

RetinaSight: A Clinician-in-the-Loop Platform for Robust, Explainable and Uncertainty-Aware Diabetic Retinopathy Screening and Risk Assessment

K.Vedhavarshini¹, P.Divyadharshini², P.M.G.Jegathambal³

^{1,2}Student of DS&IT, VISTAS, Pallavaram, Chennai

³Assistant Professor, VISTAS, Pallavaram Chennai

Abstract

Diabetic retinopathy is a major cause of preventable vision impairment, making timely retinal screening important for people with diabetes. This project presents RetinaSight, a web-based research prototype for organized, explainable diabetic retinopathy screening and future multimodal risk assessment. The application uses React and TypeScript for patient registration, examination management and visualization, while a Python FastAPI backend manages image uploads, SQLite records and PDF reports. A local dataset-import pipeline supports retinal photographs and five-level grading labels from sources such as APTOS 2019 and EyePACS, with separate pathways for external datasets. PyTorch scripts are provided for training candidate convolutional networks, including ResNet18 and EfficientNet-B0, and for generating checkpoint-dependent predictions, entropy scores and Grad-CAM visualizations. The prototype also supports technical image prechecks, clinician-review flags, longitudinal record comparison and quantitative evaluation from labeled prediction files. Fourteen backend tests passed during functional verification. However, no retinal dataset images or trained clinical model are bundled, and diagnostic accuracy, multimodal fusion, image gradability and external clinical validity remain unestablished. The system is intended for research and decision-support development, not autonomous diagnosis.

Keywords: Grad-CAM, FastAPI, EyePACS, APTOS, ResNet18, EfficientNet-B0

INTRODUCTION

Diabetic retinopathy (DR) damages the small blood vessels of the retina. In its early stages a patient usually notices no change in vision, so regular screening with retinal fundus photographs is the main way to find the disease before it causes harm. Grading the photographs takes trained specialists, who are often not available where patients live.

Deep learning has shown that DR can be graded from fundus photographs with high accuracy [1], [2], [3], and an autonomous system has been tested in primary-care offices [4]. Recent studies compare CNN backbones, build ensembles and hybrid models, and add Grad-CAM heat maps [5]; a survey is given in [6]. Studies of deployment, however, show that accuracy is not enough. A field study in clinics reported that poor image quality and workflow friction affected how a DR deep learning system was used [7], and reviews of clinical AI point to the same gap between model performance and clinical impact [8].

Three practical problems follow. First, fundus images are often blurred, badly lit or partly out of frame, and a classifier still returns an answer for them. Second, a bare class label does not tell the clinician how certain the model is, and softmax confidence is often poorly calibrated [9]. Third, a prediction must fit a clinical routine in which a doctor reviews it, records a decision and follows the patient over several visits. RetinaSight takes a system-first approach. It does not propose a new network. It places quality control, uncertainty handling, explanation and clinician review around any model that is connected to it, and it never lets the platform itself produce a clinical value. Its contributions are:

- an implemented staged workflow in which an image-quality gate must be passed before grading, and a saved, non-abstained grading must exist before a risk estimate can be requested;
- abstention by confidence, so that low-confidence results are marked “Review Required” and sent to a clinician instead of being shown as findings;
- a model-agnostic service contract with strict validation of every response, so that missing or inconsistent outputs are rejected and never replaced with invented values;
- a leakage-aware evaluation protocol and a built-in module that measures classification, ranking and calibration quality and keeps internal and external tests apart; and
- clinical workflow features: role-based access, doctor review, notifications, comparison of saved stages and PDF reports.

Section II reviews related work, Section III states the research gap and questions, Sections IV to VII describe the method, architecture, datasets and experimental design, Section VIII reports the results obtained so far.

