

An Agreement-Driven Hybrid Machine Learning Framework for Glaucoma Classification Using Retinal Biomarkers

Nibedita Kalita*, Samir Kumar Borgohain

National Institute of Technology Silchar,
Department of Computer Science and Engineering,
India

{nibedita21_rs, samir}@cse.nits.ac.in

Abstract. The growing burden of glaucoma and its irreversible nature have increased the need for trustworthy computational models that can facilitate early automated diagnosis. This work proposes an Agreement-Driven Hybrid Ensemble (ADHE) framework for glaucoma classification using pre-established biomarker features. The proposed work integrates heterogeneous machine learning classifiers using a confidence-aware agreement mechanism combined with validation-based threshold calibration. Irrespective of conventional ensemble models that optimize a single accuracy-driven decision boundary, ADHE explicitly accounts for inter-model agreement and uncertainty to reduce clinically critical false negative predictions. Results obtained from experimental evaluation demonstrate that the proposed ADHE model has an Area Under Curve (AUC) of 0.868, an F1-score of 0.788 and an accuracy of 78.32%, while consistently reducing false negatives compared to individual classifiers and ungated fusion variants. These results demonstrate that agreement-driven hybrid learning offers a robust, interpretable, and medically aligned solution for automatic glaucoma diagnosis in data-driven ophthalmic decision support systems.

Keywords. Glaucoma classification, hybrid ensemble learning, agreement-driven decision fusion, biomarkers, machine learning.

1 Introduction

Glaucoma, which is a progressive and chronic optic neuropathy is often characterized by irreversible optic nerve damage, mostly leading to permanent vision impairment if not detected at an early stage [19]. Glaucoma comprises

a severe global healthcare concern, with the impacted number of individuals projected to reach approximately 111.8 million by the year 2040 [1]. Recent data of eye patients showing increase numbers of glaucoma cases worldwide and its irreversible nature have intensified the demand for reliable computational models capable of supporting early automated diagnosis.

Initial phases of glaucoma shows no prior symptoms because of that many patients are unaware of the disease till major vision loss has already occurred. This silent progression motivates to detect even smallest development in any biomarker through computer-aided diagnostic (CAD) systems, which can support clinicians by offering objective, reliable, and scalable tools for early screening and detection.

Recent efforts for improving glaucoma diagnostics have turned to machine learning (ML) for extracting quantitative biomarkers from ocular imaging and clinical data [4, 17]. Architectures like neural networks and Support Vector Machine (SVM) have shown good performance in complex patterns extraction but they still face critical problems to be reliable in the field.

A model's performance often drops when faced with the noise or class overlaps of real-world data. To fix this generalization gap, the research community has moved toward ensemble learning, which aggregates multiple models to minimize individual biases. Common ensemble techniques like majority voting [8] assume a uniform level of accuracy that simply doesn't exist.

Data dependency is a major problem because of the patient samples. Model's consistency vary drastically between different patient samples as it can be clearly seen in confidence of these classifiers.

There are recent works for hybrid decision-fusion frameworks that combine confidence-aware mechanisms with adaptive weighting. These methods make an effort to selectively resolve confusing circumstances while incorporating model agreement. Current strategies often struggle with a lack of flexibility, oscillating between deep learning models and overly simplistic fusion tactics [18, 21]. Such methodologies struggle in real-world clinical settings. For practitioners with limited hardware, the high overhead and lack of clear interpretability in these models remain significant barriers to adoption.

We introduce the Agreement-Driven Hybrid Ensemble (ADHE), a framework designed for glaucoma classification that utilizes the biomarker feature set previously established in [11]. At its heart, this system moves beyond simple model integration; it actively weighs how much individual classifiers concur and the specific level of certainty behind their outputs. Instead of treating every model as an equal vote, ADHE prioritizes results backed by high-confidence consensus. For the more "gray-area" or ambiguous samples, the framework switches to a nuanced, weighted fusion logic. This focus on "agreement awareness" bolsters the stability and trust of the final diagnosis without the massive computational tax typically associated with deep learning. To validate the approach, we tested ADHE against a suite of traditional baselines including LR, SVM, RF, GB, and MLP. The data indicates that our model not only edges out these competitors in accuracy and F1-score but also holds its own in AUC-ROC performance. Ultimately, these results suggest that an agreement-driven logic is a highly viable, efficient path for automating glaucoma screening via clinical features.

The primary contributions of this work can be summed up as follows:

- An agreement-driven hybrid ensemble (ADHE) framework is proposed for

biomarker-based glaucoma classification, in which inter-model consensus is explicitly assessed to regulate predictions prior to fusion, surpassing classic voting and standard stacking methodologies.

- To adaptively combine heterogeneous classifier outputs, a confidence-aware decision fusion mechanism is developed. This reduces the influence of conflicting or uncertain predictions and improves robustness in borderline cases.
- Filtered probabilistic outputs are aggregated using a learned stacking strategy with validation-based threshold calibration, which then data-drivenly refines the final decision boundary.
- Extensive experimental evaluation is conducted between the proposed ADHE model and existing baseline classifiers, while emphasizing computational efficiency and maintaining interpretability suitable for practical deployment.

