ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14848301

Why Artifical Intelligence Models Struggle with Glaucoma Detection in Real-World Settings?

Yuv R Pantha

South Asia Regional Institute for Professionals in Ophthalmology
Delaware, United States

Publication Date: 2025/02/11

Abstract: Glaucoma, a leading cause of irreversible blindness, is characterized by progressive loss of retinal ganglion cells (Lee & Mackey, 2022). Early detection and intervention are crucial to prevent vision loss, but diagnosing glaucoma remains a challenge due to its heterogeneous nature and varied presentation (Križaj, 2019). One of the key challenges in diagnosing glaucoma is the lack of consensus within the medical community on standardized diagnostic criteria and treatment protocols (Kroese & Burton, 2003). In recent years, artificial intelligence (AI), particularly deep learning (DL) models, has shown promise in improving glaucoma detection by processing large datasets of ocular images and clinical data (Zhang, Tang, Xia, & Cao, 2023). In theory, AI systems could automate and enhance the accuracy of glaucoma diagnosis, reduce the burden on healthcare professionals, and enable earlier interventions that could prevent vision loss. However, variability in diagnostic thresholds and interpretation methods due to the lack of consensus contributes to inconsistencies across clinical practices. This research explores how these discrepancies in diagnosis, particularly using of retinal nerve fiber layer (RNFL) thickness data contributes to the failure of AI-based glaucoma prediction when applied to real world settings. The paper further discusses the ethical, legal, and clinical implications of AI glaucoma models and suggests that standardized diagnostic criteria and improved collaboration among ophthalmologists and AI developers are essential for enhancing the reliability and applicability of AI in glaucoma detection and management.

Keywords: *Glaucoma, Artificial Intelligence, Deep Learning, Optical Coherence Tomography (OCT), Lack of Consensus in Glaucoma, Variability in Glaucoma Diagnosis, AI-Based Glaucoma Prediction, Ethical Implications, Legal Implications.*

How to Cite: Yuv R Pantha (2025). Why Artifical Intelligence Models Struggle with Glaucoma Detection in Real-World Settings?. *International Journal of Innovative Science and Research Technology*, 10(1), 2102-2105. https://doi.org/10.5281/zenodo.14848301

I. INTRODUCTION

Glaucoma is one of the leading causes of irreversible blindness worldwide, characterized by progressive damage to the optic nerve. The disease often progresses silently without early symptoms, making it difficult to detect and manage effectively without regular eye exams (Križaj, 2019). Early detection and treatment are essential for preventing vision loss, but diagnosing glaucoma remains a challenge due to its multifaceted nature and varying presentation across individuals. The diagnosis of glaucoma traditionally relies on methods such as intraocular pressure measurement, visual field testing, and imaging techniques like optical coherence tomography (OCT). However, these approaches require specialized expertise, are time-consuming, and may sometimes be prone to human error, particularly in the interpretation of complex data (Wishart, 2009).

In response to these challenges, artificial intelligence (AI) has gained traction as a potential solution to improve glaucoma diagnosis and prediction. AI models, particularly deep learning (DL), have been trained on vast datasets of ocular images, clinical measurements, and patient data to predict glaucoma risk, and detect early-stage disease. These AI models are capable of processing large amounts of data at high speeds, identifying subtle patterns and anomalies that may be missed by human practitioners (Shen, Wu, & Suk, 2017). In theory, AI systems could automate and enhance the accuracy of glaucoma diagnosis, reduce the burden on healthcare professionals, and enable earlier interventions that could prevent vision loss.

Despite the promising potential of AI in glaucoma prediction, there are significant obstacles to its deployment in real-world clinical settings. One of the key challenges is the lack of consensus within the medical community on standardized diagnostic criteria and treatment protocols. Glaucoma is a heterogeneous disease, and there is significant variation in how it is diagnosed and managed across different healthcare systems and populations (Ghaffar, 2024). This variability can contribute to under-diagnosis in some settings where the criteria for diagnosing glaucoma are too strict. In contrast, some practices may have lower thresholds for diagnosing glaucoma based on certain criteria (like IOP, optic nerve appearance, or visual field loss) which can result in patients being diagnosed with glaucoma unnecessarily.

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14848301

Without a unified framework for diagnosis and treatment, AI models cannot be aligned with the diverse clinical practices and patient characteristics encountered in everyday real-world healthcare settings.