Related Work

A. DL for DR Grading and Clinical Deployment

Gulshan et al. trained a CNN on retinal photographs and compared it with specialist grades [1], and later evaluated a similar algorithm against manual grading in India [3]. Ting et al. tested a system on images from several ethnic groups [2]. Abramoff et al. reported a prospective trial of an autonomous AI system in primary-care offices [4]. Common backbones include ResNet [10] and EfficientNet [11]. Reddy et al. trained ResNet18, EfficientNetB0, DenseNet121 and MobileNetV2 on APTOS 2019 for five-class grading, added hard and soft voting ensembles and an EfficientNetB0–DenseNet121 hybrid, and report test accuracies of 85.29% for the ensembles and 84.02% for the hybrid on an internal split [5]. All of these studies concentrate on the model.

B. Image Quality

Beede et al. found in Thai clinics that image quality and lighting conditions affected the use of a DR system [7]. Fu et al. evaluated networks that assess retinal image quality [12], and Akter et al. describe a graph-enhanced DR model with image-quality assessment and uncertainty estimation [13]. In RetinaSight, quality is treated as a gate that precedes grading, not as a score reported afterwards.

C. Uncertainty, Calibration and Selective Prediction

Hassan and Ismail and Akram et al. apply Bayesian approaches to DR detection to estimate uncertainty [14], [15]. Monte Carlo dropout [16] and deep ensembles [17] are common estimators. Guo et al. showed that modern networks are often miscalibrated [9], and Ovadia et al. showed that predictive uncertainty degrades under dataset shift [18]. A classifier may also be allowed to reject uncertain inputs, an idea that goes back to Chow [19] and has been applied to deep networks as selective classification [20]. The Brier score [21] and the binned expected calibration error [22] measure how well confidence matches observed

frequency.

D. Explainability

Grad-CAM produces a coarse heat map of the regions that influenced a class [23]. Adebayo et al. showed that some saliency maps do not depend on the model or the data, so a map cannot be trusted without checks [24], and Rudin argues that explanation of a black box is not a substitute for interpretability in high-stakes decisions [25]. Reddy et al. add a concentration score to compare the localisation of maps [5]. RetinaSight therefore shows Grad-CAM only as an aid to inspection.

E. Dataset Shift and Leakage

Dataset shift degrades accuracy and calibration [26], [18], and data leakage inflates reported results across machine-learning-based science [27]. DR datasets differ in camera, population and grading practice. APTOS 2019 [28] and EyePACS [29] are large collections, and Messidor [30], DDR [31] and IDRiD [32] come from other sources and are natural external tests. DR severity is graded on a five-level international scale [33].

Table I compares RetinaSight with the closest work read for this paper.

COMPARISON WITH CLOSEST RELATED WORK

Work	Main focus	Quality check	Uncertainty	Grad-CAM
Reddy et al. [5]	Ensemble and hybrid CNN grading	–	–	Yes
Akter et al. [13]	Graph-enhanced grading	Yes	Yes	–
Akram et al. [15]	Bayesian uncertainty	–	Yes	–
Beede et al. [7]	Field study of a deployed DR system	Studied	–	–
RetinaSight (this work)	Platform around a model	Yes (gate)	Yes (abstain)	Yes, if supplied

A dash means not stated in the sources read, which does not rule it out. RetinaSight provides the workflow slots; the models themselves are external.

RESEARCH GAP AND RESEARCH QUESTIONS

The literature forms a chain. DR classifiers reach high accuracy on research datasets [1], [2], [5]. Deployment studies show that image quality and workflow decide whether such a system is usable [7], [8]. Uncertainty and calibration research gives the tools to abstain [9], [20], but these are rarely part of a complete screening workflow. Reddy et al. themselves name uncertainty-aware prediction, clinical metadata and multi-centre external validation as future directions [5]. In the work reviewed here we did not find a description of a platform in which an image-quality gate must be passed before grading, low-confidence results are routed to a clinician, every model output is validated before storage, and held-out evaluation including calibration is built into the same system. RetinaSight addresses this gap. The research questions are:

- RQ1: Can quality gating, grading, abstention, explanation and risk estimation be organised so that each stage depends on the one before it?
- RQ2: Can a platform accept outputs from an external model while refusing missing or inconsistent ones?