The remaining section of this paper is organised as follows. Section 2 deals with the literature review of the related works. Section 3 presents the proposed agreement-driven hybrid ensemble (ADHE) methodology including the dataset, preprocessing, experimental protocol, baseline models and ablation procedures. Section 4 reports results and discussion, featuring quantitative comparisons, visual analyses and ablation findings. Finally, Section 5 contains the outline of the paper, limitations and future directions.

2 Related Work

Glaucoma diagnosis automation from retinal fundus imaging has been an active area of research for more than two decades, driven by the clinical need for early diagnosis and scalable screening solutions. Broadly, existing approaches can be categorized into handcrafted feature-based machine learning methods, end-to-end deep learning models, and ensemble-based frameworks

that attempt to combine the strengths of multiple classifiers.

Early studies predominantly relied on handcrafted structural and textural descriptors derived from the optic nerve head (ONH) region. Clinically motivated features such as the cup-to-disc ratio (CDR), neuroretinal rim thickness, rim area asymmetry, parapapillary atrophy, and vessel distribution patterns were extensively explored due to their strong association with glaucomatous damage. Traditional machine learning algorithms such as Logistic Regression, Support Vector Machines (SVM), Random Forests, Decision Trees, and k-Nearest Neighbors were typically used for classification of these features.

For example, Kalita et al. [10] presented a shape and texture feature combination technique derived from the neuroretinal rim area and showed that carefully engineered biomarkers can achieve competitive diagnostic accuracy under controlled settings. Similarly, Kumar et al. [14] focused on parapapillary atrophy and optic disc characteristics combined with SVM classifiers for early-stage glaucoma detection. Verma et al. [20] also evaluated traditional machine learning classifiers for glaucoma detection on the Drishti-GS1 dataset, reporting moderate performance with SVM-based classification, thereby establishing a baseline for feature-driven approaches. While such feature-driven methods offer interpretability, computational efficiency, and alignment with clinical reasoning, their performance is often constrained by sensitivity to segmentation accuracy, imaging artifacts, and limited generalization across datasets acquired using different imaging protocols.

With the advent of large scale annotated retinal datasets, deep learning particularly convolutional neural networks (CNNs) has become the dominant paradigm for glaucoma classification. CNN-based models extract hierarchical features directly from fundus images and have demonstrated superior performance compared to traditional handcrafted approaches. Architectures such as VGG, ResNet, DenseNet, and Xception have been widely adopted, often leveraging transfer learning from large natural image datasets. Diaz-Pinto et al. [6] conducted one of the most extensive

validations of CNNs for glaucoma assessment, demonstrating that deep models can effectively capture subtle structural patterns associated with disease progression.

Subsequent works have incorporated attention mechanisms, multi-scale feature extraction, and region-of-interest guided learning to further enhance discriminative capability. Despite their strong performance, deep learning models however generally require large annotated datasets, substantial computational resources, and specialized hardware. Moreover, their black-box nature and limited explainability pose challenges for clinical acceptance, particularly in screening scenarios where transparency and trust are critical.

Ensemble learning algorithms have gained increasing attention in recent years in order to address robustness and generalization issues. These approaches aim to combine the outputs of multiple classifiers to exploit complementary decision patterns and reduce variance. Arsyian and Al Maki [2] proposed a feature-driven glaucoma classification method based on invariant moment descriptors, comparing k-NN and Random Forest classifiers and reporting improved performance with tree-based ensembles. Conventional ensemble techniques include majority voting, soft probability averaging, bagging, boosting, and stacking. Deep ensemble variants, such as snapshot ensembles and multi-branch CNNs, have also been explored for glaucoma detection.

For instance, Geetha et al. [7] proposed a snapshot ensemble CNN framework based on EfficientNet architectures, reporting improved performance through diversified model snapshots. Comprehensive surveys by Mithunavarshini et al. [15] highlight that ensemble-based methods consistently outperform single-model approaches across multiple glaucoma datasets. Nevertheless, most existing ensemble frameworks rely on fixed fusion rules and implicitly assume equal reliability across constituent models. As a result, uncertain or conflicting predictions may exert undue influence on the final decision, particularly in borderline or ambiguous cases.

Parallel to the rise of deep learning, there has been renewed interest in biomarker-driven

representations that leverage accurate optic disc (OD) and optic cup (OC) segmentation to derive clinically meaningful features.

Biomarker-based methods strike a balance between interpretability and performance by using structured, low-dimensional features that are directly linked to disease pathology. Such approaches reduce dependence on massive image-level training data and are well suited for small or moderately sized datasets. However, most biomarker-centric studies still employ single classifiers or conventional ensemble strategies, without incorporating adaptive decision-level reasoning, inter-model agreement analysis, or confidence-aware fusion mechanisms.

In comparison to the existing literature, the proposed Agreement-Driven Hybrid Ensemble (ADHE) framework integrates biomarker-based feature representations with an agreement-aware and confidence-adaptive fusion strategy. Rather than relying on static voting or averaging schemes, the proposed method explicitly models inter-classifier agreement, selectively suppresses uncertain predictions through gating mechanisms, and employs learned stacking with validation-driven threshold calibration. This architectural design facilitates precise control over the decision boundary, allowing sensitivity and specificity to be balanced with accordance to the clinical screening requirements. By combining interpretability, robustness, and computational efficiency, the ADHE framework addresses key limitations of prior machine learning, deep learning, and ensemble-based approaches, while delivering competitive performance on the SMDG-19 biomarker dataset.

