Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol. 4 No. 4 (2025) pp: 4356-4362
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Predictive Modeling of Blindness Risk Using RAAB 2016 for Precision Eye Health

Stepanus Silaban, Putri Ghanim Septia Habiba
D3 Optometri, Akademi Refraksi Optisi dan Optometri Gapopin
godigitaledustar@gmail.com, putrighanim@gmail.com

Abstract

The Rapid Assessment of Avoidable Blindness (RAAB) surveys provide crucial information for planning and evaluating eye health initiatives, particularly in low- and middle-income countries where data systems are often limited. RAAB results are analyzed to estimate the prevalence of visual impairment and to assess cataract surgical coverage across populations. However, despite their rich individual-level data, RAAB surveys have rarely been explored for predictive modeling that could proactively identify people most vulnerable to blindness. This study sought to address that gap by developing and validating interpretable machine-learning models capable of predicting individuals at the highest risk of avoidable blindness. We used RAAB 2016 data collected from seven provinces across Indonesia, comprising a large and diverse sample of older adults. Two modeling approaches—a calibrated Extreme Gradient Boosting (XGBoost) algorithm and a Logistic LASSO regression—were trained and evaluated. Both models demonstrated outstanding discrimination (AUC ≈ 0.96) and strong calibration performance (Brier score ≈ 0.02), ensuring that predictions corresponded well to actual outcomes. Key predictors consistently selected across methods included increasing age, presence or absence of lens opacity, self-reported functional difficulty in seeing or mobility, and lack of corrective spectacles. To enhance usability in field settings, we also derived a simplified point-score tool from the LASSO model. Decision-curve analysis confirmed that the model could offer substantial clinical and operational benefit by guiding targeted outreach where resources are limited. Overall, this work highlights predictive analytics as promising extension of the RAAB framework, enabling more precise and efficient public eye health strategies in Indonesia.

Key words: Rapid Assessment of Avoidable Blindness (RAAB); Machine Learning; Blindness Prediction; Logistic LASSO; XGBoost.

1. Introduction

Visual impairment and blindness remain major global public health challenges, particularly in low- and middle-income countries (LMICs) where avoidable causes such as cataract, refractive error, and diabetic retinopathy persist [1], [2], [3]. These conditions disproportionately affect older adults and underserved communities with limited access to eye care services. Without timely detection and intervention, preventable visual loss can lead to disability, reduced quality of life, and significant socioeconomic burden. The Rapid Assessment of Avoidable Blindness (RAAB) methodology, developed by the International Centre for Eye Health, provides standardized, population-based estimates of blindness and visual impairment to inform national eye health programs [4]. RAAB surveys have now been implemented in more than 70 countries, offering an efficient and cost-effective approach for monitoring eye health indicators. As a result, RAAB has become a cornerstone for planning, evaluation, and advocacy in global blindness prevention initiatives.

Indonesia has conducted multiple Rapid Assessment of Avoidable Blindness (RAAB) surveys across its provinces to monitor progress in eye health and evaluate cataract surgical coverage [5]. These surveys have provided essential epidemiological data on the prevalence and causes of blindness, revealing significant inter-provincial variation linked to access and service availability [2], [4]. However, most RAAB analyses have remained descriptive, focusing on population-level prevalence, cataract surgical coverage, and cause distributions rather than individual-level risk modelling [4]. Incorporating predictive analytics into RAAB data could identify individuals or clusters at highest risk of blindness, enabling more efficient and evidence-based targeting of outreach and surgical interventions [1]. Such an approach aligns with the emerging paradigm of precision public health, which leverages data-driven tools to optimize health resources and reduce avoidable vision loss in low-resource settings [6], [7].

Recent advances in machine learning (ML) offer the opportunity to develop predictive tools capable of identifying those at greatest risk [7]. Models such as Extreme Gradient Boosting (XGBoost) and Logistic LASSO regression can uncover complex, nonlinear relationships while maintaining interpretability through SHAP (SHapley Additive Explanations) and coefficient-based inference [8], [9], [10]. However, applications of ML in ophthalmology have mainly focused on imaging—such as diabetic retinopathy and glaucoma detection [11], [12]—with limited use in population-based, non-imaging datasets like RAAB.