- RQ3: Can held-out predictions be evaluated for accuracy, ranking and calibration inside the same platform, with internal and external tests kept apart?
- RQ4: What does the design guarantee, and what does it not guarantee?

PROPOSED METHODOLOGY

A. Problem Formulation

An input is a retinal fundus image x with clinical data m (age, diabetes duration, HbA1c, blood pressure, previous DR history). The platform asks an external model f for class probabilities p_c over five severity classes c , in percent. Only the model sets the values, and no later stage of the platform may change them. The predicted class and the abstention flag are

$$\hat{y} = \arg \max_c p_c \quad (1)$$

$$a = 1[\max_c p_c < \tau] \quad (2)$$

where τ is the abstention threshold, 70% by default. This default only demonstrates the workflow, and a real threshold has to be chosen on validation data at a clinically meaningful operating point.

B. Image-Quality Gate

Before grading, the model service returns a quality score from 0 to 100, a category (Good, Acceptable, Poor or Unusable) and per-factor scores for blur, illumination, contrast, field of view, retinal visibility and artefacts. If the category is Poor or Unusable, the examination is marked “Review Required”, a retake notification is created and grading is blocked. Re-assessing quality or replacing the image clears earlier grading and risk results. The backend also measures image size, mean brightness, contrast and gradient energy directly from the pixels as a technical precheck. This precheck is shown to the user but is never a quality grade, and it always states that the image is not cleared for grading.

C. Grading and Response Validation

The service returns the predicted class, the five probabilities, a confidence, an uncertainty value and optionally a model-agreement value and a heat map. Before anything is saved, the backend checks that there are exactly five probabilities between 0 and 100, that they sum to 100 within 0.5, that the confidence equals the highest probability, that the predicted class is the one with that probability, and that the uncertainty lies in $[0, 1]$. A heat map must be a valid PNG. A response that fails any check is rejected with HTTP 502 and nothing is saved.

D. Explainability and Multimodal Risk

A heat map supplied by the model service is stored privately and shown beside the original image and in the PDF report. If none is supplied, none is shown, and the platform never draws a placeholder. The risk model receives the image, the clinical data and the saved grading, and returns a score from 0 to 100, a level (Low, Moderate, High or Critical) and contributing factors. A risk request is accepted only after a saved, non-abstained grading exists, and High or Critical results notify the clinician.

E. Longitudinal Comparison and Review

For two saved examinations of one patient, the platform compares the stage labels in time order and reports Stable, Improved or Progressed, or Not available if either examination has no saved prediction. This compares recorded labels and is never presented as confirmed biological progression. A clinician can accept, request re-analysis or mark an examination for manual review, with notes. Accepting records only that the result was looked at, not that it is correct.

SYSTEM ARCHITECTURE

RetinaSight has three layers (Fig. 1). A React and TypeScript interface is used by clinical staff. A FastAPI backend stores records, enforces the workflow and calls the models. A separate Python inference service holds the trained models and is reached over HTTP with a secret key, and over HTTPS unless it runs on the same machine. If the service is not configured, the AI endpoints return HTTP 501, while patient records, uploads, reviews and reports keep working.