3 Proposed Methodology

3.1 Dataset Description and Preprocessing

In this work, the input feature representation is based on a previously established novel biomarker [11] derived from the optic nerve head (ONH) region of retinal fundus images. Biomarker extraction was carried out using the SMDG-19 dataset [13], a publicly accessible retinal fundus repository that integrates images from 19 widely

used datasets, including CRFO-v4, DRISHTI-GS1, EyePACS, and REFUGE1, comprising a total of 12,449 fundus images. However, expert-annotated optic cup and optic disc segmentation ground truths were available only for 2,873 images (2,143 glaucomatous and 730 healthy cases). Consequently, this subset of images, provided in PNG format at a spatial resolution of 512×512 pixels, was utilized for feature extraction and biomarker computation.

In order to avoid class imbalance and data leakage problems, the dataset was partitioned using stratified sampling into training, validation, and test subsets. Specifically, 60% of the samples were used for training, 20% for validation, and 20% for independent testing. The validation set was then exclusively utilized for tuning ensemble thresholds and agreement parameters.

3.2 Baseline Classifiers

To provide a fair and representative benchmark, five well-established machine learning classifiers were employed as baseline models. These methods were selected to cover linear, nonlinear, ensemble, and neural network-based learning paradigms:

- **Logistic Regression (LR):** Logistic Regression is a linear probabilistic classifier that models the posterior class probability using a sigmoid function [9]. Owing to its simplicity and interpretability, LR is frequently adopted as a baseline in biomedical classification problems and serves as a reference for assessing the linear separability of the biomarker features.
- **Support Vector Machine with Radial Basis Function (SVM-RBF):** Support Vector Machines gives an optimal separating hyperplane by maximizing the margin between classes [5]. The modeling of nonlinear decision boundary is enabled by using an RBF kernel, thus making SVM-RBF a popular choice for glaucoma detection from fundus-derived features [14].

- **Random Forest (RF):** Random Forest is a bagging-based ensemble learning technique. It combines predictions from several decision trees constructed on randomised subsets of samples and features [3]. RF has demonstrated strong performance in retinal disease classification because of the robustness to noise and ability to extract complex feature interactions [2].
- **Histogram-based Gradient Boosting (HGB):** An efficient variant of gradient boosting is Histogram-based Gradient Boosting. HGB discretizes continuous features into histograms to reduce computational cost while preserving predictive accuracy [12]. This method is particularly effective for structured and tabular biomedical data, offering a favorable trade-off between performance and efficiency.
- **Multilayer Perceptron (MLP):** The Multilayer Perceptron is a feed-forward neural network consisting of multiple fully connected layers trained using backpropagation [16]. Although less complex than deep convolutional networks, MLPs can model nonlinear relationships in biomarker feature spaces and have been employed in several glaucoma screening studies.

All baseline classifiers were trained using identical training, validation, and test partitions and evaluated under the same experimental protocol. This controlled setup ensures an unbiased comparison and allows the performance gains of the proposed Agreement-Driven Hybrid Ensemble (ADHE) framework to be attributed solely to its architectural and decision-fusion mechanisms rather than differences in data handling or evaluation conditions.

3.3 Proposed Model: Agreement-Driven Hybrid Ensemble (ADHE) Framework

The proposed Agreement-Guided Hybrid Ensemble (ADHE) framework is designed as a modular and interpretable machine learning pipeline for glaucoma classification using tabular

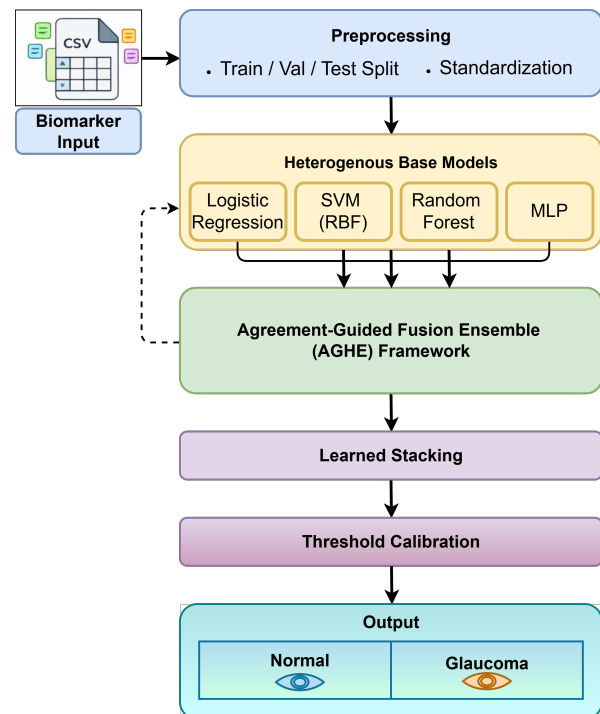


Fig. 1. Architectural representation of the proposed ADHE framework

biomarker features. The overall architecture, illustrated in Fig. 1, consists of six main stages: preprocessing, heterogeneous base learning, probability extraction, agreement-guided fusion, learned stacking, and threshold-calibrated decision making.

Stage 1: Preprocessing. The proposed framework operates on a pre-established set of biomarker which is extracted from the ONH region of the retinal image. The input data is initially partitioned into training, validation, and testing subsets for avoiding information leakage during model construction. Feature standardization is applied using statistics computed from the training data to ensure scale consistency across all classifiers.