This study bridges that gap by developing and validating predictive models for individual-level blindness risk using pooled RAAB 2016 data from seven Indonesian provinces. The analysis leverages a large and diverse sample of older adults, enabling robust modeling that reflects real-world population characteristics. The models integrate demographic, clinical, and functional variables—such as age, lens status, spectacle use, and self-reported difficulties in seeing or mobility—to generate accurate and interpretable predictions of blindness risk [4], [5]. By applying modern machine-learning algorithms, including Logistic LASSO regression and XGBoost, the study advances beyond traditional descriptive epidemiology toward predictive analytics in community eye health [8], [9]. Importantly, the framework emphasizes interpretability through SHAP values and simplified point-score derivation, ensuring practical utility for non-specialist field users [10]. This approach aligns with the growing need for tools that support rapid, actionable decision-making in low-resource environments. Overall, the predictive modeling strategy supports a precision public health paradigm, where data-driven insights can guide more efficient targeting of blindness prevention and cataract outreach programs in resource-limited settings [6], [7].

2. Method

2.1. Data Source

Data for this study were obtained from the Rapid Assessment of Avoidable Blindness (RAAB) 2016 surveys conducted in seven Indonesian provinces: North Sumatra, West Sumatra, West Papua, South Kalimantan, Nusa Tenggara Timur, North Sulawesi, and Maluku [5]. Each provincial survey adhered to standardized RAAB version 6 protocols developed by the International Centre for Eye Health, ensuring methodological consistency across diverse geographic and healthcare settings [4]. The surveys employed multistage cluster sampling to recruit participants aged 50 years and older, representing the population group at highest risk for blindness and visual impairment [2]. Visual acuity testing, lens status assessment, and interviews using structured questionnaires were conducted by trained ophthalmic personnel. These combined data form a robust foundation for developing and evaluating individual-level predictive models of presenting blindness in Indonesia.

Data collection included demographic characteristics, distance visual acuity measurements, lens status classification, and self-reported functional difficulties such as problems with seeing or mobility. All information was recorded using structured electronic questionnaires to support accuracy and completeness. The resulting harmonized datasets were pooled into a unified database for this analysis, enabling cross-provincial modeling of individual-level risk of presenting blindness and supporting broader generalizability of the findings across Indonesia.

2.2. Variables and Outcome

The primary outcome variable, presenting blindness, was defined according to World Health Organization criteria as best-eye presenting visual acuity of ≤3/60, measured using standardized RAAB assessment protocols to ensure reliability and comparability across settings [2], [4]. Presenting vision was used rather than best-corrected vision to reflect the practical level of visual function experienced in everyday activities. This definition captures blindness that could potentially be mitigated through interventions such as cataract surgery or refractive correction. It also aligns with global programmatic priorities focused on reducing avoidable causes of severe visual loss. Overall, the outcome reflects a clinically meaningful threshold that informs public health planning and access to eye care services.

Predictors included demographic characteristics such as age and sex, clinical variables including lens status, spectacle use, and self-reported history of diabetes, and functional indicators capturing daily visual and physical capacity. These functional variables comprised self-reported difficulty in seeing, mobility, hearing, memory, communication, and self-care, following the standardized Washington Group Short Set framework for disability assessment [13]. Incorporating these measures allowed the model to reflect not only clinical eye health status but

DOI: https://doi.org/10.31004/riggs.v4i4.4235

also the broader functional implications of visual loss. The selected predictors represent information that can be feasibly collected in rapid community surveys, making them suitable for deployment in low-resource settings. They were chosen based on established or hypothesized associations with vision loss documented in population-based ophthalmic research [1], [3]. Together, these variables provide a comprehensive and practical foundation for risk stratification in community eye health programs.

2.3. Model Development

Two complementary predictive models were developed to assess individual risk of presenting blindness: a Logistic Least Absolute Shrinkage and Selection Operator (LASSO) regression model for interpretable feature selection [8], and an Extreme Gradient Boosting (XGBoost) model for high-performance ensemble learning [9]. The LASSO model provided a sparse and transparent set of predictors, facilitating clinical interpretation and potential field implementation. In contrast, the XGBoost model leveraged non-linear interactions and ensemble learning to maximize predictive accuracy. The use of both linear and non-linear approaches allowed evaluation of model performance from complementary perspectives while balancing interpretability and predictive power. To ensure robust and unbiased estimation, cluster-aware cross-validation was applied, preventing data leakage between geographically grouped samples. This approach also preserved the survey design structure, ensuring that model evaluation accurately reflected real-world applicability across provinces.

