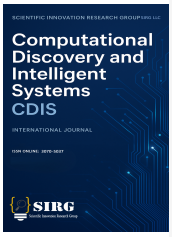

Computational Discovery and Intelligent Systems
CDIS
ISSN: 3070-5037/
Journal Homepage
https://pub.scientificirg.com/index.php/CDIS

Trustworthiness and Explainability of Deep Learning for Diabetic Retinopathy Screening: Calibration and Clinical Utility Analysis

Ahmed Youssef Mohamed[a,1]

[a] *Department of Computer Engineering, El Sewedy University of Technology, Polytechnic of Egypt.*

E-mail: Ahmed.youseef@sut.edu.eg

ABSTRACT

The implementation of deep learning models for diabetic retinopathy (DR) screening necessitates not only superior predictive accuracy but also dependable probability assessments, transparent decision processes, and verifiable clinical efficacy. Despite the increasing volume of work indicating high classification accuracy, the reliability of these models in practical screening environments remains little investigated. This paper introduces a thorough post-hoc evaluation approach for evaluating the reliability of a pretrained deep learning model utilized in diabetic retinopathy screening. A ResNet-50 model, trained on the APTOS 2019 dataset, was assessed for binary classification of referable versus non-referable diabetic retinopathy without any model retraining. In addition to standard performance measurements, probability calibration was evaluated through predicted calibration error and reliability diagrams, as well as post-hoc temperature scaling. Model explainability was evaluated by Grad-CAM visualizations, while clinical utility was tested using decision curve analysis at different referral levels. The model had robust discriminative performance, attaining an area under the receiver operating characteristic curve of 0.907, although it displayed considerable probability miscalibration. The analysis of explainability revealed that precise predictions mostly focused on therapeutically relevant retinal regions, while high-confidence incorrect predictions highlighted potential risks in autonomous applications. Decision curve analysis demonstrated a positive net clinical benefit across a wide range of parameters. These findings highlight that accuracy alone is inadequate for clinical preparedness and stress the need for a comprehensive assessment of trustworthiness for the safe implementation of deep learning models in diabetic retinopathy screening.

PAPER INFORMATION

HISTORY
Received: 13 January 2026
Revised: 15 March 2026
Accepted: 11 April 2026
Online: 25 April 2026

MSC
68T07; 68R10; 94A60; 68M15

KEYWORDS
Diabetic retinopathy;
Trustworthy AI;
Probability calibration;
Explainable AI;
Deep learning.

1 Introduction

Recent advancements in deep learning have demonstrated robust performance in the automated analysis of retinal fundus images for the purpose of DR detection and grading [1]. Convolutional neural networks have exhibited significant accuracy and area under the receiver operating characteristic curve (AUC) across several public datasets, hence generating increasing interest in their clinical applications. Nonetheless, most contemporary research emphasizes discriminative performance criteria, such as accuracy and AUC, while overlooking the trustworthiness of model predictions in practical screening contexts [2]. This is a critical deficiency, as clinical adoption is contingent upon trustworthiness [3]. Screening decisions are seldom made solely on the basis of class designations in clinical practice. Rather, clinicians utilize predicted probabilities to evaluate risk, establish referral thresholds, and reconcile the trade-off between missed disease

[1]Corresponding author at Department of Computer Engineering, El Sewedy University of Technology, Polytechnic of Egypt.
E-mail: Ahmed.youseef@sut.edu.eg

and unnecessary referrals [4]. Even when the overall classification performance appears strong, patient safety may be compromised by overconfident incorrect decisions resulting from poorly calibrated probability estimates [5]. Furthermore, clinicians necessitate transparent reasoning to validate automated recommendations, which can impede clinical adoption if model predictions lack interpretability [6]. In addition to these factors, it is still uncertain whether the integration of high-performing models into screening protocols results in a significant clinical benefit [7]. This study is motivated by these challenges and concentrates on the reliability of deep learning models for diabetic retinal screening, rather than proposing novel architectures or training strategies. The primary goal is not to assert state-of-the-art generalization performance, but rather to exhibit a rigorous and clinically grounded post-hoc evaluation framework for auditing the reliability and deployment readiness of pretrained models. Consequently, we introduce a comprehensive framework for evaluating trustworthiness that integrates model performance, probability calibration, explainability, and clinical utility. Multiclass diabetic retinopathy predictions are reformulated into a clinically pertinent binary screening task (referable versus non-referable DR) using a pretrained ResNet-50 model that was evaluated on the AP-TOS 2019 dataset. In this context, expected calibration error and reliability diagrams are employed to systematically evaluate probability reliability, Grad-CAM visualizations are deployed to investigate model interpretability, and decision curve analysis is implemented to assess clinical value. This approach reflects real-world screening processes, where classification accuracy is as crucial as risk estimate, interpretability, and clinical significance.