Stage 2: Heterogeneous Base Learning. Several classical machine learning models are trained independently to extract complementary decision characteristics. These include LR, SVM-RBF, RF, and a Multi-Layer Perceptron

Algorithm 1 Agreement-Driven Hybrid Ensemble (ADHE)

Require: Biomarker feature matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$,
class labels $\mathbf{y} \in \{0, 1\}^N$

Ensure: Final class prediction $\hat{y} \in \{0, 1\}$

- 1: Split $\mathbf{X}, \mathbf{y}$ into training, validation, and test sets
- 2: Standardize features using training statistics
- 3: Initialize base classifiers: $\mathcal{M} = \{\text{LR}, \text{SVM}_{\text{RBF}}, \text{RF}, \text{HGB}, \text{MLP}\}$
- 4: **for** each model $m \in \mathcal{M}$ **do**
- 5: Train m on the training set
- 6: Obtain posterior probabilities $\mathbf{p}_m$ on validation and test sets
- 7: **end for**
- 8: Form stacked probability matrix $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{|\mathcal{M}|}]$
- 9: Train a meta-classifier $\mathcal{M}_{\text{meta}}$ on validation probabilities to learn agreement-aware fusion weights
- 10: Obtain fused probability $\hat{\mathbf{p}} = \mathcal{M}_{\text{meta}}(\mathbf{P})$
- 11: Tune decision threshold τ on validation set by maximizing balanced accuracy (sensitivity–specificity trade-off)
- 12: Assign final prediction:

$$\hat{y} = \begin{cases} 1, & \text{if } \hat{p} \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

return $\hat{y}$

(MLP). Each model represents diverse inductive biases spanning linear, kernel-based, tree-based, and neural paradigms.

Stage 3: Probability Estimation. Each base classifier generates a posterior probability indicating the likelihood of glaucoma class after completion of training phase.

Obtained probability outputs reflect the confidence of individual models, provides a uniform representation which can be directly compared and combined across classifiers. Inputs are further provided for fusion and stacking stages.

Stage 4: Agreement-Driven Fusion. An agreement mechanism is introduced to assess inter-model consensus using expected probability dispersion. When predicted probabilities are

almost same, the decision is considered reliable whereas larger discrepancies signal uncertainty or borderline cases.

This agreement information directs the fusion strategy by prioritizing consistent predictions over unreliable ones.

Stage 5: Learned Stacking. The agreement-filtered probability outputs are combined using a learned stacking strategy rather than relying on fixed voting criteria.

A lightweight meta-classifier is trained on the validation set to properly weight the base model predictions, enabling adaptive and data-driven fusion while preserving interpretability.

Stage 6: Threshold-Calibrated Decision Making. The final decision boundary is established by applying a validation-driven, post-hoc threshold to the probabilistic output of the stacked model. This calibration is tuned on the validation fold to maximise balanced performance (e.g., accuracy and AUC) and produces the final binary labels in a clinically meaningful manner.

Overall, the ADHE framework integrates confidence awareness, inter-model agreement, and learned fusion within a unified and computationally efficient architecture.

The complete procedural workflow is summarized in Algorithm 1 providing a robust yet interpretable solution for biomarker based glaucoma classification.

4 Results and Discussion

This section evaluates the comprehensive quantitative analysis of the proposed glaucoma classification framework in comparison with five widely used baseline classifiers: LR, SVM-RBF, RF, HistGradientBoosting, and MLP.

Performance assessment is done using accuracy, AUC, and F1-score, along with confusion matrices and ROC characteristics to provide class-wise and threshold-independent insights.

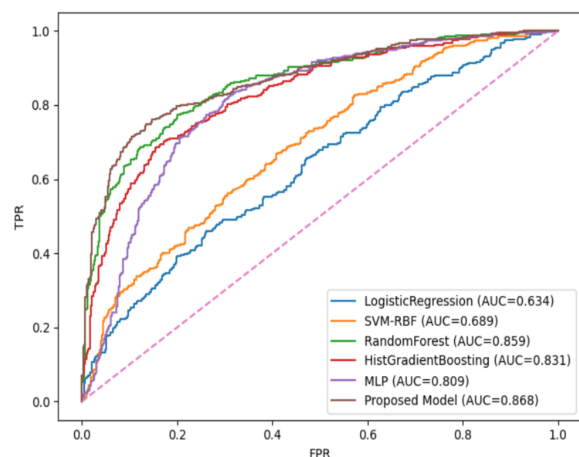


Fig. 2. ROC curves of baseline classifiers and proposed model, highlighting the superior AUC of the proposed model

4.1 Experimental Setup

All experiments were implemented in Python (v3.11.5) using Jupyter Notebook and the scikit-learn library on a Fujitsu workstation equipped with an Intel Xeon W-2133 CPU, 64 GB RAM, 2 TB HDD, NVIDIA Quadro RTX 4000 GPU, and Windows 11 Pro. Despite the available GPU, all experiments were conducted in a CPU-based setting to ensure reproducibility and practical clinical applicability.

Table 1. Comparative analysis of the proposed ADHE model against baseline classifiers

Model	Accuracy	AUC	F1-score
Random Forest	0.7774	0.8589	0.7787
HistGradientBoosting	0.7517	0.8312	0.7526
MLP	0.7506	0.8093	0.7476
SVM-RBF	0.6224	0.6890	0.6058
Logistic Regression	0.5944	0.6338	0.5433
Proposed Model	0.7832	0.8679	0.7877

Model performance was evaluated on an independent test set using evaluation matrices namely accuracy, F1-Score and AUC, complemented by ROC curves, confusion matrices and radar plots for comprehensive performance analysis.