The XGBoost model's predicted probabilities were further calibrated using isotonic regression to enhance reliability of probabilistic outputs. Model performance was assessed using standard discrimination and calibration metrics, including Area Under the Receiver Operating Characteristic Curve (AUC), Average Precision (AUPRC), and Brier score. In addition, model interpretability was evaluated using SHapley Additive exPlanations (SHAP) values to quantify the contribution of each predictor to the risk of presenting blindness [10].

2.4. Decision and Scoring Tools

Decision-curve analysis (DCA) was conducted to assess the clinical and operational utility of the predictive models by quantifying their net benefit across a range of probability thresholds [14], [15]. This approach enabled evaluation of whether model-driven screening strategies offered greater advantage than default options such as "screen all" or "screen none." By incorporating real-world trade-offs between false positives and false negatives, DCA provided a more comprehensive view of model performance beyond traditional accuracy metrics. The analysis also allowed consideration of resource-limited settings, where unnecessary screening carries substantial operational costs. Through comparing net benefit curves, DCA helped identify the threshold ranges where model-guided decisions were most advantageous. Overall, DCA offered an evidence-based framework for determining the practical value and feasibility of implementing predictive models in applied settings.

The final Logistic LASSO regression coefficients were translated into a simplified integer point-score algorithm. This transformation enables easy calculation of individual blindness risk in the field. Health workers can use the point-score system without relying on complex software or computational resources. As a result, the tool supports rapid and practical risk assessment in community-based screening programs.

The resulting score-to-risk lookup table provides field workers with an intuitive tool to approximate predicted probabilities in community screening programs. This approach allows rapid identification of individuals at highest risk of presenting blindness without the need for digital devices. By facilitating manual risk estimation, the point-score system improves the efficiency of screening workflows in resource-limited settings. It also enhances scalability and adoption of targeted blindness prevention strategies across diverse communities. Overall, the tool supports more practical and data-driven allocation of limited eye health resources [16].

3. Result and Discussion

3.1. Dataset Overview

The pooled analytical dataset comprised 20,382 participants aged 50 years and older from seven Indonesian provinces included in the 2016 RAAB surveys. The mean participant age was 60 years (SD ± 8.9]), and females represented 58.3% of the sample.

DOI: https://doi.org/10.31004/riggs.v4i4.4235

3.2. Model Performance

Both the XGBoost and Logistic LASSO models demonstrated excellent predictive performance in identifying individuals at risk of presenting blindness. Figure 2 depicts that the XGBoost model achieved an area under the receiver operating characteristic curve (AUC) of 0.957, while the Logistic LASSO model achieved an AUC of 0.956, indicating near-perfect discrimination. The average precision (AUPRC) for both models was 0.996, reflecting exceptional ranking ability in the context of class imbalance. The Brier score of 0.024 further confirmed high overall calibration accuracy and reliability of predicted probabilities. Visual inspection of calibration plots as shown in Figure 1, showed close alignment between predicted and observed risks across probability deciles, underscoring the robustness and clinical interpretability of both models [8], [9].

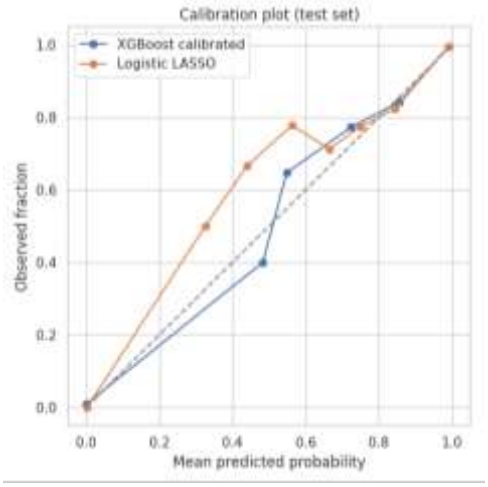

Figure 1. Calibration (Test Set)