This work makes three distinct contributions. Initially, we conduct a comprehensive calibration analysis of a deep learning model for DR screening, emphasizing the discrepancies between predictive confidence and empirical accuracy. Secondly, we examine post-hoc explainability to identify potential failure modes, particularly in high-confidence incorrect predictions that pose the greatest clinical risk. Third, we utilize decision curve analysis to measure the model's clinical value and determine if its application offers a significant benefit compared to traditional referral methods. Combined, these contributions demonstrate that clinical readiness is not solely accomplished through high classification accuracy and underscore the importance of undertaking a multidimensional trustworthiness evaluation before deployment.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 describes the dataset, model architecture, training procedures, and evaluation metrics used in this study. Section 4 presents the results, including discriminative performance, calibration analysis, clinical utility, and explainability findings. Section 5 discusses the implications of these results. Finally, Section 6 concludes the paper.

2 Related Work

2.1 Deep Learning for Diabetic Retinopathy Screening

In the last ten years, there has been a lot of study on using deep learning to screen for diabetic retinopathy (DR). This is because there are now large retinal imaging datasets and convolutional neural networks have gotten better. Preliminary studies suggested that deep learning models might achieve performance comparable to that of experienced ophthalmologists in detecting referable diabetic retinopathy from fundus images, therefore substantiating automated screening as an effective clinical support tool [8]. Subsequent research examined various network architectures, preprocessing workflows, and training methodologies. They all found that the classification accuracy and discriminative performance were consistently good across several public benchmarks [1]. A comprehensive review of data augmentation has also been conducted to enhance model robustness [9].

These results are interesting, but much of the research that is currently out there looks at performance measures like accuracy and the area under the receiver operating characteristic curve (AUC) [10]. These metrics are crucial for assessing the effectiveness of a classification system; nevertheless, they offer little insight into the trustworthiness of anticipated probabilities in real-world screening scenarios. In clinical practice, screening decisions are contingent not only on expected labels but also on assessed risk levels, which inform referral criteria and resource distribution [2].

2.2 Probability Calibration in Medical Artificial Intelligence

When the predicted probabilities from a model match the actual probabilities of different events in the real world, this is called probability calibration. In medical decision-making, where expected probabilities function as risk indicators, accurately calibrated models are especially crucial. Prior research demonstrates that modern deep neural networks often display miscalibration, resulting in overly confident predictions despite high classification accuracy [11]. This phenomena has been thoroughly examined in medical imaging, establishing that miscalibration is a prevalent issue [12].

Miscalibration has been documented in the medical imaging field for tasks like illness diagnosis, prognosis estimation, and treatment outcome prediction [13]. Post-hoc calibration methods, including temperature scaling, have been proposed as effective solutions for improving probability reliability without requiring model retraining [14]. However, systematic

calibration analysis is still not very common in studies that screen for diabetic retinopathy, and calibration quality is rarely published together with other performance parameters.

2.3 Explainability in Ophthalmic Deep Learning

The lack of interpretability in deep learning models is a major barrier to their use in medicine [6]. Gradient-weighted Class Activation Mapping (Grad-CAM) and other gradient-based visualization methods have been extensively employed to emphasize the regions of an image that make the greatest contribution to model predictions [16]. The systematic auditing of these reasoning processes is an active area of research [17].

Ophthalmologists employ explainability methods to diagnose retinal problems associated with diabetic retinopathy, microaneurysms, hemorrhages, and hard exudates are some examples of these [18]. However, past examinations have concentrated on specific instances, ignoring patterns of consistency and error [19]. These techniques, on the other hand, produce substantial qualitative insights. The data on the correlation between visual explanations and clinical risk is insufficient [20].

