RetinaTriage: Explainable Multi-Disease Retinal Screening and Confidence-Aware Triage from Color Fundus Photographs

Bidipta Biswas[1], Anitha D[2]
[1,2]*Department of Computational Intelligence, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India – 603203*

***Abstract*—Retinal illnesses such as diabetic retinopathy, glaucoma, cataract, and age-related macular degeneration are major sources of preventable vision impairment, particularly when early screening is limited by paucity of specialists and variability in imaging conditions. Automated analysis of color fundus photographs offers a scalable approach for supporting early detection; however, practical deployment requires not only accurate classification but also interpretable predictions and reliable confidence estimation. This paper presents RetinaTriage, an explainable and confidence-aware retinal screening framework for multi-class ocular disease recognition from fundus images. The proposed system performs five-class classification of age-related macular degeneration (ARMD), cataract, diabetic retinopathy (DR), glaucoma, and normal retina using a transfer-learning approach based on EfficientNet-B3. To improve clinical transparency, Grad-CAM is integrated to generate visual explanations highlighting image regions influencing model decisions. A confidence-aware inference mechanism using temperature scaling further supports triage-oriented decision making by distinguishing reliable predictions from uncertain outputs. The framework is designed not merely as a benchmark classifier but as a workflow-ready screening support system that combines prediction, interpretability, and uncertainty handling in a unified pipeline.**

***Index Terms*—Fundus Photography, Diabetic Retinopathy, Glaucoma, Age-Related Macular Degeneration, Cataract, Transfer Learning, EfficientNet, Explainable AI, Grad-CAM, Temperature Scaling, Calibration, Triage, ODIR.**

I. INTRODUCTION

Retinal and optic nerve disorders such as diabetic retinopathy (DR), glaucoma, cataract, and age-related macular degeneration (ARMD) are among the leading causes of vision impairment worldwide [1]. A major clinical challenge is that several of these conditions may progress silently during early stages, delaying diagnosis until irreversible damage has already occurred. Early screening therefore plays a critical role in reducing avoidable blindness and improving long-term visual outcomes. In this context, color fundus photography has emerged as a practical and non-invasive imaging modality for large-scale ocular assessment, because it is relatively low-cost, widely available, and suitable for tele-ophthalmology and community-level screening workflows.

Recent advances in deep learning have substantially improved automated analysis of retinal images. Convolutional neural networks have demonstrated strong performance in identifying ophthalmic abnormalities from fundus photographs, in several cases approaching specialist-level agreement. However, most existing systems are optimized primarily for classification accuracy and are presented as benchmark models rather than as deployable screening tools. In practical screening environments, a model output must be accompanied by an indication of how the decision was reached and how much trust should be placed in it, particularly when image quality is variable or when the case lies close to a decision boundary.

This work addresses that gap by presenting RetinaTriage, a retinal screening framework that integrates multi-class disease recognition, visual explanation, and confidence-aware triage within a single pipeline. The system is built on a transfer-learning backbone and is designed so that predictions

are not only accurate but also inspectable and reliably quantified, making outputs more useful for downstream screening decisions rather than merely correct in a benchmarking sense.

Reliable confidence estimation has become particularly important as deep classifiers are increasingly deployed in clinical contexts. Conventional softmax outputs tend to be overconfident, meaning that a model may report high certainty on an incorrect prediction. In a screening workflow, such behaviour can translate into unsafe referral decisions. By contrast, calibrated confidence values make it possible to distinguish outputs that are likely to be correct from those that warrant manual review, and this distinction is central to the design of RetinaTriage. The system is therefore intended to behave less like an opaque classifier and more like a triage assistant that indicates when a case should be escalated, routed routinely, or deferred for re-imaging.

