amira_soltani@yahoo.fr

Imed JABRI
National Higher Engineering School of Tunis,
LR: LATICE
imedjabri@yahoo.com

Abstract— **This paper analyzes the subgroup performance of an interpretable multimodal framework for glaucoma screening based on retinal fundus biomarkers and clinical data. Using an AdaBoost classifier selected through GridSearchCV with five-fold cross-validation, the model achieved 0.9516 accuracy, 0.9492 F1-score, and 0.9937 AUC on the test set. Subgroup analysis showed lower performance in younger patients, while results were more stable across sex. These findings highlight the importance of subgroup-level evaluation for assessing the reliability of lightweight glaucoma screening systems on small clinical datasets.**

Keywords— Glaucoma; subgroup analysis; multimodal classification; interpretability; fundus biomarkers; model robustness.

I. INTRODUCTION

Glaucoma is one of the leading causes of irreversible blindness worldwide, making early screening essential for preventing permanent vision loss [1], [2]. In recent years, artificial intelligence methods have shown promising results for glaucoma detection from retinal fundus images, especially when structural optic nerve head biomarkers are used for classification [3], [4].

However, most studies report performance mainly through global metrics such as accuracy, F1-score, or AUC. Although these measures are useful, they do not indicate whether the model performs consistently across different patient groups. In medical applications, high overall performance may still hide important subgroup variability, which can affect the practical reliability of the system [5], [6].

This issue is particularly relevant in small clinical datasets, where subgroup imbalance and patient heterogeneity may influence model behavior. In such cases, subgroup-level evaluation can provide a more realistic view of generalization and help identify populations for which the classifier is less stable [7].

In our previous work, we proposed an interpretable multimodal framework for glaucoma screening that combines retinal fundus biomarkers with demographic variables in a lightweight machine learning setting. Although the final AdaBoost classifier achieved strong overall performance, its behavior across patient subgroups was not investigated.

This paper extends that work by analyzing subgroup performance across age, sex, and origin categories. The goal is not to introduce a new predictive model, but to assess whether the strong overall performance remains consistent across patient profiles and to better understand the stability of the proposed framework.

The remainder of this paper is organized as follows. Section II reviews related work on glaucoma screening and subgroup-level evaluation in medical AI. Section III describes the dataset and evaluation protocol. Section IV presents the experimental results and discusses subgroup variability. Finally, Section V concludes the paper.

Related work

Automated glaucoma screening has been extensively studied through retinal fundus image analysis, as structural changes in the optic nerve head play a central role in glaucoma assessment. Early epidemiological and clinical studies highlighted both the global burden of glaucoma and the importance of early diagnosis, motivating the development of computer-aided screening systems [1], [2]. In this context, fundus imaging became particularly attractive because it provides a practical and non-invasive means of assessing optic disc morphology and cup-to-disc ratio in screening settings.

As the field evolved, glaucoma diagnosis progressively shifted from handcrafted feature extraction toward machine learning and deep learning approaches. A review by Hagiwara et al. showed that computer-aided glaucoma diagnosis from fundus images developed from traditional structural descriptors to more advanced learning-based frameworks while preserving clinically meaningful optic nerve head information [3]. More recently, Ashtari-Majlan et al. surveyed deep learning methods for glaucoma diagnosis across fundus, OCT, and visual field modalities, highlighting both the strong performance of recent models and the continuing challenges related to interpretability, data heterogeneity, and clinical deployment [8]. Similarly, Ling et al. reported in a systematic review and meta-analysis that deep learning achieves high diagnostic performance for glaucoma detection, while also emphasizing that external validation remains a major issue for reliable clinical translation [9].

Another important trend in the literature is the development of hybrid and multimodal screening frameworks. Sharma et al. proposed a hybrid multi-model artificial intelligence system for fundus-based glaucoma screening, showing that structured integration of several computational components can improve screening performance [10]. In parallel, Kovalyk-Borodyak et al. demonstrated that combining fundus-derived information with clinical variables improves patient representation and can enhance glaucoma detection [7]. These studies support the growing relevance of multimodal systems, especially in medical screening applications where demographic and clinical information may complement image-based biomarkers.