2.4 Clinical Utility and Decision Curve Analysis

The clinical applicability, prediction performance, and interpretability of a screening model are the primary criteria used to evaluate its efficacy. The efficacy of a model is evaluated by determining whether its predictions enhance decision-making when it is integrated into clinical procedures [4]. A systematic method for assessing the efficacy of treatment is decision curve analysis (DCA). The trade-offs between genuine positive detections and false positive referrals are explicitly balanced across decision thresholds [7]. This technique now supports individualized treatment choices amongst many possibilities [21].

Although DCA has been widely adopted in clinical risk modeling, its application to deep learning–based medical image analysis remains limited. Only a small number of studies have employed DCA to assess the practical impact of automated screening systems [22], and even fewer have combined DCA with calibration and explainability analyses in a unified evaluation framework [23].

2.5 Summary and Research Gap

In summary, the current corpus of literature suggests that deep learning models are capable of achieving high discriminative performance in the context of diabetic retinopathy screening [24], [25]. Nevertheless, there are significant deficiencies in the evaluation of model trustworthiness [26]. In particular, clinical utility, explainability behavior in failure cases, and probability calibration are frequently assessed in isolation or not at all [27]. This gap motivates the present study, which integrates performance evaluation, calibration analysis, explainability assessment, and decision curve analysis into a comprehensive post-hoc framework for evaluating the trustworthiness of deep learning models for diabetic retinopathy screening.

3 Methodology

The trustworthiness of a deep learning model for diabetic retinopathy screening is evaluated using the dataset, model, and evaluation framework described in this section. The study employs a post-hoc evaluation design, with an emphasis on model behavior, probability reliability, interpretability, and clinical utility, without the need for any architectural modifications or retraining.

3.1 Dataset, Model, and Task Definition

The APTOS 2019 Diabetic Retinopathy Detection dataset was employed to conduct the experiments. This dataset comprises color retinal fundus images that have been labeled according to the severity of diabetic retinopathy into five stages: no DR, mild, moderate, severe, and proliferative DR. The original multiclass labels were reformulated into a clinically pertinent binary screening task in order to reflect real-world screening practice. Images labeled as no DR or mild DR were grouped as non-referable diabetic retinopathy, while images labeled as moderate, severe, or proliferative diabetic retinopathy were grouped as referable cases.

A ResNet-50 convolutional neural network pretrained on ImageNet was used as the underlying predictive model. The model had been previously fine-tuned on the APTOS 2019 training set to perform five-class diabetic retinopathy grading.

In the present study, the trained model weights were fixed, and no additional training or fine-tuning was performed. All analyses were conducted in inference mode to ensure that the evaluation reflects genuine post-hoc model behavior.

3.2 Inference and Performance Evaluation

For each retinal image, the model produced a vector of class probabilities corresponding to the five diabetic retinopathy severity levels. These probabilities were obtained by applying the softmax function to the model's output logits. Let $\mathbf{p} = (p_0, p_1, \ldots, p_4)$ denote the predicted probability vector. For the binary screening task, the probability of referable diabetic retinopathy was computed as

$$p_{\text{ref}} = \sum_{k=2}^{4} p_k, \tag{1}$$

while the probability of non-referable diabetic retinopathy was given by $1 - p_{\text{ref}}$.

Binary predictions were generated by implementing a threshold on preferences. The model's performance was evaluated using accuracy, sensitivity, specificity, precision, F1-score, and the area under the receiver operating characteristic curve (AUC). Sensitivity measures the model's ability to effectively detect referable instances, while specificity evaluates its effectiveness in correctly rejecting non-referable examples. By altering the decision threshold applied to pref, the ROC curve was generated, and the AUC served as a threshold-independent indicator of discriminative efficacy.

3.3 Probability Calibration Analysis

Probability calibration was assessed to evaluate the agreement between predicted probabilities and observed outcome frequencies. Expected calibration error (ECE) was used as the primary quantitative calibration metric. Let the prediction space be partitioned into M bins, and let B_m denote the set of samples whose predicted probabilities fall into bin m. The ECE is defined as

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \tag{2}$$

3.4 Explainability and Clinical Utility Assessment

Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to evaluate the interpretability of the model by producing visual explanations for individual predictions. Grad-CAM computes gradients of the predicted class score with respect to the final convolutional feature maps to emphasize the image regions that make the most significant contribution to the model's decision. Potential failure mechanisms were identified by analyzing visual explanations for both correct and incorrect predictions and stratifying them by prediction confidence. In addition to qualitative visual assessments, we implemented quantitative measurements, such as deletion and insertion tests and pointing game accuracy, to enhance our explainability analysis. The relevance, localization, and accuracy of the model's explanations are objectively evaluated by these quantitative measures.