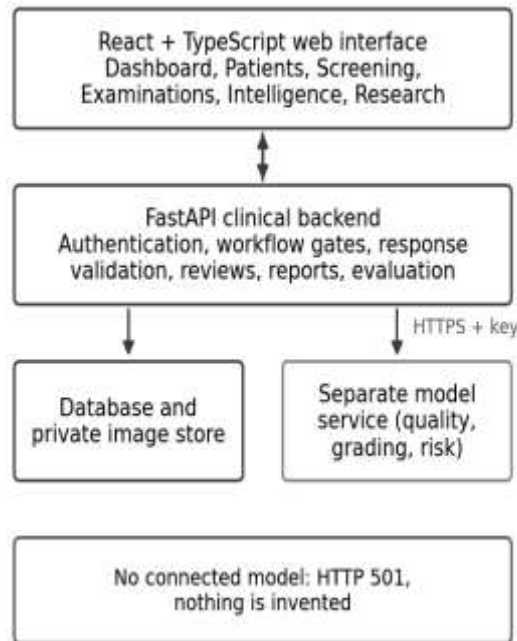


Fig. 1. **Three-layer architecture of RetinaSight.**

Fig. 2 shows the workflow. Each stage starts only after the previous one is saved, and the backend enforces this order, not the interface. Three roles can register: Doctor, Ophthalmologist and Screening Staff. An Administrator needs an invitation. Passwords are stored as salted scrypt hashes, reset links are signed, single-use and expire after 20 minutes, and images are stored privately and served only to authenticated users. The database has nine tables: users, patients, examinations, AI results, risk results, progression, reviews, notifications and dataset evaluations.

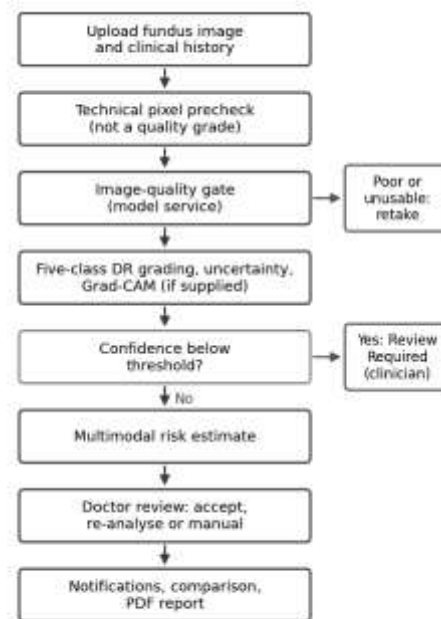


Fig. 2. Staged workflow with gates. Red boxes route to a retake or to a clinician.

DATASETS AND LEAKAGE PREVENTION

A. Planned Datasets

Table II lists the datasets planned for training and testing. None of them has been downloaded for the work reported here, and the platform itself never downloads or generates a dataset.

TABLE I. PLANNED EVALUATION DATASETS

Dataset	Planned role	Note
APTOS 2019 [28]	Training, internal test	No patient identifiers [5]
EyePACS [29]	Training, internal test	Screening-programme images
Messidor-2 [30]	External test	Different source
DDR [31]	External test	Other cameras
IDRiD [32]	External test	Lesion annotations

B. Class Imbalance

APTOS 2019 is strongly imbalanced. Reddy et al. report 3,662 labelled images of which 1,805 are No DR, 999 Moderate, 370 Mild, 295 Proliferative and 193 Severe [5]. A classifier can therefore reach a high accuracy while missing the severe classes, so the evaluation reports macro-averaged metrics, per-class specificity and the full confusion matrix, not accuracy alone.

C. Split Protocol and Leakage Checks

Images are to be split into training, validation and test sets before any augmentation, stratified by class. Where patient identifiers exist, all images of one patient must stay in one subset. APTOS 2019 has no patient identifiers [5], so image-level leakage between the two eyes of one patient cannot be excluded there. This is an inherent limitation of the dataset and is stated with the results. Exact and near-duplicate images

are to be removed with image hashes before splitting. The test set is opened once, after the model, the abstention threshold and the referable-DR operating point have been fixed on validation data.

D. Known Dataset Cues

Datasets built from different cameras can differ in field of view, colour balance and border, and a model may learn such cues instead of disease [27]. Testing on external datasets from other sources is the planned way to measure how much performance depends on them.

EXPERIMENTAL DESIGN

A user with the Doctor, Ophthalmologist or Administrator role uploads a UTF-8 file of held-out predictions with one row per examination, holding the true class and the five probabilities, and enters the dataset name, whether the test is internal or external, the model version and the data source. Each row is checked, probabilities must lie in $[0, 1]$ and sum to 1, and a file may hold at most 100,000 rows. The module reports accuracy and macro precision, recall, F1-score and specificity, macro one-versus-rest AUROC for the classes present, and the 5×5 confusion matrix. Calibration is measured by the multiclass Brier score [21] and the expected calibration error with ten bins [22]:

$$\text{Brier} = (1/N) \sum_n \sum_c (p_{nc} - y_{nc})^2 \quad (3)$$

$$\text{ECE} = \sum_b (n_b / N) | \text{acc}(b) - \text{conf}(b) | \quad (4)$$

where p_{nc} is the predicted probability of class c for image n , y_{nc} is 1 for the true class and 0 otherwise, and n_b is the number of images in confidence bin b . For screening, a referable-DR task is defined as Moderate, Severe or Proliferative DR against No DR or Mild, with the score

$$r = p_{\text{Moderate}} + p_{\text{Severe}} + p_{\text{PDR}} \quad (5)$$

and ROC and precision–recall curves are drawn from r . If a class or the referable task has only one kind of case, the curve is left out instead of being computed wrongly. Internal and external results are saved with their provenance and reported separately. The module labels every result as unverified, because it depends entirely on the uploaded file. Binned ECE is an estimate and is reported together with the Brier score and the calibration curve [22].

RESULTS

No trained model is connected, so this section contains no accuracy or sensitivity values, and none are claimed. It reports what has been implemented and how the safety rules behave. These behaviours follow from the implemented checks and from the automated backend tests, which cover registration and login, role rules, password reset, image upload, the quality gate, abstention, metric import and comparison, using mock model responses that exist only inside the test code.

A. User Interface

The application provides a clinician-in-the-loop review queue, which states that acknowledging an AI result is not a final diagnosis. The multi-dataset validation page likewise clarifies that uploaded metrics constitute research evidence rather than independent verification. All patients and values shown in demonstration mode are synthetic and labelled accordingly.

B. Implementation Status

TABLE II. IMPLEMENTATION STATUS OF PLATFORM COMPONENTS

Component	Status	Still needs
Login, roles, patients, uploads	Implemented	–
Staged workflow gates	Implemented	–
Technical pixel precheck	Implemented	–
Quality, grading and risk inference	Interface implemented	Trained model service
Abstention and review routing	Implemented	Validated threshold
Grad-CAM storage and display	Implemented	Model that makes maps
Longitudinal comparison	Implemented	Saved predictions
Held-out evaluation module	Implemented	Prediction files
Doctor review, alerts, PDF	Implemented	–

Implemented means the code is present in the repository.

C. Behaviour of the Safety Rules

TABLE III. BEHAVIOUR IN CRITICAL SITUATIONS

Situation	Platform behaviour	HTTP
No model service configured	Refuses; integration required	501
Service is not HTTPS	Refuses to send the image	503
Inconsistent probabilities	Rejects; nothing saved	502
Invalid heat map	Rejects; nothing saved	502
Grading before quality check	Blocked	409
Quality Poor or Unusable	Grading blocked; retake alert	409
Confidence below threshold	“Review Required”; clinician alerted	–
Risk after abstention	Blocked	409
Comparing different patients	Refused	422
No saved predictions	Result “Not available”	–

CONCLUSION AND FUTURE WORK

This paper presented RetinaSight, a clinician-in-the-loop platform for diabetic retinopathy screening. It joins an image-quality gate, five-class grading, uncertainty-based abstention, Grad-CAM explanation, multimodal risk estimation, doctor review, longitudinal comparison and reporting in one enforced workflow. It validates every model output, refuses to create results without a model, and includes a leakage-aware module for measuring accuracy, ranking and calibration on held-out data, with internal and external tests kept apart. The platform is a working prototype of the system around a screening model. Its value will depend on the models connected to it and on the validation that is still to be carried out.

The next step is to train and connect real models: an EfficientNet [11] or ResNet [10] backbone fine-tuned on APTOS 2019 and EyePACS for five-class grading, with a separate image-quality model. Calibration such as temperature scaling and an uncertainty method such as Monte Carlo dropout [16] or an ensemble [17] should then be added, and the abstention threshold chosen on validation data. The finished model

should be tested on Messidor-2, DDR and IDRiD with the evaluation module and reported with confidence intervals. Later work can add a risk model trained on a dataset that has clinical metadata, a usability study with eye-care staff, a security review, and support for diabetic macular oedema.