Beyond image-only pipelines, recent studies have also examined the diagnostic value of structured clinical and tabular data. Shahriari et al. reviewed machine learning approaches for glaucoma diagnosis based on tabular data and showed that demographic, clinical, and imaging-derived variables can provide strong predictive value when used in structured models [11]. This is particularly relevant for interpretable frameworks, where feature contribution and

subgroup behavior can be analyzed more directly than in purely end-to-end image-based models.

In addition to predictive performance, fairness and subgroup reliability have become increasingly important in medical AI. Shi *et al.* addressed this issue by proposing an equitable artificial intelligence framework for glaucoma screening with fair identity normalization, explicitly targeting disparities in model behavior across patient groups [12]. Their work illustrates a broader shift in evaluation philosophy: strong global performance alone is no longer considered sufficient, and subgroup consistency must also be examined when assessing the practical reliability of a screening system.

A related recent direction is the evaluation of foundation models for glaucoma screening. Chuter et al. assessed a foundation artificial intelligence model for glaucoma detection using color fundus photographs and analyzed its behavior across race, age, disease severity, and training size [13]. This work is particularly relevant to the present study because it highlights the importance of subgroup-level analysis and shows that model performance may vary across clinically meaningful patient categories even when overall results are strong.

Taken together, the literature points to three major trends: first, glaucoma screening has progressed from conventional feature-based analysis to high-performing deep learning systems; second, multimodal and tabular-data-based frameworks are increasingly recognized as clinically relevant; and third, fairness, subgroup consistency, and generalization are becoming essential evaluation criteria. However, relatively few studies have examined how a lightweight and interpretable multimodal framework behaves across age, sex, and origin categories in a small clinical dataset. The present work addresses this gap by focusing on subgroup-level performance analysis rather than proposing a new predictive architecture.

II. Materials and Methods

A. Dataset

This study was conducted using a clinical dataset of retinal fundus images collected at the Ophthalmology Hospital Hedi Raies in Tunis. The dataset includes glaucomatous and normal cases together with demographic and structural information relevant to glaucoma screening. For each sample, the available variables were age, sex, origin, cup-to-disc ratio (CDR), optic disc diameter (ODD), and the corresponding diagnostic label. The distribution of the full dataset and the subgroup composition considered in this study are summarized in Table I.

B. Feature Set and Final Model

The subgroup analysis was based on the final multimodal framework developed in our previous work, which combines explicit structural optic nerve head biomarkers with demographic variables for glaucoma classification. The features retained in the present analysis were age, sex, origin, CDR, and ODD. These variables were selected because they preserve clinical interpretability while allowing structured multimodal prediction.

The final classifier retained for subgroup evaluation was AdaBoost, an ensemble method that combines weak learners through iterative reweighting of misclassified samples and remains a strong baseline for structured binary classification tasks [14]. The model was selected because it achieved the best overall performance in our previous experiments and remains compatible with an interpretable feature-based screening framework.

C. Data Preprocessing and Model Training

Before training, categorical variables were encoded and numerical features were normalized using MinMax scaling. The dataset was then divided into 70% for training and 30% for testing using a stratified random split in order to preserve class proportions across subsets. Hyperparameter optimization was performed using GridSearchCV combined with five-fold cross-validation. This choice was intended to reduce model-selection bias and improve the reliability of the selected configuration, in line with established findings on overfitting in model selection and performance evaluation [15].

TABLE I. Distribution of the independent test set across classes and demographic subgroups

Category	Subgroup	N
Class Label	Normal	34
	Glaucoma	28
Age group	<50	17
	50-65	24
	>65	21
Gender	Male	30
	Female	32
Origin group	White	53
	Other	9
Test	Test set size	62

The hyperparameters explored for AdaBoost included the maximum depth of the decision tree base learner, the number of estimators, and the learning rate. The best-performing configuration used a maximum depth of 3, a learning rate of 1.0, and 150 estimators, with a mean cross-validation score of 0.9721. The final model was then evaluated on the independent test set. The overall workflow of the subgroup analysis is illustrated in Fig. 1.