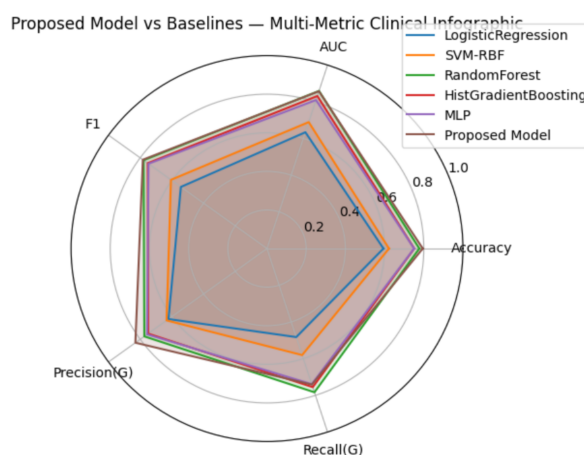


Fig. 3. Multi-metric radar plot for comparison of proposed model with baselines

4.2 Evaluation Metrics and Visualization

Several complementary metrics, such as accuracy, F1-score, and area under the receiver operating characteristic curve (AUC-ROC), were used to evaluate the proposed model's performance. Confusion matrices and precision-recall matrices were computed to analyze class-wise performance. The ROC curves were also plotted to see the discriminative capability. Also, a radar-based infographic visualisation was plotted to show the relative advantages of the suggested framework over baseline classifiers and to provide an easy-to-understand multi-metric comparison.

4.3 Performance Comparison of the Proposed ADHE Model

A quantitative comparison of the proposed framework against several baseline classifiers is provided in Table 1. The proposed model achieves the highest overall performance, having an accuracy of 78.32%, an AUC of 0.8679, and an F1-score of 0.7877, indicating superior discriminative capability and balanced classification behavior. These results indicate that the model not only separates glaucomatous and normal samples more effectively, but also maintains a balanced trade-off between precision and recall. Random Forest performs competitively

(Accuracy = 0.7774, AUC = 0.8589), followed by HistGradientBoosting and MLP among the baseline classifiers.

In contrast, SVM with an RBF kernel and Logistic Regression exhibit noticeably weaker performance, suggesting that their decision boundaries are less suited to the underlying biomarker feature distribution. The ROC curves shown in Fig. 2 further support this observation. The curve analysis reinforces these findings, with the proposed model achieving the highest AUC (0.8679) among all evaluated classifiers, indicating greater separability between glaucomatous and normal samples across different decision thresholds. Though Random Forest and HistGradientBoosting demonstrate almost similar ROC trends, their curves are consistently below that of the proposed method.

A more comprehensive comparison is provided in Fig. 3 which depicts the multi-metric radar plot advantage by showing that the proposed model consistently dominates across accuracy, AUC, precision, recall, and F1-score, highlighting its robustness and balanced behavior from a clinical screening perspective. This is particularly important for glaucoma screening, where uneven trade-offs between sensitivity and specificity can lead to clinically adverse outcomes.

Confusion-matrix analysis provides detailed class-wise insight into the behavior of all evaluated models (Figs. 4–9). As depicted in Fig. 9, the proposed ADHE model achieves the highest numbers of true positives and true negatives while substantially reducing false negatives.

4.4 Proposed Model Performance Comparison with State-of-the-Art (SOTA) Models

Table 2 represents a comparative overview of the proposed Agreement-Driven Hybrid Ensemble (ADHE) model against SOTA glaucoma classification approaches reported in the literature. Existing methods predominantly rely on fundus image-based learning using either classical machine learning models, such as SVM and Random Forest, or deep convolutional architectures including Xception and ResNet-50, evaluated on benchmark datasets

such as Drishti-GS1, REFUGE, ACRIMA, and sjchoi86-HRF. These approaches typically achieve classification accuracies in the range of 70%–75.86%, with limited reporting of complementary metrics like AUC and F1-score.

In contrast, the proposed ADHE framework, evaluated on the SMDG-19 dataset using biomarker-level features, achieves a higher accuracy of 78.32%, along with an F1-score of 0.7877 and an AUC of 0.8679. Notably, while many existing methods emphasize image-level feature learning through deep architectures, the proposed approach demonstrates that agreement-aware ensemble learning with confidence-guided fusion on clinically interpretable biomarkers can achieve competitive or superior performance without reliance on computationally intensive deep models. This highlights the practicality and efficacy of the suggested framework for scalable and interpretable glaucoma screening applications.

4.5 Ablation Study

An ablation analysis examined the contribution of each key component of the proposed Agreement Driven Hybrid Ensemble framework. The ablation results were shown in Table 3. Also, Fig. 10 provided a visual illustration of these results. Each design element contributed to the final performance of the proposed framework. The number of false negative cases represented the undetected glaucoma cases and changes in these cases showed the clinical importance of the agreement and the fusion mechanism. The removal of individual components or the simplification of the individual components caused performance degradation. These findings confirmed the relevance of each component for reliable glaucoma screening.