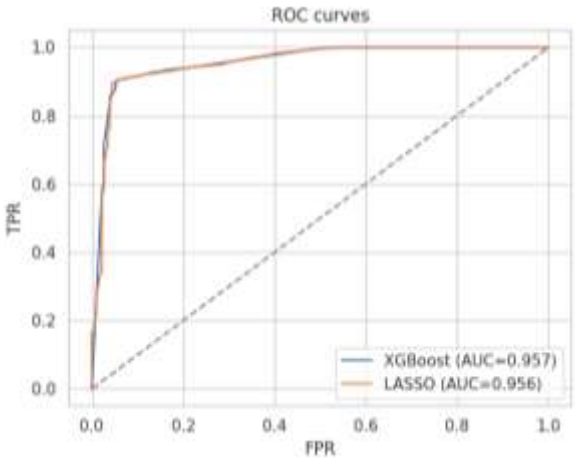

Figure 2. ROC Curve

3.3. Feature Importance

Feature importance analysis revealed that the most influential predictors of presenting blindness included lens opacity or absence, older age, self-reported difficulty in seeing or mobility, and non-use of spectacles. These variables were consistently identified as high-impact features across both the XGBoost and Logistic LASSO models, indicating strong agreement between linear and nonlinear modeling approaches (Figure 3). Figure 4 shows that SHAP summary plots further confirmed the relative importance and directionality of these predictors, illustrating that impaired lens status and functional limitations substantially increased predicted blindness risk [10]. The convergence of findings across models underscores the robustness and interpretability of the proposed predictive framework for practical application in community eye health.

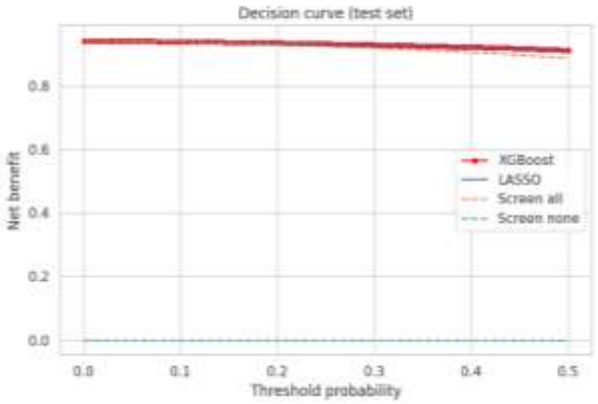

Figure 3. SHAP Summary

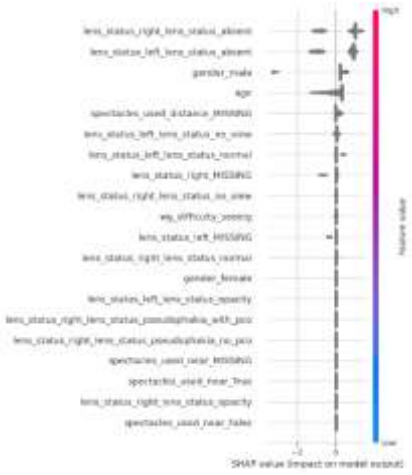

Figure 4. SHAP Summary

DOI: https://doi.org/10.31004/riggs.v4i4.4235

3.4. Operational Tools

A simplified integer point-score system as depicted in Table 1, derived from the Logistic LASSO model was developed to translate complex model coefficients into a practical, field-ready tool for estimating individual blindness risk. Each retained predictor was assigned a weighted integer score proportional to its coefficient magnitude, allowing manual calculation of total risk without computational resources [16]. The resulting score-to-risk lookup table enables community health workers to estimate probabilities of blindness directly from patient characteristics, supporting scalable screening strategies in low-resource settings. Decision-curve analysis demonstrated that targeting the top 1% of individuals predicted at highest risk could identify approximately 40 true-positive cases with minimal false positives, markedly improving screening efficiency. These operational tools bridge advanced analytics and real-world applicability, facilitating data-driven prioritization in blindness prevention and outreach programs [15], [17].

Table 1. Point-score System (10 of 100 rows)

score_raw	pred_prob
-6.959218428	0.0005284136138
-6.823701343	0.0006134617320
-6.688184257	0.0007121885824
-6.552667172	0.0008267907953
-6.417150087	0.0009598165726
-6.281633001	0.0011142215480
-6.146115916	0.0012934333780
-6.010598831	0.0015014263820
-5.875081745	0.0017428077290
-5.739564660	0.0020229168570

3.5. Discussion

This study demonstrates the feasibility and utility of applying machine learning (ML) techniques to Rapid Assessment of Avoidable Blindness (RAAB) datasets for predictive assessment of individual blindness risk across multiple Indonesian provinces. Both XGBoost and Logistic LASSO models achieved exceptionally high discrimination and strong calibration, validating their robustness for population-level eye health prediction. These findings extend the traditional descriptive focus of RAAB toward predictive analytics, enabling a shift from population prevalence estimation to individual risk stratification. The approach underscores the potential of ML as a practical tool to inform targeted interventions in low- and middle-income country (LMIC) contexts where the burden of blindness remains high.