D. Subgroup Definition

To assess subgroup-level behavior, the independent test set was divided according to demographic categories. Age was grouped into three intervals: patients younger than 50 years, patients aged 50 to 65 years, and patients older than 65 years. Sex was analyzed through male and female categories. Origin was also considered; however, because of strong imbalance between categories, the corresponding analysis was treated as exploratory and interpreted cautiously.

This subgroup-oriented design was motivated by recent work showing that a model with strong average performance may still underperform specific patient groups and that subgroup-aware evaluation is essential for reliable clinical AI [16], [17].

E. Evaluation Protocol

The final AdaBoost classifier was evaluated using accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). These metrics were computed both globally and separately for each subgroup. ROC-AUC was included because it provides a threshold-independent measure of discriminative ability and remains one of the standard tools for evaluating binary classifiers in medical prediction tasks [15].

The objective of this evaluation was not to introduce a new predictive model, but to determine whether the strong global performance of the final multimodal classifier remained stable across age, sex, and origin groups. This type of analysis is increasingly important in medical AI, where subgroup consistency and reliability are becoming key complements to global predictive performance [17], [18].

III. Results

A. Overall Performance of the Final Model

The final AdaBoost classifier selected through GridSearchCV and five-fold cross-validation achieved strong overall performance on the independent test set. The best hyperparameter configuration used a maximum tree depth of 3, a learning rate of 1.0, and 150 estimators, with a mean cross-validation score of 0.9721. On the test set, the model reached 0.9516 accuracy, 0.9032 precision, 1.0000 recall, 0.9492 F1-score, and 0.9937 AUC. These results confirm the strong discriminative ability of the multimodal framework and provide a reliable basis for subgroup-level analysis.

The corresponding performance values are summarized in Table II, while the ROC curve of the final classifier is shown in Fig. 2.

B. Performance Across Age Groups

To assess the stability of the model across age categories, the test set was divided into three groups: patients younger than 50 years, patients aged 50 to 65 years, and patients older than 65 years. The subgroup results revealed noticeable performance differences across these age intervals.

The model showed its lowest performance in the group younger than 50 years, with an accuracy of 0.824 and an F1-score of 0.667. In contrast, performance was substantially higher in the 50–65 and >65 groups, which achieved 0.958 and 0.952 accuracy, respectively. The corresponding subgroup metrics are reported in Table III, and the comparative values are illustrated in Fig. 3.

These findings indicate that the classifier did not behave uniformly across age groups. While the model remained highly effective in middle-aged and older patients, its predictive stability was lower in the youngest subgroup. This suggests that strong global performance may partially mask age-related variability that becomes visible only through subgroup-level analysis.

C. Performance Across Sex

The subgroup analysis by sex showed a more consistent pattern than the age-based evaluation. In female patients, the model achieved 0.900 accuracy and 0.880 F1-score, whereas in male patients it reached 0.938 accuracy and 0.938 F1-score. These results are summarized in Table IV and illustrated in Fig. 4.

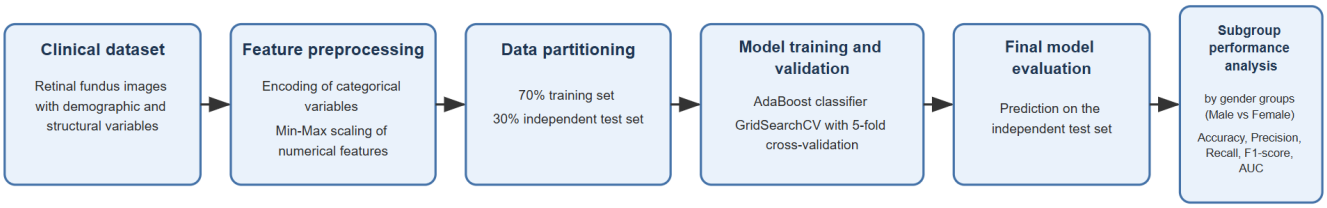

Fig. 1. Workflow of the subgroup analysis

TABLE II. Overall performance of the final AdaBoost classifier on the independent test set.