Alongside confidence, interpretability is treated as a first-class requirement rather than an optional addition. Clinicians are understandably reluctant to act on predictions that provide no supporting evidence, and post-hoc heatmaps offer a practical means of exposing the regions that influenced a decision. In retinal imaging this is especially useful, because most of the relevant pathology is localized to identifiable anatomical structures such as the optic disc, the macula, or the vascular network. An explanation that highlights a clinically plausible region is more trustworthy than one that highlights an imaging artifact, and access to that distinction supports human oversight of automated outputs.

The proposed work is developed using the ODIR benchmark dataset [4], which contains fundus photographs collected under real-world conditions with substantial variability in image quality, acquisition devices, and disease distribution. By focusing on a five-class setting, this work aims to provide a balanced and implementable formulation for computer-aided retinal screening while preserving practical disease coverage.

The main contributions of this paper are as follows:

- A transfer-learning-based multi-class retinal disease classification pipeline is developed for automated ocular screening from color fundus images.
- An explainability layer based on Grad-CAM is integrated to improve transparency of model decisions.
- A confidence-aware decision mechanism is incorporated to support safer triage by identifying low-certainty predictions.
- The complete framework is positioned as a clinically oriented screening support system that combines prediction, interpretability, and uncertainty handling in a unified pipeline.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the dataset and problem formulation. Section IV presents the proposed methodology. Section V details the experimental setup. Section VI reports results and analysis. Section VII discusses limitations and future directions, and Section VIII concludes the paper.

II. RELATED WORKS

A. Deep Learning for Fundus-Based Ocular Disease Detection

Deep learning has become a major approach for automated retinal image analysis, particularly for screening tasks based on color fundus photographs. Early landmark studies established that convolutional neural networks can achieve high diagnostic performance for diabetic retinopathy detection when trained on large annotated fundus datasets [2]. These results demonstrated that image-based learning methods could approach specialist-level performance for specific ophthalmic tasks and created the foundation for broader research in computer-aided retinal screening. Subsequent studies extended this direction beyond diabetic retinopathy to related ocular conditions such as glaucoma and ARMD, showing that deep learning models can learn disease-specific structural and textural patterns from retinal photographs.

A key progression in the literature has been the movement from single-disease detection toward multi-disease screening. While many earlier systems focused on binary or narrow diagnostic tasks, more recent work has explored broader classification settings in which a single model identifies multiple ocular conditions from one fundus image [6]. This shift is clinically important because real screening workflows do not encounter diseases in isolation. In practice, a patient image may exhibit signs of more than one

pathology, and referral decisions often depend on distinguishing among several plausible conditions rather than answering a single yes/no question.

B. Transfer Learning in Retinal Image Analysis

Transfer learning has been widely adopted in medical imaging because labeled clinical datasets are often limited compared with large natural-image benchmarks. In fundus-image analysis, pretrained convolutional backbones such as EfficientNet, ResNet, and related architectures have been fine-tuned to retinal data to improve convergence, stability, and classification accuracy [7]. Several studies on ocular disease recognition have shown that transfer learning is especially useful when class imbalance and inter-class similarity are present. At the same time, literature also shows that architecture choice alone is not sufficient; performance remains sensitive to preprocessing strategy, data quality, class distribution, and label formulation.

C. Explainability in Medical Image Classification

Despite strong classification performance, deep learning systems in healthcare face an important trust problem: clinicians are often reluctant to rely on predictions that do not provide interpretable evidence. Among explainability methods, Gradient-weighted Class Activation Mapping (Grad-CAM) has emerged as one of the most widely used approaches for visual explanation in convolutional neural networks [9]. Grad-CAM produces class-specific heatmaps that highlight image regions contributing most strongly to a prediction, thereby allowing users to inspect whether the model is attending to clinically plausible structures. In retinal imaging, explainability is especially valuable because disease interpretation is linked to visible anatomical patterns such as lesions, optic disc changes, vascular abnormalities, and macular alterations.