The most influential predictors—lens opacity, older age, and functional difficulties in mobility or seeing—are consistent with established determinants of visual impairment and blindness reported in global studies [1], [18]. Lens opacity, primarily caused by cataract, directly reduces visual acuity and remains a leading cause of avoidable blindness in Indonesia. Older age reflects cumulative biological changes and increased vulnerability to ocular conditions, compounding the risk of vision loss. Functional difficulties in mobility or seeing capture the broader impact of visual impairment on daily life and may indicate both severity and access-related barriers. By identifying these key risk factors, the models provide actionable guidance for targeted interventions and efficient allocation of eye health resources.

This study introduces two key methodological innovations. First, it applies advanced machine learning algorithms to non-imaging, population-based RAAB data, demonstrating that high-accuracy predictive modeling is achievable without dependence on costly diagnostic imaging. Second, it develops a field-adaptable, integer-based point-score system for estimating blindness risk, enabling non-specialist health workers to conduct data-driven screening. Together, these innovations combine algorithmic precision with operational interpretability. This integration effectively bridges the gap between data science and practical community ophthalmology.

DOI: https://doi.org/10.31004/riggs.v4i4.4235

Compared with traditional RAAB analyses, which focus mainly on estimating the prevalence and causes of blindness, this predictive targeting framework represents a paradigm shift toward precision public health. By directing outreach and screening efforts toward individuals or clusters identified as highest risk, programs can substantially increase the detection yield per unit of effort. This approach also enables more efficient allocation of scarce resources in low-capacity health systems. Decision-curve analysis further confirmed the operational efficiency of the models under realistic programmatic constraints. The results demonstrated measurable gains in net benefit compared with conventional "screen all" or "screen none" strategies. Collectively, these findings highlight the potential of predictive analytics to transform community eye-health planning.

Overall, these results suggest that predictive RAAB models could be integrated into national blindness prevention programs, supporting evidence-based prioritization for cataract surgery and refractive services. The approach provides a scalable framework for LMICs, where resources are constrained, and highlights the value of combining robust epidemiologic data with machine-learning techniques to enhance public eye health strategies.

4. Conclusion

This study demonstrates that interpretable machine-learning models trained on Rapid Assessment of Avoidable Blindness (RAAB) data can accurately predict individual-level risk of presenting blindness and guide targeted community outreach. By integrating demographic, clinical, and functional indicators, the models achieved excellent discrimination and calibration across diverse Indonesian provinces, confirming their robustness for population-based applications. These findings highlight the feasibility of using non-imaging, survey-based data for predictive modeling in eye health. The models provide actionable insights that can inform resource allocation and improve the efficiency of cataract and vision-restoration programs. Importantly, the framework supports the identification of both high-risk individuals and clusters, enabling more precise targeting of interventions. Overall, this approach represents a significant advancement toward precision public health in low- and middle-income country (LMIC) contexts. The ability to identify individuals and clusters at highest risk provides a valuable foundation for precision public health planning. By focusing on those most vulnerable, health programs can design interventions that are both timely and effective. Targeted strategies based on these predictions can optimize resource allocation and improve the efficiency of cataract and vision-restoration programs. This approach ensures that limited resources are directed toward populations with the greatest need, maximizing the overall impact of eye health initiatives. The combination of high-performance algorithms, transparent explainability techniques, and field-ready operational tools represents a scalable innovation for blindness prevention, particularly in low- and middle-income countries (LMICs) such as Indonesia. These approaches integrate advanced data science with practical considerations, ensuring applicability in real-world community health settings. The LASSO-derived point-score enables frontline health workers to estimate individual risk without relying on advanced technology or software. Meanwhile, decision-curve analysis provides an evidence-based framework to guide prioritization of interventions under resource constraints. Together, these tools enhance both the efficiency and effectiveness of targeted blindness prevention programs. Collectively, these advances bridge the gap between data-driven modeling and practical eye-care delivery. By combining high-performance machine-learning models with interpretable and field-ready tools, the approach ensures that predictive insights can be effectively translated into community action. The framework allows health workers to identify individuals and clusters at highest risk, enabling more focused and efficient interventions. It offers a replicable method for integrating predictive analytics into future RAAB surveys, enhancing the utility of population-based eye health data. Additionally, this approach can inform national vision health strategies by supporting evidence-based planning and resource allocation. Ultimately, these innovations promote more precise, equitable, and impactful blindness prevention initiatives across low- and middle-income country contexts.