A consistent pattern emerges across the ablation experiments with an observed monotonic trend. Eliminating key components leads to a steady rise in false negatives, accompanied by a decline in overall discriminative performance. The *HGB Only* variant performs the poorest, with the smallest AUC of 0.8312, the lowest accuracy of 75.17%, along with the highest number of missed glaucoma cases i.e., FN = 105. This clearly shows the

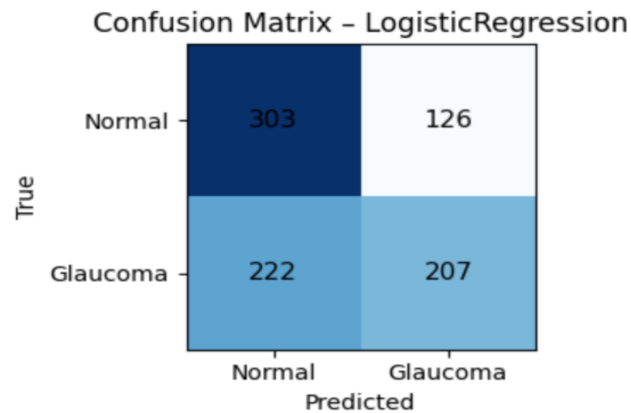


Fig. 4. Confusion matrix of Logistic Regression

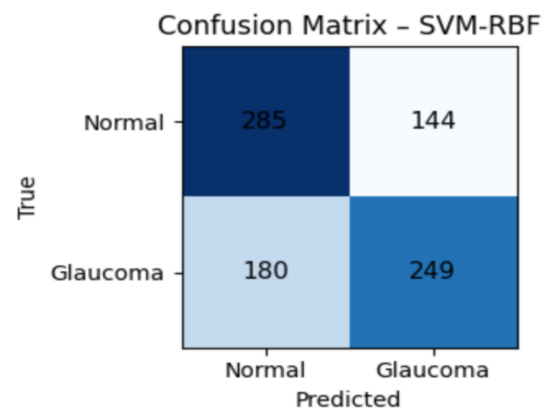


Fig. 5. Confusion matrix of SVM-RBF

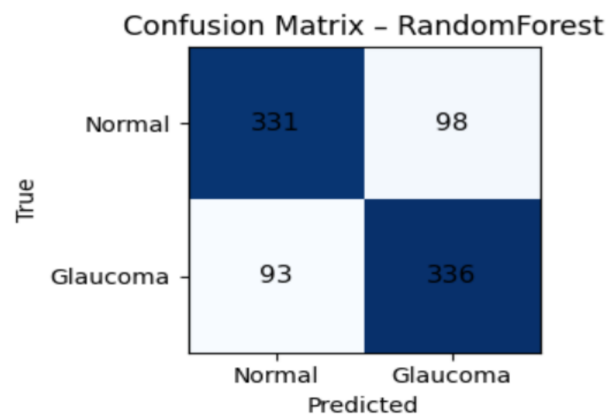


Fig. 6. Confusion matrix of Random Forest

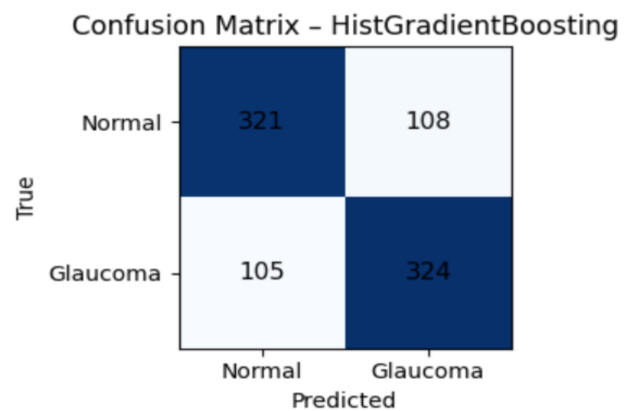


Fig. 7. Confusion matrix of Hist Gradient Boosting

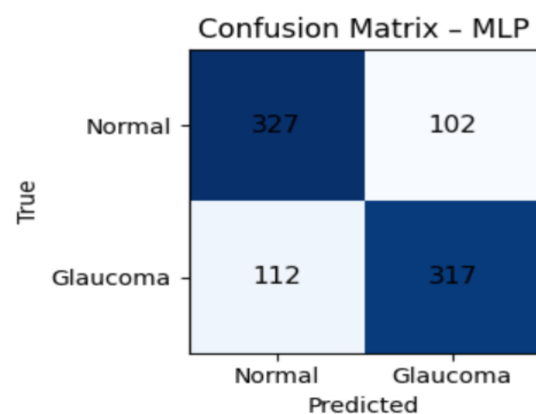


Fig. 8. Confusion matrix of MLP

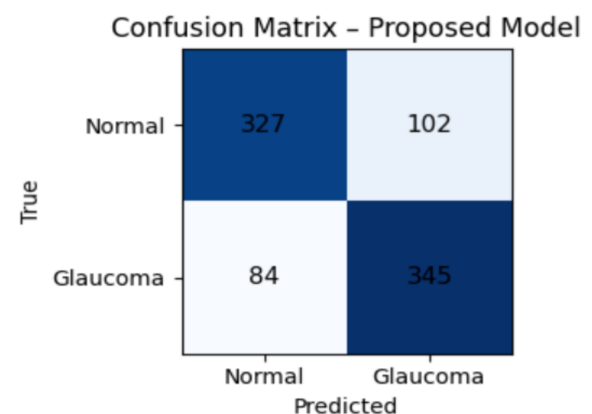


Fig. 9. Confusion matrix of Proposed ADHE Model

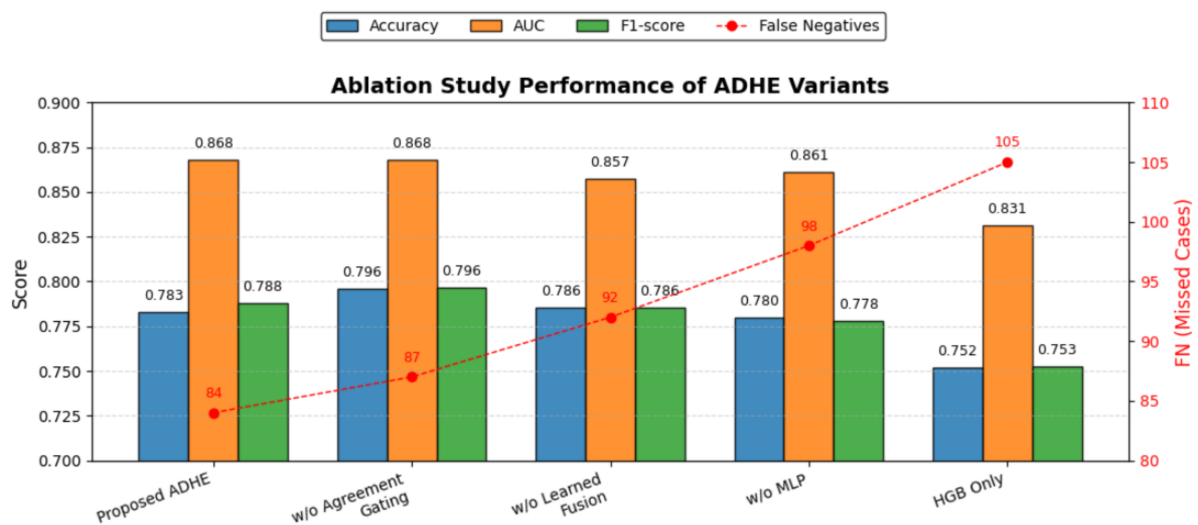
Table 2. Indicative proposed framework comparison with SOTA glaucoma classification methods

Method	Dataset used	Input Modality	Accuracy (%)	AUC	F1-score
SVM [20]	Drishti-GS1	Fundus Images	70.77%	–	–
Random Forest [2]	REFUGE	Fundus Images	75.86%	–	0.7142
Xception [6]	ACRIMA	Fundus Images	70.21%	0.7678	–
ResNet-50 [6]	sjchoi86-HRF	Fundus Images	70%	0.85	–
Proposed ADHE Framework	SMDG-19	Biomarker Features	78.32%	0.8679	0.7877

Table 3. Ablation study evaluating the contribution of individual variants of the proposed ADHE framework

Variant	Acc*(%)	AUC	F1-score	FN
HGB Only	75.17	0.8312	0.7526	105
w/o MLP	77.97	0.8613	0.7779	98
w/o Learned Fusion	78.55	0.8572	0.7855	92
w/o Agreement Gating	79.60	0.8679	0.7963	87
Proposed ADHE	78.32	0.8679	0.7877	84

*Note: Acc denotes classification accuracy, AUC is area under ROC curve, F1-score represents the harmonic mean of precision and recall, while FN indicates the number of false negative cases (missed glaucoma samples)