D. Calibration and Confidence in Clinical AI

Another major limitation of deep neural classifiers is that predicted probabilities are often poorly calibrated. A model may assign very high confidence to a wrong prediction, which is particularly dangerous in clinical screening systems where confidence can influence referral urgency or downstream decisions [3]. Research on calibration has shown that modern neural networks frequently produce overconfident softmax outputs, even when classification accuracy is strong [10]. In response, post-hoc methods such as temperature scaling have been proposed to improve agreement between predicted confidence and empirical correctness without altering the underlying classifier. For healthcare applications, calibration directly affects operational safety.

E. ODIR and Multi-Disease Retinal Screening

The Ocular Disease Intelligent Recognition (ODIR) dataset has become a common standard for classifying retinal images across multiple disease categories. It contains standard fundus photographs of the left and right eyes obtained at different hospitals and camera systems, along with diagnostic labels [4]. The dataset reflects the variety of real-life acquisition conditions and includes a range of clinically important disease categories. A number of recent studies have proposed advanced architectures and feature-fusion methods for ODIR-based classification, but many emphasize classification accuracy while giving less attention to interpretability, confidence calibration, and screening-oriented workflow design [5].

F. Research Gap and Positioning of the Present Work

From the literature, three clear gaps remain. First, many prior fundus-image systems are optimized primarily for classification accuracy and do not adequately address interpretability. Second, predicted probabilities are often used directly despite known miscalibration, creating risk in triage-oriented settings. Third, existing studies frequently stop at model evaluation and do not frame the system as a workflow-ready screening support tool that can handle uncertain cases more safely [6]. The present work addresses these gaps through RetinaTriage, a five-class retinal screening framework that combines transfer learning, explainability, and confidence-aware inference in a unified pipeline.

III. DATASET AND PROBLEM FORMULATION

A. Dataset Description

The proposed RetinaTriage framework is developed using the Ocular Disease Intelligent Recognition (ODIR) benchmark dataset, a publicly available retinal imaging dataset designed for automated ocular disease analysis. ODIR comprises fundus photographs from

5,000 patients, with one image for each eye, resulting in a total of 10,000 color fundus images [4]. The dataset was acquired from multiple clinical sites using different camera systems, resulting in substantial diversity in image resolution, illumination, contrast, focus quality, and overall acquisition conditions [5]. This variability makes ODIR a suitable benchmark for studying real-world retinal screening problems.

The present work adopts a five-class formulation that is better aligned with the project objective of practical retinal disease screening. The selected classes are Age-Related Macular Degeneration (ARMD), Cataract, Diabetic Retinopathy (DR), Glaucoma, and Normal retina. These categories represent clinically important and visually distinguishable screening targets while allowing the development of a tractable and interpretable multi-class classification pipeline.

B. Data Characteristics and Challenges

The ODIR dataset presents several challenges for model development and evaluation:

- Domain variability: Photographs are sourced from different hospitals and imaging systems, introducing heterogeneous acquisition conditions.
- Class imbalance: A disparate number of samples exists across disease groups, as shown in Table I, with Normal class samples far outnumbering Glaucoma samples.
- Inter-class similarity: Retinal diseases may exhibit similar or overlapping visual patterns between classes, making certain conditions difficult to distinguish purely from image inspection.
- Multi-label complexity: Certain patients may exhibit more than one anomaly, but a basic single-label categorization setup requires the model to select one primary class.

TABLE I. CLASS DISTRIBUTION IN THE ODIR FIVE-CLASS PARTITION

Class	Count	Imbalance Ratio
Normal	4285	1.00× (reference)
Diabetic Retinopathy	2611	1.64×
Cataract	1038	4.13×
ARMD	727	5.89×
Glaucoma	339	12.64×
Total	9000	—

C. Problem Formulation

Let an input fundus image be denoted by $x \in R^{Hxwx3}$, where H and W represent the image height and width, and the three channels correspond to RGB color information. Let the set of five target classes be:

$$\Omega = \{ARMD, Cataract, DR, Glaucoma, Normal\} \quad (1)$$

The objective is to learn a mapping from the input image space to a probability distribution over the five disease classes:

$$f: \mathbb{R}^{H\times W\times 3} \rightarrow \Delta^5 \quad (2)$$

where Δ_5 denotes the 5-dimensional probability simplex. For each input image x, the model produces a probability vector:

$$\hat{p}(x) = [p_1, p_2, \dots, p_5], p_c \geq 0, \sum_{c=1}^{5} \quad p_c = 1 \quad (3)$$