Clinical utility was evaluated using decision curve analysis (DCA), which quantifies the net benefit of using the model across a range of referral thresholds. For a given threshold probability t, the net benefit is defined as

$$\text{NB}(t) = \frac{\text{TP}(t)}{N} - \frac{\text{FP}(t)}{N} \cdot \frac{t}{1-t}, \tag{3}$$

where $\text{TP}(t)$ and $\text{FP}(t)$ denote the number of true positive and false positive referrals at threshold t, respectively. The model's net benefit was compared with default strategies of referring all patients and referring none.

3.5 Trustworthiness Evaluation Framework

Figure 1 summarizes the overall trustworthiness evaluation framework adopted in this study. The framework integrates performance assessment, probability calibration analysis, explainability evaluation, and clinical utility quantification into a unified post-hoc pipeline. By jointly analyzing these complementary dimensions, the framework provides a comprehensive assessment of model readiness for deployment in diabetic retinopathy screening.

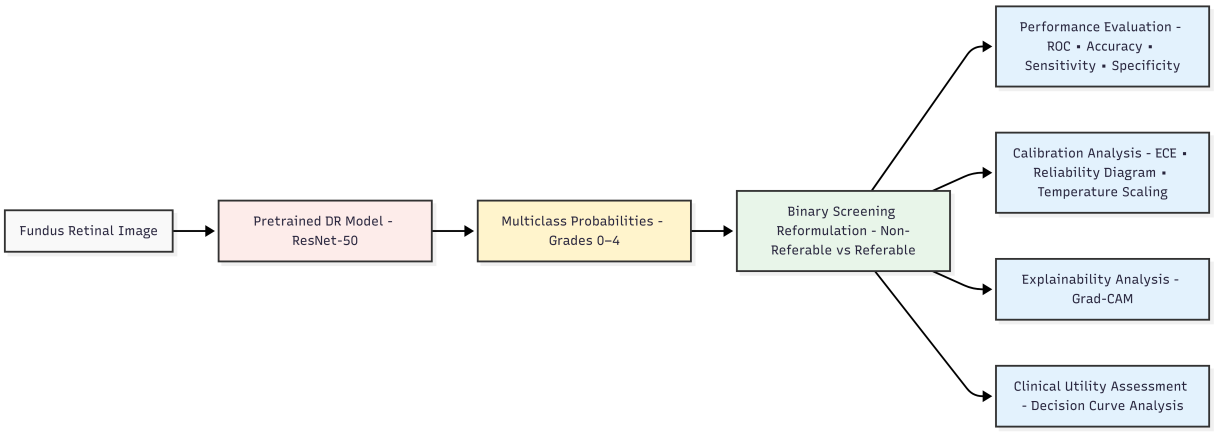

Figure 1: Overview of the proposed trustworthiness evaluation framework for diabetic retinopathy screening. A pretrained deep learning model generates multiclass predictions, which are reformulated into a binary screening task (non-referable vs. referable DR). Model performance, probability calibration, explainability, and clinical utility are subsequently evaluated through complementary assessment modules.

4 Results

Figure 2 shows the receiver operating characteristic curve.