References

[1] S. R. Flaxman *et al.*, "Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis," *Lancet Glob. Heal.*, vol. 5, no. 12, pp. e1221–e1234, Dec. 2017, doi: 10.1016/S2214-109X(17)30393-5.
[2] "World health statistics 2019: monitoring health for the SDGs, sustainable development goals." Accessed: Oct. 12, 2025. [Online]. Available: https://www.who.int/publications/i/item/9789241565707
[3] D. Pascolini and S. P. Mariotti, "Global estimates of visual impairment: 2010," *Br. J. Ophthalmol.*, vol. 96, no. 5, pp. 614–618, May 2012, doi: 10.1136/BJOPHTHALMOL-2011-300539.
[4] I. Mactaggart, H. Limburg, A. Bastawrous, M. J. Burton, and H. Kuper, "Rapid Assessment of Avoidable Blindness: Looking back, looking forward," *Br. J. Ophthalmol.*, vol. 103, no. 11, pp. 1549–1552, Nov. 2019, doi: 10.1136/BJOPHTHALMOL-2019-314015.
[5] "RAAB | Rapid Assessment of Avoidable Blindness." Accessed: Oct. 13, 2025. [Online]. Available: https://www.raab.world/
[6] M. J. Khoury, G. L. Armstrong, R. E. Bunnell, J. Cyril, and M. F. Iademarco, "The intersection of genomics and big data with public health: Opportunities for precision public health," *PLoS Med.*, vol. 17, no. 10, p. e1003373, Oct. 2020, doi:

DOI: https://doi.org/10.31004/riggs.v4i4.4235

10.1371/JOURNAL.PMED.1003373.
[7] D. S. W. Ting *et al.*, "Deep learning in ophthalmology: The technical and clinical considerations," *Prog. Retin. Eye Res.*, vol. 72, p. 100759, Sep. 2019, doi: 10.1016/J.PRETEYERES.2019.04.003.
[8] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *J. R. Stat. Soc. Ser. b-methodological*, vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/J.2517-6161.1996.TB02080.X.
[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, Mar. 2016, doi: 10.1145/2939672.2939785.
[10] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, pp. 4766–4775, May 2017, Accessed: Oct. 05, 2025. [Online]. Available: https://arxiv.org/pdf/1705.07874
[11] R. Poplin *et al.*, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nat. Biomed. Eng.*, vol. 2, no. 3, pp. 158–164, Mar. 2018, doi: 10.1038/S41551-018-0195-0;TECHMETA.
[12] S. Keel *et al.*, "Keeping an eye on eye care: monitoring progress towards effective coverage," *Lancet Glob. Heal.*, vol. 9, no. 10, pp. e1460–e1464, Oct. 2021, doi: 10.1016/S2214-109X(21)00212-6.
[13] "UNICEF/ Washington Group on Disability Statistics Child Functioning Module 1," 2016.
[14] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," *Med. Decis. Mak.*, vol. 26, no. 6, pp. 565–574, Nov. 2006, doi: 10.1177/0272989X06295361.
[15] A. J. Vickers, B. van Calster, and E. W. Steyerberg, "A simple, step-by-step guide to interpreting decision curve analysis," *Diagnostic Progn. Res.*, vol. 3, no. 1, pp. 1–8, Dec. 2019, doi: 10.1186/S41512-019-0064-7/FIGURES/3.
[16] "Applying 'Lasso' Regression to Predict Future Visual Field Progression in Glaucoma Patients | IOVS | ARVO Journals." Accessed: Oct. 12, 2025. [Online]. Available: https://iovs.arvojournals.org/article.aspx?articleid=2289280
[17] A. J. Vickers and E. B. Elkin, "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models," *Med. Decis. Mak.*, vol. 26, no. 6, pp. 565–574, Nov. 2006, doi: 10.1177/0272989X06295361.
[18] M. J. Burton *et al.*, "The Lancet Global Health Commission on Global Eye Health: vision beyond 2020," *Lancet Glob. Heal.*, vol. 9, no. 4, pp. e489–e551, Apr. 2021, doi: 10.1016/S2214-109X(20)30488-5.

DOI: https://doi.org/10.31004/riggs.v4i4.4235