REFERENCES

1. V. Gulshan et al., "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.
2. D. S. W. Ting et al., "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.
3. V. Gulshan et al., "Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India," *JAMA Ophthalmol.*, vol. 137, no. 9, pp. 987–993, 2019.
4. M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *npj Digit. Med.*, vol. 1, Art. no. 39, 2018.
5. G. P. Reddy, P. D. Javali, Ramyashree, S. Raghavendra, C. B. Chandrakala, and P. S. Venugopala, "A hybrid deep learning approach with explainable AI for diabetic retinopathy classification," *IEEE Access*, vol. 14, pp. 28819–28839, 2026, doi: 10.1109/ACCESS.2026.3665564.
6. M. Z. Atwany, A. H. Sahyoun, and M. Yaqub, "Deep learning techniques for diabetic retinopathy classification: A survey," *IEEE Access*, vol. 10, pp. 28642–28655, 2022.
7. E. Beede et al., "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2020, pp. 1–12.
8. C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, Art. no. 195, 2019.
9. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, PMLR vol. 70, 2017, pp. 1321–1330.
10. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
11. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR vol. 97, 2019, pp. 6105–6114.
12. H. Fu et al., "Evaluation of retinal image quality assessment networks in different color-spaces," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2019.
13. Z. Akter, J. I. Ahad, M. M. Farhan, and R. Khan, "Graph-enhanced deep learning for diabetic retinopathy diagnosis: A quality-aware and uncertainty-driven approach," *PLOS Comput. Biol.*, vol. 21, no. 12, Dec. 2025, Art. no. e1013745.
14. M. M. Hassan and H. R. Ismail, "Bayesian deep learning applied to diabetic retinopathy with uncertainty quantification," *Heliyon*, vol. 11, no. 2, Jan. 2025, Art. no. e41802.
15. M. Akram et al., "Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by Bayesian approaches," *Sci. Rep.*, vol. 15, no. 1, p. 1342, Jan. 2025.
16. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, PMLR vol. 48, 2016, pp. 1050–1059.
17. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.

18. Y. Ovadia et al., “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.
19. C. K. Chow, “On optimum recognition error and reject tradeoff,” *IEEE Trans. Inf. Theory*, vol. 16, no. 1, pp. 41–46, 1970.
20. Y. Geifman and R. El-Yaniv, “Selective classification for deep neural networks,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.
21. G. W. Brier, “Verification of forecasts expressed in terms of probability,” *Mon. Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950.
22. M. P. Naeni, G. F. Cooper, and M. Hauskrecht, “Obtaining well calibrated probabilities using Bayesian binning into quantiles,” in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2901–2907.
23. R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.
24. J. Adebayo et al., “Sanity checks for saliency maps,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018.
25. C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.
26. J. Quiñero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Eds., *Dataset Shift in Machine Learning*. Cambridge, MA, USA: MIT Press, 2009.
27. S. Kapoor and A. Narayanan, “Leakage and the reproducibility crisis in machine-learning-based science,” *Patterns*, vol. 4, no. 9, Art. no. 100804, 2023.
28. Asia Pacific Tele-Ophthalmology Society, “APTOS 2019 Blindness Detection,” Kaggle competition dataset, 2019. [Online]. Available: <https://www.kaggle.com/competitions/aptos2019-blindness-detection>
29. J. Cuadros and G. Bresnick, “EyePACS: An adaptable telemedicine system for diabetic retinopathy screening,” *J. Diabetes Sci. Technol.*, vol. 3, no. 3, pp. 509–516, 2009.
30. E. Decencière et al., “Feedback on a publicly distributed image database: The Messidor database,” *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, 2014.
31. T. Li et al., “Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening,” *Inf. Sci.*, vol. 501, pp. 511–522, 2019.
32. P. Porwal et al., “IDRiD: Diabetic retinopathy – segmentation and grading challenge,” *Med. Image Anal.*, vol. 59, Art. no. 101561, 2020.
33. C. P. Wilkinson et al., “Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales,” *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003.