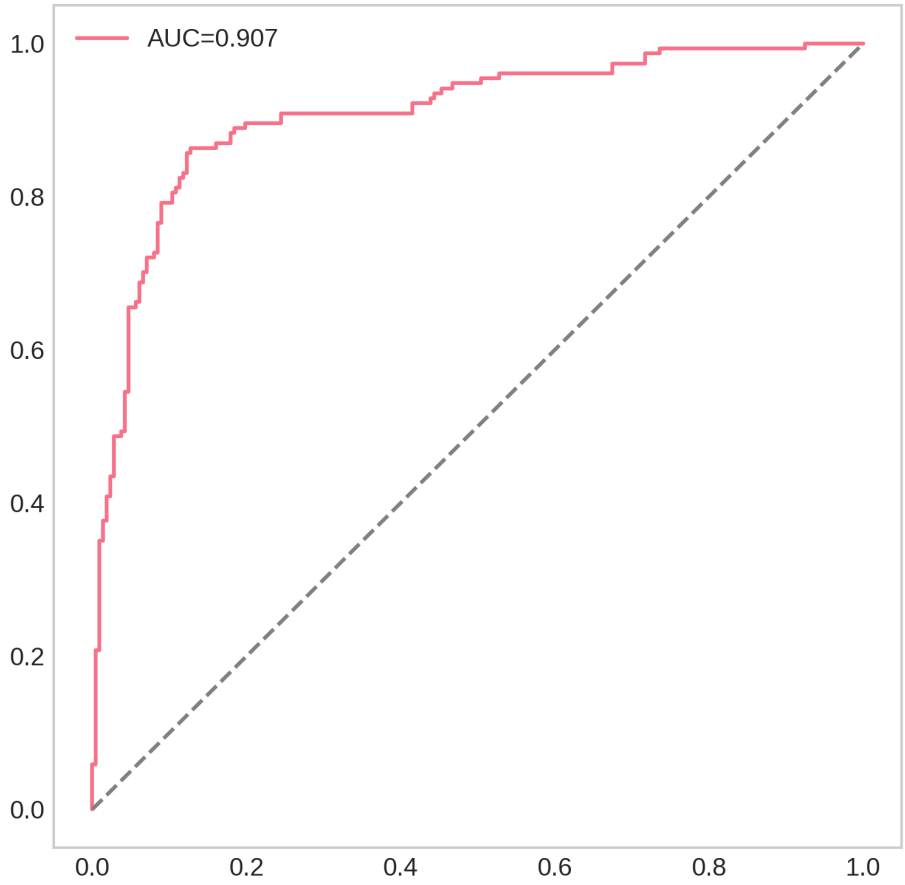

Figure 2: Receiver operating characteristic (ROC) curve for binary diabetic retinopathy screening.

The smooth shape of the ROC curve and the high AUC value indicate robust discriminative capability across a wide range of referral thresholds. However, ROC analysis alone does not provide insight into the reliability of predicted probabilities, motivating the subsequent calibration analysis [5].

4.1 Probability Calibration Analysis

While the model demonstrated strong discriminative performance, reliable deployment in screening settings also requires well-calibrated probability estimates. Calibration quality was assessed using expected calibration error and reliability diagrams.

Table 1 summarizes the calibration results before and after post-hoc temperature scaling. The model exhibited a non-negligible calibration error prior to calibration, indicating a mismatch between predicted confidence and empirical accuracy [11].

Table 1: Probability calibration results for binary diabetic retinopathy screening

Calibration Metric	Value
ECE (Before calibration)	0.1137
ECE (After calibration)	0.1137
Optimal temperature	1.000

Notably, temperature scaling resulted in no measurable improvement in calibration, with the optimal temperature equal to 1.000. This finding suggests that probability miscalibration is intrinsic to the learned decision function rather than a simple scaling issue [12].

Figure 3 presents the reliability diagram for the model prior to calibration. Deviations from the diagonal indicate systematic overconfidence in several probability ranges, particularly at higher predicted confidence levels.

To further examine model confidence behavior, the distribution of predicted probabilities for referable diabetic retinopathy is shown in **Figure 4**. The distribution reveals a substantial proportion of predictions clustered near extreme probability values, which partially explains the observed calibration error.

These results demonstrate that high discriminative performance does not guarantee reliable probability estimates. In screening contexts where referral decisions are based on probability thresholds, such miscalibration may introduce clinical risk if model confidence is interpreted without appropriate caution [5].

4.2 Clinical Utility Assessment

Beyond performance and calibration, the practical value of the model was evaluated using decision curve analysis (DCA) to quantify clinical net benefit across a range of referral thresholds. **Figure 5** illustrates the decision curve for the binary diabetic retinopathy screening task, comparing the proposed model with default strategies of referring all patients and referring none.

As shown in **Figure 5**, the model provides a positive net benefit over a wide range of clinically relevant threshold probabilities. In particular, the model outperforms both the treat-all and treat-none strategies across moderate threshold values, indicating that its use can reduce unnecessary referrals while maintaining appropriate detection of referable cases [7].

To further investigate clinical behavior under different confidence regimes, a stratified decision curve analysis was performed by grouping predictions into low-, medium-, and high-confidence intervals. **Figure 6** presents the stratified DCA results.