Accuracy	Precision	Recall	F1-score	AUC
0.9516	0.9032	1.0000	0.9492	0.9937

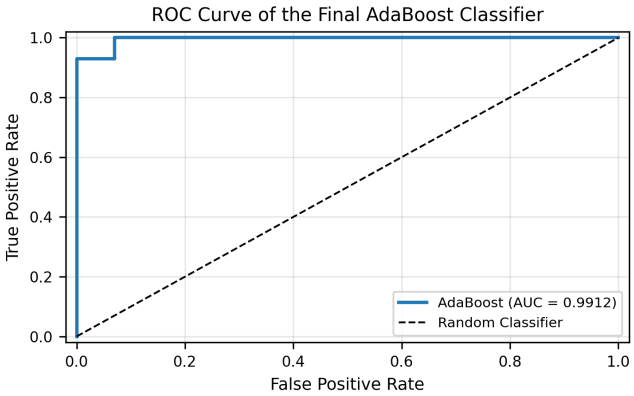

Fig. 2. ROC curve of the final AdaBoost classifier

Although performance was slightly higher in male patients, the overall difference between the two groups remained moderate. Compared with the stronger variability observed across age groups, the sex-based analysis suggests that the final classifier was relatively stable across this demographic factor.

D. Exploratory Analysis by Origin

An additional analysis was performed across origin categories. However, the distribution of the test set was strongly imbalanced, with most cases belonging to a single dominant category. For this reason, the origin-based analysis was considered exploratory rather than conclusive.

When the data were grouped into dominant and non-dominant origin categories, the classifier appeared to maintain high performance. Nevertheless, because the minority groups contained only a small number of cases, these results should be interpreted with caution. The subgroup imbalance prevents strong conclusions regarding origin-related consistency and highlights the need for larger and more balanced datasets for this type of analysis.

Table V reports the exploratory subgroup results across origin categories, which should be interpreted cautiously because of subgroup imbalance.

E. Summary of Subgroup Findings

Taken together, the subgroup analysis shows that the final AdaBoost classifier remained globally strong, but its behavior was not perfectly homogeneous across all patient categories. The most substantial variability was observed across age groups, with lower performance in younger patients.

By contrast, performance across sex remained broadly consistent, while the origin-based analysis remained limited by subgroup imbalance.

These results indicate that global performance metrics alone are not sufficient to fully characterize the reliability of the model. Even when the overall classifier performs well, subgroup-level evaluation can reveal clinically relevant variability that may affect practical screening use.

IV. DISCUSSION

The results of this study show that the final AdaBoost classifier achieved strong overall performance, but that this performance was not perfectly homogeneous across patient subgroups. This observation is important because high global accuracy and AUC do not necessarily imply stable behavior across clinically relevant patient categories. Recent work in clinical AI has increasingly emphasized that fairness, subgroup reliability, and practical robustness should be considered alongside global predictive performance when evaluating medical models [19],[20].

In the present study, the largest variability was observed across age groups. The model remained highly effective in the 50–65 and >65 groups, whereas lower performance was found in patients younger than 50 years. This suggests that the predictive behavior of the model may depend on subgroup composition, representation, or feature distribution. Such variability is consistent with recent studies showing that AI systems may underperform on specific patient groups even when their global results appear strong [20.