**Fig. 10.** Performance comparison of ADHE variants together with false negatives (red dashed line, right axis)

limitation of relying on a single classifier without incorporating complementary representations or consensus-based decision logic. When the MLP head or the learned fusion strategy is excluded, performance also degrades, reflected by higher FN counts of 98 and 92 and corresponding drops in AUC and F1-score.

These results underscore the value of nonlinear feature transformation and adaptive fusion for capturing the complex relationships among heterogeneous biomarker features. While the variant *w/o Agreement Gating* achieves marginally higher accuracy and F1-score, this improvement comes at a cost i.e., it misses a substantially larger

number of glaucoma cases (FN = 87) compared to the complete framework.

This shows that the marginal improvements in aggregate metrics do not necessarily correspond to safer clinical outcomes. Contrary to the fact, the full ADHE configuration yields the lowest false-negative count (FN=84) cases while preserving competitive accuracy and the strongest AUC, suggesting a better balance between sensitivity and specificity.

When analyzed through a clinical prospect, reducing false negative cases is especially important in glaucoma screening, since undetected cases may progress silently and result in vision loss which is irreversible. Taken together, the results summarized in Table 3 and Fig. 10 show that agreement-aware gating and learned hybrid fusion play a central role in improving diagnostic reliability, making the proposed ADHE framework better aligned with the safety requirements of real-world glaucoma screening systems.

5 Conclusion

The work presented here is a pre-established biomarker input based Agreement-Driven Hybrid Ensemble (ADHE) framework for the glaucoma classification task. ADHE combines outputs from various machine learning classifiers by taking into account both inter-model agreement and prediction confidence. Instead of relying on simple voting or naive averaging, the proposed framework favors predictions supported by multiple models and uses a weighted fusion rule to resolve uncertain cases.