The final predicted label is obtained using the maximum-probability decision rule:

$$\hat{y} = \backslash argmax_{c\in\Omega}\, p_c \quad (4)$$

D. Confidence-Aware Screening Perspective

Although the primary formulation is multi-class classification, the intended application is not limited to assigning a label. In a screening context, the reliability of a prediction is equally important. The system is therefore designed to interpret the predicted probability distribution not only as a classification output but also as a signal for confidence-aware triage. High-confidence abnormal predictions can support referral-oriented recommendations, while uncertain outputs can be flagged for manual review instead of being treated as equally trustworthy.

IV. METHODOLOGY

A. Overall Pipeline

The RetinaTriage pipeline begins with a standard preprocessing stage that normalizes input color fundus photographs. The processed image then passes through a deep convolutional classifier based on EfficientNet-B3, which produces class probabilities for the five target categories: ARMD, Cataract, DR, Glaucoma, and Normal. After classification, a Grad-CAM module generates a visual heatmap highlighting the image regions most influential for the predicted class. Finally, a confidence-aware decision layer evaluates the projected probability distribution to differentiate between stronger predictions and uncertain cases that may require manual review.

B. Image Preprocessing

The fundus photographs in the ODIR collection show a wide range of differences in resolution, lighting, border artifacts, and overall image quality. The unprocessed image is first converted to a standard RGB format. Non-informative dark peripheral areas are minimized through center-focused cropping, and the cropped image is subsequently scaled to a spatial resolution appropriate for EfficientNet-B3. Pixel intensities are normalized on a channel-wise basis to ensure compatibility with pretrained ImageNet weights:

$$x^{\sim}_c = \frac{x_c - \mu_c}{\sigma_c}, c \in \{R, G, B\} \quad (5)$$

This normalization enhances optimization stability and increases the efficacy of transfer learning. Data augmentation during training comprises random flips, minor angle rotations, regulated color disturbances, and subtle blur modifications to mimic real variability in fundus capture while preserving diagnostically significant retinal features.

C. Transfer Learning with EfficientNet-B3

EfficientNet-B3 is adopted as the primary backbone network. EfficientNet balances model depth d, width w, and input resolution r through compound scaling [7]:

$$d = \alpha^{\phi}, w = \beta^{\phi}, r = \gamma^{\phi} \quad (6)$$

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \quad (7)$$

The model is initialized with pretrained ImageNet weights [8]. The original classification head is replaced with a new fully connected layer corresponding to the five target classes:

$$z = W_{cls}^{\top} h + b_{cls}, z \in \mathbb{R}^C \quad (8)$$

During training, the network maps each fundus image to a probability distribution over disease classes using the softmax function:

$$\hat{p}_c = \frac{exp(z_c)}{\sum_{j=1}^{C} exp(z_j)} \quad (9)$$

D. Learning Objective and Optimization

Because the dataset exhibits class imbalance, training employs a class-weighted categorical cross-entropy loss so that errors on less common classes affect optimization more strongly than errors on majority classes [11]:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} \sum_{c=1}^{C} w_c \, y_{ic} log(\hat{p}_{ic}) \quad (10)$$

$$w_c = \frac{N_{train}}{C \cdot N_c} \quad (11)$$

A modern gradient-based optimizer with weight decay and a validation-contingent learning rate schedule is employed. Early stopping based on validation performance is used to prevent overfitting and select the best model checkpoint.