II. VARIABILITY IN DIAGNOSIS OF GLAUCOMA

The diagnosis of glaucoma is often inconsistent across clinical practices due to the lack of a universal agreement on diagnostic criteria and disease management protocols. Although there are established diagnostic tests, such as measuring intraocular pressure (IOP), assessing the optic nerve head, and conducting visual field tests, the interpretation of these results can vary significantly between clinicians and institutions (Ahmed et al., 2016). One major issue is the absence of standardized thresholds for key diagnostic parameters. For instance, the IOP threshold for diagnosing glaucoma is not universally agreed upon. Some practitioners consider an IOP above 21 mmHg as indicative of potential glaucoma, while others may diagnose the condition with lower or higher values based on individual patient risk factors and clinical judgment.

Furthermore, glaucoma is a heterogeneous disease, manifesting differently across patient populations, which complicates its diagnosis. There is considerable variability in how practitioners assess optic nerve damage or retinal nerve fiber layer (RNFL) thinning, with differences in imaging technologies and assessment methods influencing diagnostic outcomes. Even within the same institution, the subjective interpretation of fundus photographs or optical coherence tomography (OCT) images can lead to over diagnosis or under diagnosis of the condition. Additionally, variations in the interpretation of visual field tests such as the criteria for defining glaucomatous damage can result in inconsistent diagnoses.

III. DIAGNOSING GLAUCOMA USING RNFL THICKNESS

While OCT is a valuable diagnostic tool, it has several limitations that can contribute to over diagnosis. OCT measurements vary depending on age, ethnicity, and optic disc anatomy, and the thresholds for defining abnormal findings can differ between devices and protocols (Budenz et al., 2007). This variability makes it challenging to establish universally accepted diagnostic cutoffs that reliably distinguish between true glaucomatous changes and normal anatomical variation. For instance, an RNFL thickness considered normal in one population (e.g., a thicker RNFL in younger individuals) might be interpreted as abnormal in another group (e.g., elderly patients or those of a different ethnic background). Without universally accepted criteria or population-specific reference ranges, OCT findings may be over-interpreted, leading to diagnoses of glaucoma in individuals who do not have the condition.

Changes in the RNFL can occur for reasons other than glaucoma, such as diabetic retinopathy, optic neuropathies, or retinal vascular diseases (Oshitari, Hanawa, & Adachi-Usami, 2009). OCT does not always differentiate between glaucomatous and non-glaucomatous causes of optic nerve changes, which could lead to over diagnosis if clinicians fail to consider other potential diagnoses. Misinterpretation of RNFL thinning due to non-glaucomatous causes could result in incorrect diagnoses and unnecessary treatment for glaucoma.

IV. USING OCT IN DEEP LEARNING MODEL

Using OCT data from over diagnosed or underdiagnosed cases of glaucoma can significantly affect the performance of AI models designed to predict glaucoma. In deep learning, the quality and accuracy of the training data are crucial, as the model learns to make predictions based on patterns observed in the dataset (Whang & Lee, 2020). When training data includes over diagnosed or underdiagnosed cases—patients incorrectly diagnosed with glaucoma based on OCT scans—the model will learn from these faulty data points, leading to several potential issues in its predictions.

A. Training Model With Over-Diagnosed Cases & Its Impact

Over diagnosis involves incorrectly labeling patients as having glaucoma, even if they do not have the disease. These false positives represent cases where structural changes in the optic nerve or RNFL are interpreted as glaucomatous damage, even though they are not.

When OCT data from over diagnosed cases are used to train an AI model, the model may learn to associate certain features (such as mild RNFL thinning, optic nerve head changes, or other subtle OCT findings) with glaucoma, even though these features are not truly indicative of the disease. As a result, the AI model may "learn" the wrong patterns from these over diagnosed cases, causing it to incorrectly classify healthy individuals (or those with only normal aging or non-glaucomatous changes) as having glaucoma. This leads to false positives and reduces the model's specificity, making it less accurate at distinguishing healthy individuals from those with actual glaucoma.

B. Imbalanced Training Data

Over diagnosis can lead to an imbalanced dataset, where there are too many false positives (over diagnosed cases) and not enough true negatives (healthy patients) or true positives (actual glaucoma patients). This imbalance can significantly affect the AI model's performance in predicting inaccurately (Johnson & Khoshgoftaar, 2019). For instance: False positives may dominate the dataset, leading the model to incorrectly predict glaucoma in many patients. There may not be enough true negatives, making it difficult for the model to distinguish between healthy individuals and those with actual glaucoma, which reduces the model's specificity.