TABLE III. PERFORMANCE OF THE FINAL ADABOOST CLASSIFIER ACROSS AGE GROUPS

Age group	N	Accuracy	Precision	Recall	F1-score	AUC
<50	17	0.824	0.600	0.750	0.667	0.904
50–65	24	0.958	0.917	1.000	0.957	0.965
>65	21	0.952	1.000	0.923	0.960	1.000

TABLE IV. PERFORMANCE OF THE FINAL ADABOOST CLASSIFIER ACROSS SEX GROUPS

Sex	N	Accuracy	Precision	Recall	F1-score	AUC
Female	30	0.900	0.917	0.846	0.880	0.946
Male	32	0.938	0.882	1.000	0.938	0.959

TABLE V. PERFORMANCE OF THE FINAL ADABOOST CLASSIFIER ACROSS SEX GROUPS

Sex	N	Accuracy	Precision	Recall	F1-score	AUC
White	53	0.906	0.889	0.923	0.906	0.945
Other	9	1.000	1.000	1.000	1.000	1.000

By contrast, the sex-based analysis showed more consistent behavior. Although performance was slightly better in male patients than in female patients, the difference remained moderate and did not indicate a major instability across this factor. This is encouraging from a screening perspective, but it should still be interpreted cautiously. Recent fairness studies in medical AI have shown that even limited subgroup differences may become clinically important depending on the deployment setting, the subgroup size, and the cost of classification errors [19].

The origin-based analysis was less conclusive because of strong subgroup imbalance. Most cases in the test set belonged to a dominant category, while the remaining origin groups were represented by only a few samples. Under such conditions, subgroup metrics may become unstable and cannot support firm conclusions. This limitation is well aligned with broader concerns in the medical AI literature, where underrepresented groups often prevent reliable subgroup-level interpretation and may lead to overoptimistic conclusions if imbalance is not explicitly acknowledged [19], [20].

Another important aspect of the present study is the nature of the evaluated framework. Unlike large end-to-end black-box systems, the proposed model relies on explicit structural and demographic variables, namely CDR, ODD, age, sex, and origin. This makes subgroup analysis more transparent, since variations in performance can be interpreted in relation to understandable clinical and demographic inputs. Recent work on tabular-data-based glaucoma diagnosis supports the relevance of such interpretable approaches, particularly when subgroup behavior and variable contribution are part of the research objective.

The present findings also resonate with recent glaucoma AI studies showing that model evaluation is moving beyond simple global metrics. Foundation-model and hybrid-AI studies in glaucoma increasingly assess behavior across age, race, disease severity, or healthcare context, reflecting a broader shift toward more clinically grounded evaluation. In this sense, the current study contributes to this direction by showing that even a lightweight and interpretable multimodal model benefits from subgroup-level analysis.

This study nevertheless has several limitations. First, the dataset is relatively small and originates from a single clinical center, which limits the generalizability of the conclusions. Second, subgroup imbalance was substantial for origin categories, preventing robust interpretation for this factor. Third, the analysis was performed on one final model rather than comparing subgroup behavior across several classifiers. Finally, the present study identifies subgroup variability but does not attempt to mitigate it through fairness-aware training, subgroup rebalancing, or subgroup-aware optimization strategies. These directions remain important for future work.

Overall, the main contribution of this paper is to show that subgroup-level evaluation provides information that global metrics alone cannot capture. Even when a glaucoma screening system appears strong overall, meaningful variation may still exist across patient categories. For this reason, subgroup analysis should be considered an essential complement to global performance evaluation, particularly in small clinical datasets and screening-oriented medical AI applications.

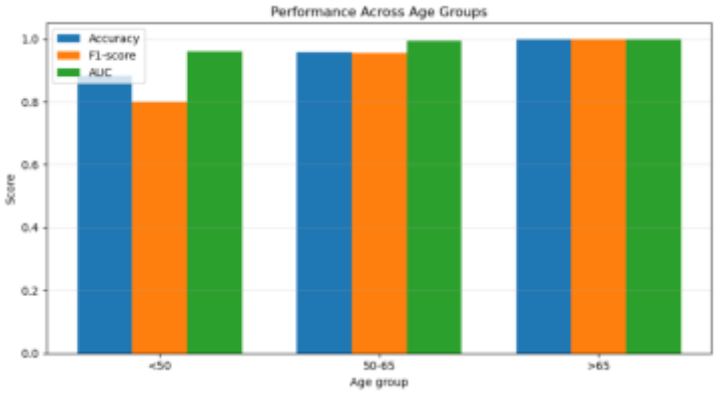

Fig. 3. Comparative performance of the final AdaBoost classifier across age groups.