E. Explainability through Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is used to provide post-hoc visual explanations [9]. Grad-CAM computes the gradient of the predicted class score with respect to the feature maps of the last convolutional layer:

$$\alpha_k^{\hat{y}} = \frac{1}{Z}\sum_i \quad \sum_j \frac{\partial \hat{y}}{\partial A_{ij}^k} \quad (12)$$

These gradients weight the activation channels to produce a class activation map:

$$L_{GC}^{\hat{y}} = ReLU(\sum_{k=1}^{K} \alpha_k^{\hat{y}} A^k) \quad (13)$$

The resulting heatmap is resized and superimposed on the original fundus image, producing a visual explanation for the predicted class. This allows inspection of whether the network focuses on clinically plausible regions such as the macula, optic disc, vascular structures, or diffuse pathological patterns.

F. Confidence-Aware Inference and Triage

A central component of RetinaTriage is that classification is not treated as the sole endpoint. To address known overconfidence in deep neural networks, confidence handling is paired with temperature scaling [10]:

$$\hat{q}_c = \frac{exp(z_c/T)}{\sum_{j=1}^{C} exp(z_j/T)} \quad (14)$$

The optimal temperature T* is found by minimizing the negative log-likelihood on the validation set:

$$T^* = \backslash argmin_{T>0}(-\frac{1}{N_{val}}\sum_{i=1}^{N_{val}} log\hat{q}_{y_i}(T)) \quad (15)$$

The calibrated probabilities support a three-tier triage action function:

$$a(x) = \{Urgent\ referral \hat{y} \neq Normal, \hat{q}_{\hat{y}} \geq \tau_H Routine\ review \hat{y} \neq Normal, \tau_L \leq \hat{q}_{\hat{y}} < \tau_H Retake\ /\ defer \hat{q}_{\hat{y}} < \tau_L \quad (16)$$

This design reduces the risk of overconfident misclassification and aligns the model more closely with practical screening use.

G. Modular System Integration

The proposed methodology is structured as a modular pipeline. The classification result, confidence score,

Grad-CAM visualization, and associated metadata are stored in a shared workflow state, enabling integration with downstream functions such as interactive explanation, disease progression analysis, and report generation. Such separation of preprocessing, feature learning, classification, explainability, confidence assessment, and downstream presentation improves maintainability, testing, and future scalability.

V. EXPERIMENTAL SETUP

A. Data Partitioning

To ensure valid model evaluation, the ODIR dataset is divided at the patient level rather than at the image level. This is necessary because each patient contributes both left-eye and right-eye fundus photographs; splitting images independently could lead to information leakage between training and test sets, which would artificially inflate performance estimates. The dataset is partitioned into training, validation, and test sets. All model design choices and hyperparameter adjustments are based only on training and validation data so that the test set remains an unbiased estimate of generalization performance.

B. Training Configuration

The backbone is EfficientNet-B3, initialized with ImageNet weights and subsequently fine-tuned for the five designated classes. Training uses mini-batch optimization with class-weighted categorical cross-entropy as the primary loss function [11]. The optimizer incorporates an adjustable learning rate, weight decay for regularization, and a validation-contingent learning rate schedule. Early stopping is employed to terminate training when validation performance ceases to improve, preventing overfitting.

C. Implementation Environment

The proposed system is implemented in Python using standard scientific computing and deep learning libraries. The training and inference pipeline includes modular components for preprocessing, model training, Grad-CAM generation, and confidence-level output management. Training is performed on GPU hardware with mixed-precision tensor operations to accelerate fine-tuning and inference. The modular design ensures reproducibility and supports component-wise upgrades in future iterations.

D. Evaluation Metrics

Standard multi-class classification metrics are used. Let TP_e, FP_e, TN_e, and FN_e denote the true positives, false positives, true negatives, and false negatives for class c, respectively. The metrics are defined as:

$$Precision_c = \frac{TP_c}{TP_c+FP_c} \quad (17)$$

$$Sensitivity_c = \frac{TP_c}{TP_c+FN_c} \quad (18)$$

$$Specificity_c = \frac{TN_c}{TN_c+FP_c} \quad (19)$$

$$F1_c = 2 \cdot \frac{Precision_c \cdot Sensitivity_c}{Precision_c+Sensitivity_c} \quad (20)$$

Macro-averaged metrics are presented to provide a balanced picture across all classes without allowing majority categories to dominate:

$$Macro - F1 = \frac{1}{C}\sum_{c=1}^{C} F1_c \quad (21)$$

Additionally, macro-AUC is computed using a one-vs-rest approach for each class [13].