C. Model Sensitivity

Over diagnosis can cause the AI model to become highly sensitive, meaning it may flag many potential glaucoma cases but with poor specificity. While the model may correctly identify many true glaucoma patients, it will also incorrectly flag many healthy individuals or people with other eye conditions as having glaucoma. When such model is deployed, it could result in unnecessary follow-up tests,

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14848301

treatments, or surgeries for individuals who do not actually have the disease.

The model may become overly sensitive to small changes in the RNFL or optic nerve, which could be due to factors other than glaucoma (e.g., normal variations, ethnic differences, or age-related changes). In such cases, the model may misinterpret these normal variations as signs of glaucoma, leading to over diagnosis.

D. Risk of Overfitting

AI models, particularly those based on deep learning or convolutional neural networks (CNNs), are highly susceptible to over fitting if the training dataset is not carefully curated. Over fitting occurs when a model learns the specific details or "noise" in the training data rather than the underlying patterns that generalize to unseen data (Montesinos López, Montesinos López, & Crossa, 2022). In the case of over diagnosis, the model may "memorize" features that are irrelevant to true glaucoma, such as minor RNFL changes that are normal for a particular age group or ethnic background. As a result, the model may perform poorly on new or diverse data, as it will rely on spurious correlations that do not hold in different populations.

E. Reduced Model Generalization

If the AI model is trained primarily on over diagnosed cases, it may fail to generalize well to other populations (Jakubovitz, Giryes, & Rodrigues, 2019), leading to poor performance when encountering data from underdiagnosed or true glaucoma patients. For example, the model may perform well in identifying glaucoma in a biased sample where many individuals have been over diagnosed, but it might fail to identify true glaucoma in patients with less obvious signs or in populations not represented in the training set.

F. Failure to Detect True Glaucoma

Over diagnosis can skew the model's understanding of what constitutes "true" glaucoma. The model might focus on features present in over diagnosed cases that are not typically seen in more accurate, true glaucoma cases. This could result in missed diagnoses of patients who have glaucoma but do not exhibit the features commonly associated with over diagnosed cases.

G. Misleading Performance Metrics

- *Accuracy:*

The presence of over diagnosed cases in the training data set can distort evaluation metrics commonly used to assess model performance. If the dataset contains many over-diagnosed cases, the model may appear highly accurate. It is simply because it correctly identifies false positives.

- *Precision and Sensitivity:*

The precision may be reduced due to the high number of false positives. Similarly, sensitivity might be impaired because the model is trained on over-diagnosed cases.

- *Positive Predictive Value (PPV) and Negative Predictive Value (NPV):*

Over-diagnosis can lower the PPV, meaning the model will incorrectly flag more healthy individuals as having glaucoma. At the same time, the NPV may be inflated.

V. IMPACT ON CLINICAL DECISION MAKING

If an AI model trained on over diagnosed cases is deployed in a clinical setting, it could lead to unnecessary treatments, such as IOP-lowering medications, surgeries, or laser treatments, for patients who do not need them. Over diagnosis may also result in unnecessary follow-up visits, patient anxiety, additional costs, and potential risks associated with incorrect treatment. Clinicians may lose confidence in the AI system's predictions if they notice frequent over diagnoses. This could undermine the role of AI in clinical decision-making, particularly in fields like glaucoma; where over diagnosis can have serious consequences.

VI. ETHICAL & LEGAL IMPLICATIONS

The negative impact of inconsistent diagnoses on prediction models is fundamentally an ethical issue. One might argue that if the algorithm has been validated and is used by competent providers, all should be well. However, this is not entirely true. If the original diagnoses were inconsistent and biased, what can an algorithm learn from such data? Whether human or machine, must produce unbiased, discrimination-free assessments for reliable, equitable, and resilient health care delivery.

Ethically, algorithms should complement, not replace, consistent human assessments that enhance human welfare. Over-reliance on biased and inconsistent data can lead to further complications. Information inconsistency and bias can erode trust between doctors and patients, especially if patients or other professionals believe the assessments are biased, ultimately polluting decisions.

Machine learning has led to the development of pattern recognition models that can unintentionally propagate and amplify bias and discrimination. Allowing this to continue is unacceptable. A deeper understanding of the strengths and limitations of AI technologies in socio-politically charged environments is crucial to addressing these challenges.

VII. CLINICAL IMPLEMENTATION STRATEGIES

Although AI-based glaucoma prediction models face significant challenges in being deployed for initial diagnosis due to the lack of consensus in glaucoma diagnostic criteria, they still hold potential in other areas of glaucoma care. These models can be effectively used in monitoring disease progression, as they can process vast amounts of imaging and clinical data to track subtle changes over time, providing early indicators of worsening conditions. Additionally, AI models can be valuable in screening, particularly for large populations, where they can help identify individuals at risk or those needing further evaluation by specialists. By automating routine tasks such as image analysis, AI can

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14848301

reduce clinician workload and improve efficiency, helping prioritize patients who require immediate attention. In these contexts, AI can complement human expertise and enhance glaucoma management, even without a fully standardized diagnostic approach.