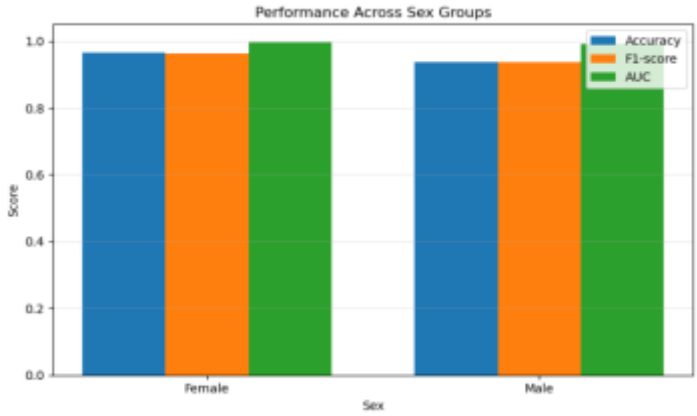

Fig. 4. Comparative performance of the final AdaBoost classifier across sex groups.

V. Conclusion

This paper examined the subgroup performance of an interpretable multimodal framework for glaucoma screening based on retinal fundus biomarkers and clinical data. Although the final AdaBoost classifier achieved strong overall performance, subgroup analysis showed that this performance was not fully uniform across patient categories.

The main variability was observed across age groups, with lower performance in patients younger than 50 years, while results were more stable in middle-aged and older patients. By contrast, performance across sex remained broadly consistent, whereas origin-based findings remained exploratory because of subgroup imbalance.

These results show that global metrics alone are not sufficient to assess the practical reliability of glaucoma screening models. Subgroup-level evaluation provides a more informative view of model stability and should be considered an essential complement to overall performance analysis. Future work will focus on larger and more balanced datasets to further assess and improve subgroup consistency.

References

[1] H. A. Quigley and A. T. Broman, "The number of people with glaucoma worldwide in 2010 and 2020," Br. J. Ophthalmol., vol. 90, no. 3, pp. 262–267, 2006, doi: 10.1136/bjo.2005.081224.

[2] F. Topouzis, A. L. Coleman, and A. I. Harris, "Factors Associated with Undiagnosed Open-Angle Glaucoma: The Thessaloniki Eye Study," Am. J. Ophthalmol., vol. 145, no. 2, pp. 327–335.e1, 2008, doi: 10.1016/j.ajo.2007.09.013.

[3] Y. Hagiwara, J. E. W. Koh, J. H. Tan, S. V. Bhandary, A. Laude, E. J. Ciaccio, L. Tong, and U. R. Acharya, "Computer-aided diagnosis of glaucoma using fundus images: A review," Comput. Methods Programs Biomed., vol. 165, pp. 1–12, 2018, doi: 10.1016/j.cmpb.2018.07.012.

[4] R. Hemelings, B. Elen, A. K. Schuster, M. B. Blaschko, J. Barbosa-Breda, P. Hujanen, A. Junglas, S. Nickels, A. White, N. Pfeiffer, P. Mitchell, P. De Boever, A. Tuulonen, and I. Stalmans, "A generalizable deep learning regression model for automated glaucoma screening from fundus images," npj Digit. Med., vol. 6, Art. no. 112, 2023, doi: 10.1038/s41746-023-00857-0.

[5] M. Liu, Y. Ning, S. Teixayavong, X. Liu, M. Mertens, Y. Shang, X. Li, D. Miao, J. Liao, J. Xu, *et al*., "A scoping review and evidence gap analysis of clinical AI fairness," *npj Digit. Med.*, vol. 8, Art. no. 360, 2025, doi: 10.1038/s41746-025-01667-2.

[6] Z. Xu, J. Li, Q. Yao, H. Li, M. Zhao, and S. K. Zhou, "Addressing fairness issues in deep learning-based medical image analysis: A systematic review," *npj Digit. Med.*, vol. 7, Art. no. 286, 2024, doi: 10.1038/s41746-024-01276-5.