E. Calibration Assessment

The Expected Calibration Error (ECE) is computed to evaluate the reliability of predicted probabilities [12]:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \quad (22)$$

where B_m denotes the set of samples whose predicted confidence κ_i falls in the m-th bin, and:

$$acc(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} 1[\hat{y}_i = y_i], conf(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} \kappa_i \quad (23)$$

Calibration is assessed before and after post-hoc temperature scaling to ascertain whether the system's probability outputs improve reliability for triage.

F. Explainability and Triage Assessment

A qualitative assessment of explainability is conducted by examining Grad-CAM heatmaps produced for representative photographs from each category. Heatmaps are analyzed to ascertain whether the model focuses on areas of the retina that are visually coherent and clinically relevant for the expected disease category. The confidence-aware triage analysis additionally evaluates whether uncertainty estimates behave sensibly in borderline or ambiguous cases, particularly when disease patterns overlap or image quality is suboptimal.

VI. RESULTS AND ANALYSIS

A. Classification Performance

The proposed RetinaTriage framework demonstrates strong performance on the five-class retinal disease

classification task. The EfficientNet-B3-based model achieves high overall accuracy while maintaining balanced recognition across the disease categories [14]. Performance is evaluated through class-wise sensitivity, specificity, precision, F1-score, and macro-averaged metrics to account for class imbalance and differences in disease prevalence.

The model performs best on the Normal category, which is expected given the larger training representation and stable visual features of healthy retinal images. The model also performs well on Diabetic Retinopathy and Cataract, demonstrating that the learned feature representations capture disease-relevant patterns effectively. Performance for ARMD and Glaucoma remains competitive but is constrained by fewer samples and subtler visual cues. Macro-averaged metrics are the primary reported figures, as they treat each class equally and are more informative under class imbalance than overall accuracy alone.

B. Calibration Performance

Before calibration, the model exhibits the typical overconfidence behavior common in modern deep neural classifiers. The Expected Calibration Error (ECE) decreases substantially after temperature scaling while classification accuracy remains unchanged [15]. This confirms that post-hoc calibration makes confidence estimates more reliable without altering the ranking of class predictions. The reduction in ECE demonstrates that calibrated confidence values are better aligned with actual prediction correctness and are therefore more suitable for triage-oriented decision support.

C. Explainability Analysis

The Grad-CAM component provides qualitative interpretability of classification results. For diabetic retinopathy, heatmaps highlight retinal areas with lesions and vascular abnormalities [16]. For glaucoma, the optic disc region receives dominant attention. For ARMD, the macular area shows stronger activation. Cataract cases demonstrate more diffuse attention patterns, consistent with the global degradation of image quality caused by lens opacification rather than a localized retinal lesion.

These observations confirm that the classifier is learning disease-relevant visual information rather than relying on image artifacts. Notably, not all heatmaps are anatomically precise, particularly when prediction confidence is low. The correlation between uncertain predictions and lower-quality explanations is informative, suggesting that the model's uncertainty signal and its explainability behavior are mutually consistent.

D. Confidence-Aware Screening Behavior

A practically significant finding is that confidence-aware inference improves the screening orientation of the system. In borderline examples, the class-probability distribution often shows that multiple disease categories receive non-trivial scores even when one class is selected as the final output. This exposes ambiguity instead of hiding it behind a single hard label. Low-confidence outputs can therefore be interpreted as warnings that the image may contain overlapping pathology, poor image quality, atypical presentation, or out-of-distribution visual characteristics.