This selective fusion improves decision stability and robustness while keeping computational overhead low, since it builds on classical learners and a lightweight meta-classifier. Experimental results showed that ADHE outperformed several standard baselines, yielding higher accuracy and F1 while retaining competitive AUC values.

These findings suggest that agreement-guided fusion is well suited to tabular biomedical problems, where individual model behavior can change across different feature regimes, and that ADHE's model-agnostic design makes it practically deployable for resource-constrained settings.

Despite the proposed model's effectiveness, there are some limitations. This work is limited to a single dataset and binary classification setting.

Extensive clinical validation on larger, multi-center cohorts is needed to confirm generalizability.

Future work focuses on extending the approach to multi-class glaucoma staging by incorporating explainability tools to increase clinical transparency and trust.

References

1. **Allison, K., Patel, D., Alabi, O. (2020).** Epidemiology of glaucoma: the past, present, and predictions for the future. *Cureus*, Vol. 12, No. 11.
2. **Arsyan, A. R., Al Maki, W. F. (2022).** Classification of glaucoma using invariant moment methods on k-nearest neighbor and random forest models. *Building of Informatics, Technology and Science (BITS)*, Vol. 3, No. 4, pp. 466–472.
3. **Breiman, L. (2001).** Random forests. *Machine Learning*, Vol. 45, No. 1, pp. 5–32.
4. **Civit-Masot, J., Domínguez-Morales, M. J., Vicente-Díaz, S., Civit, A. (2020).** Dual machine-learning system to aid glaucoma diagnosis using disc and cup feature extraction. *IEEE Access*, Vol. 8, pp. 127519–127529.
5. **Cortes, C., Vapnik, V. (1995).** Support-vector networks. *Machine Learning*, Vol. 20, No. 3, pp. 273–297.
6. **Díaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J. M., Navea, A. (2019).** Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical engineering online*, Vol. 18, No. 1, pp. 29.
7. **Geetha, A., Sobia, M. C., Santhi, D., Ahilan, A. (2025).** Deep gd: Deep learning based snapshot ensemble cnn with efficientnet for glaucoma detection. *Biomedical Signal Processing and Control*, Vol. 100, pp. 106989.

8. **Guntreddi, V., Sivakumar, V. (2025).** Deep learning based glaucoma detection using majority voting ensemble of resnet50, vgg16, and swin transformer. *Results in Engineering*, pp. 107229.
9. **Hosmer, D. W., Lemeshow, S., Sturdivant, R. X. (2013).** *Applied Logistic Regression*. Wiley.
10. **Kalita, N., Borgohain, S. K. (2024).** Glaucoma diagnosis based on fusion of shape and texture features from neuroretinal rim area of retinal fundus image. *International Conference on Machine Intelligence, Tools, and Applications*, Springer, pp. 201–208.
11. **Kalita, N., Borgohain, S. K. (2024).** An ocular feature-based novel biomarker determination for glaucoma diagnosis using supervised machine learning and fundus imaging. *IEEE Sensors Letters*.
12. **Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y. (2017).** Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, Vol. 30.
13. **Kiefer, R., Abid, M., Steen, J., Ardali, M. R., Amjadian, E. (2023).** A catalog of public glaucoma datasets for machine learning applications: A detailed description and analysis of public glaucoma datasets available to machine learning engineers tackling glaucoma-related problems using retinal fundus images and oct images. *Proceedings of the 2023 7th International Conference on Information System and Data Mining*, pp. 24–31.
14. **Kumar, S., Kumar, B. (2022).** Automatic early glaucoma detection by extracting parapapillary atrophy and optic disc from fundus image using svm. *Multimedia Tools and Applications*, Vol. 81, No. 10, pp. 13513–13535.
15. **Mithunavarshini, A., Deepa, S. (2025).** A comprehensive survey of deep learning and ensemble techniques in glaucoma detection. *International Journal of Preventive Medicine and Health (IJPMH)*, Vol. 5, No. 2, pp. 9–13.
16. **Rumelhart, D. E. e. a. (1986).** Learning representations by back-propagating errors. *Nature*, Vol. 323, pp. 533–536.
17. **Septiarini, A., Khairina, D. M., Kridalaksana, A. H., Hamdani, H. (2018).** Automatic glaucoma detection method applying a statistical approach to fundus images. *Healthcare informatics research*, Vol. 24, No. 1, pp. 53–60.
18. **Serte, S., Serener, A. (2019).** A generalized deep learning model for glaucoma detection. *2019 3rd International symposium on multidisciplinary studies and innovative technologies (ISMSIT)*, IEEE, pp. 1–5.
19. **Tham, Y.-C., Li, X., Wong, T. Y., Quigley, H. A., Aung, T., Cheng, C.-Y. (2014).** Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*, Vol. 121, No. 11, pp. 2081–2090.
20. **Verma, R., Shrinivasan, L., Hiremath, B. (2023).** Machine learning classifiers for detection of glaucoma. *IAES International Journal of Artificial Intelligence*, Vol. 12, No. 2, pp. 806.
21. **Yi, S., Zhou, L. (2025).** Multi-step framework for glaucoma diagnosis in retinal fundus images using deep learning. *Medical & Biological Engineering & Computing*, Vol. 63, No. 1, pp. 1–13.

Article received on 09/02/2026; accepted on 03/04/2026.

**Corresponding author is Nibedita Kalita.*