VIII. CONCLUSION

While AI-based glaucoma prediction models hold significant promise for advancing early diagnosis and personalized treatment, we are not yet ready for deployment in real world setting. The main barrier lies in the lack of consensus in glaucoma diagnosis, which remains highly variable across practitioners, regions, and even diagnostic tools. Different interpretations of diagnostic data, such as imaging and intraocular pressure readings, can lead to inconsistencies that AI models cannot account for without standardized criteria. Until there is agreement on a unified, evidence-based approach to glaucoma diagnosis, AI systems risk perpetuating or even amplifying diagnostic errors. Furthermore, comprehensive validation across diverse populations and clinical settings is essential to ensure that these models are not only accurate but also equitable and generalizable. Therefore, a collaborative effort toward standardizing diagnostic protocols and fostering broader consensus is crucial before AI can be effectively and safely integrated into glaucoma care.

REFERENCES

[1]. Kroese M, Burton H. Primary open angle glaucoma. The need for a consensus case definition *Journal of Epidemiology & Community Health* 2003; **57:752**-754.
[2]. Zhang, L., Tang, L., Xia, M., & Cao, G. (2023). The application of artificial intelligence in glaucoma diagnosis and prediction. *Frontiers in Cell and Developmental Biology*, *11*, 1173094.
[3]. Lee, S. S.-Y., & Mackey, D. A. (2022). Glaucoma – Risk factors and current challenges in the diagnosis of a leading cause of visual impairment. *Maturitas*, *163*, 15–22.
[4]. Križaj D. What is glaucoma? 2019 May 30. In: Kolb H, Fernandez E, Jones B, et al., editors. Webvision: The Organization of the Retina and Visual System [Internet]. Salt Lake City (UT): University of Utah Health Sciences Center; 1995.
[5]. Wishart, P. K. (2009). Interpretation of the glaucoma "landmark studies." *British Journal of Ophthalmology*, *93*(5), 561–562.
[6]. Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, *19*, 221–248.
[7]. Ghaffar, F. (2024). Diagnostic decision-making variability in glaucoma: Comparative analysis of novice and expert optometrists to inform AI system design. *Borealis*, *1*.
[8]. Ahmed, S., Khan, Z., Si, F., Mao, A., Pan, I., Yazdi, F., Tsertsvadze, A., Hutnik, C., Moher, D., Tingey, D., Trope, G. E., Damji, K. F., Tarride, J. E., Goeree, R., & Hodge, W. (2016). Summary of glaucoma diagnostic testing accuracy: An evidence-based meta-analysis. *Journal of Clinical Medicine Research*, *8*(9), 641–649.
[9]. Budenz, D. L., Anderson, D. R., Varma, R., Schuman, J., Cantor, L., Savell, J., Greenfield, D. S., Patella, V. M., Quigley, H. A., & Tielsch, J. (2007). Determinants of normal retinal nerve fiber layer thickness measured by Stratus OCT. *Ophthalmology*, *114*(6), 1046–1052.
[10]. Oshitari, T., Hanawa, K. & Adachi-Usami, E. Changes of macular and RNFL thicknesses measured by Stratus OCT in patients with early-stage diabetes. *Eye* **23**, 884–889 (2009).
[11]. Whang, S. E., & Lee, J.-G. (2020). Data collection and quality challenges for deep learning. *Proceedings of the VLDB Endowment*, *13*(12), 3429–3432.
[12]. Montesinos López, O.A., Montesinos López, A., Crossa, J. (2022). Overfitting, Model Tuning, and Evaluation of Prediction Performance. In: Multivariate Statistical Machine Learning Methods for Genomic Prediction. Springer, Cham.
[13]. Jakubovitz, D., Giryes, R., Rodrigues, M.R.D. (2019). Generalization Error in Deep Learning. In: Boche, H., Caire, G., Calderbank, R., Kutyniok, G., Mathar, R., Petersen, P. (eds) Compressed Sensing and Its Applications. Applied and Numerical Harmonic Analysis. Birkhäuser, Cham.
[14]. Johnson, J.M., Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J Big Data* **6**, 27 (2019). https://doi.org/10.1186/s40537-019-0192-5