E. System-Level Interpretation

Beyond numerical performance, the results demonstrate that the proposed framework functions effectively as a system-level screening support pipeline. The integration of classification, explanation, and confidence estimation gives model outputs greater practical value than a conventional closed-set classifier. The classifier provides the disease prediction, Grad-CAM adds a visual rationale, and calibration-aware confidence helps determine how much trust should be placed in the result. This combined behavior is especially important for healthcare-oriented AI, where usability depends not only on correctness but also on transparency and reliability.

F. Error Patterns and Observed Limitations

The results also reveal several limitations. First, minority classes such as Glaucoma remain more challenging than majority classes, reflecting the influence of class imbalance and limited representation. Second, some ambiguous retinal images share visual features across diseases, making hard single-label classification inherently difficult. Third, Grad-CAM does not provide exact lesion segmentation and should not be interpreted as definitive pathology localization. Finally, confidence-aware behavior improves safety but does not eliminate

the need for human oversight in uncertain or clinically complex cases.

VII. DISCUSSION

The results of this study show that effective retinal disease screening from color fundus photographs requires more than strong classification accuracy alone. The proposed RetinaTriage framework demonstrates that combining transfer learning, visual explainability, and confidence-aware inference produces a system more aligned with practical screening needs than a conventional closed-set image classifier. In clinical settings, a model is valuable not only when it predicts correctly, but also when it expresses uncertainty responsibly and provides outputs that can be meaningfully reviewed by a human user.

One of the main findings of the work is that confidence reliability matters as much as discrimination quality. The classifier shows strong predictive performance on the five target classes, but the calibration analysis confirms that raw neural-network confidence should not be accepted uncritically. Temperature scaling improves the reliability of probability outputs without changing classification accuracy, making the system's confidence estimates more usable for triage-oriented decision support [17]. This strengthens the argument that clinical AI systems should explicitly account for calibration before using confidence in downstream decision workflows.

The explainability results also demonstrate practical utility. Grad-CAM visualizations often highlight retinal regions that are anatomically plausible for the predicted disease category [18], suggesting that the model is learning meaningful visual representations rather than relying on spurious correlations. At the same time, the study reinforces an important limitation of post-hoc explainability: attention heatmaps are helpful for qualitative interpretation but do not constitute certified lesion localization or causal reasoning. Therefore, Grad-CAM should be viewed as an interpretability aid, not as a substitute for expert clinical examination.

Another important contribution is the screening-oriented framing of the work. Rather than stopping at benchmark classification performance, this project positions the model as part of a broader workflow in which uncertain cases can be flagged for manual review. This design choice is important because real screening environments operate under imperfect imaging conditions, heterogeneous populations, and varying clinical contexts [19].

Despite these strengths, the present study has several limitations. First, the evaluation is conducted within a single benchmark dataset, and although ODIR contains useful real-world heterogeneity, this does not replace true external validation. Generalization under domain shift remains a critical challenge for retinal AI [20]. Second, the present formulation simplifies the task into a single-label five-class classification problem, whereas real clinical cases may involve multiple concurrent conditions. Third, minority-class difficulty remains visible, particularly for glaucoma and ARMD. Fourth, the study evaluates explainability qualitatively rather than against pixel-level expert annotations, which limits the strength of conclusions about pathological localization.

These limitations point directly to future work. A more advanced version of RetinaTriage should include external multi-center validation, multi-label disease modeling, and image-quality-aware screening logic. Incorporating bilateral eye context, longitudinal clinical history, or lesion-level annotation could further improve both predictive performance and interpretability. Prospective workflow evaluation would also be necessary to determine how confidence-aware outputs actually influence clinician behavior, referral efficiency, and trust in real deployment settings.

VIII. CONCLUSION

This paper presented RetinaTriage, an explainable and confidence-aware retinal disease screening framework for automated analysis of color fundus photographs. The proposed system addresses a five-class classification problem involving ARMD, cataract, diabetic retinopathy, glaucoma, and normal retina using a transfer-learning strategy based on EfficientNet-B3 [21]. In addition to classification, the framework incorporates Grad-CAM for visual interpretability and a confidence-aware inference mechanism to support safer screening-oriented decision making under uncertainty [22].