[7] O. Kovalyk-Borodyak, J. Morales-Sánchez, R. Verdú-Monedero, and J.-L. Sancho-Gómez, "Glaucoma detection: Binocular approach and clinical data in machine learning," *Artif. Intell. Med.*, vol. 160, Art. no. 103050, 2025, doi: 10.1016/j.artmed.2024.103050.

[8] M. Ashtari-Majlan, M. M. Dehshibi, and D. Masip, "Glaucoma diagnosis in the era of deep learning: A survey," *Expert Syst. Appl.*, vol. 256, Art. no. 124888, 2024, doi: 10.1016/j.eswa.2024.124888.

[9] X. C. Ling, H. S.-L. Chen, P.-H. Yeh, Y.-C. Cheng, C.-Y. Huang, S.-C. Shen, and Y.-S. Lee, "Deep learning in glaucoma detection and progression prediction: A systematic review and meta-analysis," *Biomedicines*, vol. 13, no. 2, Art. no. 420, 2025, doi: 10.3390/biomedicines13020420.

[10] P. Sharma, N. Takahashi, T. Ninomiya, M. Sato, T. Miya, S. Tsuda, and T. Nakazawa, "A hybrid multi model artificial intelligence approach for glaucoma screening using fundus images," *npj Digit. Med.*, vol. 8, Art. no. 130, 2025, doi: 10.1038/s41746-025-01473-w.

[11] M. H. Shahriari, F. Asadi, H. Moghaddasi, A. Roshanpour, F. Sharifipour, and Z. Khorrami, "Applications of machine learning in glaucoma diagnosis based on tabular data: A systematic review," *BMC Biomed. Eng.*, vol. 7, Art. no. 9, 2025, doi: 10.1186/s42490-025-00095-3.

[12] M. Shi, Y. Luo, Y. Tian, L. Q. Shen, N. Zebardast, M. Eslami, S. Kazeminasab, M. V. Boland, D. S. Friedman, L. R. Pasquale, and M. Wang, "Equitable artificial intelligence for glaucoma screening with fair identity normalization," *npj Digit. Med.*, vol. 8, Art. no. 46, 2025, doi: 10.1038/s41746-025-01432-5.

[13] B. Chuter, J. Huynh, S. Hallaj, E. Walker, J. M. Liebmann, M. A. Fazio, C. A. Girkin, R. N. Weinreb, M. Christopher, and L. M. Zangwill, "Evaluating a foundation artificial intelligence model for glaucoma detection using color fundus photographs," *Ophthalmol. Sci.*, vol. 5, no. 1, Art. no. 100623, 2025, doi: 10.1016/j.xops.2024.100623.

[14] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997, doi: 10.1006/jcss.1997.1504.

[15] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[16] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, 2010.

[17] A. Subbaswamy, B. Sahiner, N. Petrick, V. Pai, R. Adams, M. C. Diamond, and S. Saria, "A data-driven framework for identifying patient subgroups on which an AI/machine learning model may underperform," *npj Digit. Med.*, vol. 7, Art. no. 334, 2024, doi: 10.1038/s41746-024-01275-6.

[18] Z. Xu, J. Li, Q. Yao, H. Li, M. Zhao, and S. K. Zhou, "Addressing fairness issues in deep learning-based medical image analysis: A systematic review," *npj Digit. Med.*, vol. 7, Art. no. 286, 2024, doi: 10.1038/s41746-024-01276-5.

[19] J. Pruthi, K. Khanna, and S. Arora, "Optic cup segmentation from retinal fundus images using Glowworm Swarm Optimization for glaucoma detection," Biomed. Signal Process. Control, vol. 60, Art. no. 102004, 2020, doi: 10.1016/j.bspc.2020.102004.

[20] J. Nayak, U. R. Acharya, P. S. Bhat, N. Shetty, and T.-C. Lim, "Automated diagnosis of glaucoma using digital fundus images," J. Med. Syst., vol. 33, no. 5, pp. 337–346, 2009, doi: 10.1007/s10916-008-9195-z.