The experimental results show that the proposed approach achieves strong multi-class retinal disease recognition while improving the practical usability of

model outputs. The explainability component provides qualitative insight into the image regions influencing model predictions, and the calibration-aware confidence framework makes it possible to distinguish reliable predictions from cases that require cautious review. These elements make the system more meaningful for real screening workflows than a conventional classifier that outputs only a label [23].

At the same time, the current study has important limitations, including single-dataset evaluation, simplified single-label classification, and qualitative rather than lesion-level explainability assessment. Future work should focus on external validation, multi-label disease modeling, stronger robustness under domain shift, and prospective workflow evaluation in realistic clinical settings.

In summary, RetinaTriage shows that an effective retinal screening framework should not only predict disease, but also communicate how confidently and on what visual basis that prediction is made. This combination of performance, interpretability, and calibrated uncertainty is essential for building safer and more usable AI systems for ophthalmic screening.

REFERENCES

[1]. R. R. A. Bourne, S. R. Flaxman, T. Braithwaite, M. V. Cicinelli, A. Das, J. B. Jonas, J. Keeffe, J. H. Kempen, J. Leasher, H. Limburg, K. Naidoo, K. Pesudovs, A. Silvester, G. A. Stevens, N. Tahhan, T. Y. Wong, and H. R. Taylor, "Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis," Lancet Global Health, vol. 5, no. 9, pp. e888–e897, 2017.

[2]. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

[3]. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. International Conference on Machine Learning (ICML), pp. 1321–1330, 2017.

[4]. Peking University, "Peking University International Competition on Ocular Disease Intelligent Recognition (ODIR-2019)," Grand Challenge, Shanggong Medical Technology Co., Ltd., China. [Online]. Available: https://odir2019.grand-challenge.org, 2019.

[5]. M. D. Abramoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," IEEE Reviews in Biomedical Engineering, vol. 3, pp. 169–208, 2010.

[6]. N. Gour and P. Khanna, "Multi-class multi-label ophthalmological disease detection using transfer learning based convolutional neural network," Biomedical Signal Processing and Control, vol. 66, p. 102329, 2021.

[7]. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. International Conference on Machine Learning (ICML), pp. 6105–6114, 2019.

[8]. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255, 2009.

[9]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 618–626, 2017.

[10]. J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in Advances in Large Margin Classifiers, A. J. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, pp. 61–74, 1999.

[11]. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. International Conference on Learning Representations (ICLR), 2019.

[12]. M. P. Naeini, G. F. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in Proc. AAAI Conference on Artificial Intelligence, pp. 2901–2907, 2015.

[13]. T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[14]. D. S. Kermany, M. Goldbaum, W. Cai et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, pp. 1122–1131, 2018.

[15]. M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic, "Revisiting the calibration of modern neural networks," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 15682–15694, 2021.

[16]. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. International Conference on Machine Learning (ICML), 2015.

[17]. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 6402–6413, 2017.

[18]. D. S. W. Ting, C. Y. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, X. Baskaran, F. Alam, R. Bhargava, V. A. Barathi, W. T. Fong, F. Tan, Z. Quek, N. M. Tai, H. H. Gan, and T. Y. Wong, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," JAMA, vol. 318, no. 22, pp. 2211–2223, 2017.

[19]. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sanchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.

[20]. B. van Ginneken, B. M. ter Haar Romeny, and M. A. Viergever, "Computer-aided diagnosis in chest radiography: A survey," IEEE Transactions on Medical Imaging, vol. 20, no. 12, pp. 1228–1241, 2001.

[21]. M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in Proc. International Conference on Machine Learning (ICML), pp. 10096–10106, 2021.

[22]. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, pp. 115–118, 2017.

[23]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. International Conference on Learning Representations (ICLR), 2015.