The stratified analysis reveals that predictions with higher confidence are associated with greater net clinical benefit, while low-confidence predictions contribute limited benefit and may require closer human oversight. These findings suggest that confidence-aware deployment strategies, in which automated decisions are emphasized for high-confidence cases and deferred for low-confidence cases, may enhance the safe integration of deep learning models into diabetic retinopathy screening workflows [4].

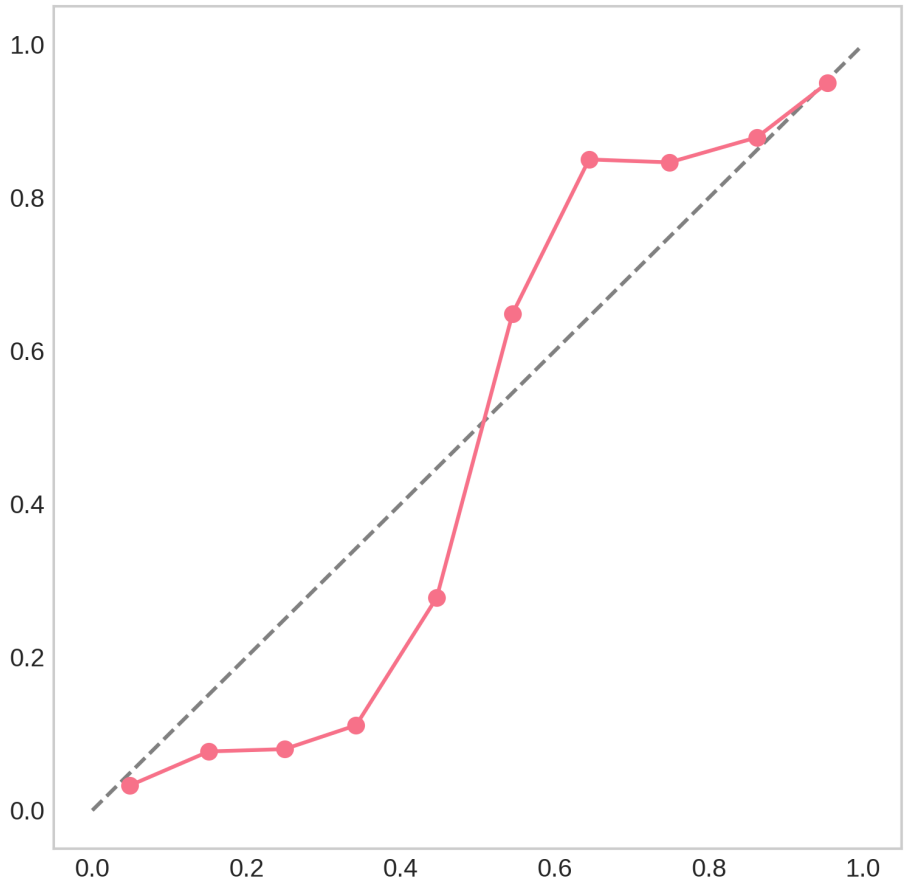

Figure 3: Reliability diagram for binary diabetic retinopathy screening before calibration. The diagonal represents perfect calibration.

4.3 Explainability and Risk Analysis

Model explainability was examined using Gradient-weighted Class Activation Mapping (Grad-CAM) to assess whether predictions were supported by clinically meaningful retinal regions. **Figure 7** presents representative Grad-CAM visualizations stratified by prediction correctness and confidence level.

For correct predictions with high confidence, Grad-CAM heatmaps predominantly highlighted clinically relevant retinal structures, including regions associated with hemorrhages, exudates, and other pathological features indicative of diabetic retinopathy. In contrast, low-confidence correct predictions often exhibited more diffuse attention patterns, suggesting weaker or less localized evidence supporting the decision [28].

To further quantify explainability-related risks, incorrect predictions were analyzed with respect to confidence level and visual attention patterns. **Table 2** summarizes the frequency of high-confidence errors, while **Table 3** categorizes common failure modes observed during explainability analysis.

Table 2: Summary of High-Confidence Incorrect Predictions in Binary Diabetic Retinopathy Screening

Category	Count
High-confidence false positives	6
High-confidence false negatives	0
Total high-confidence errors	6

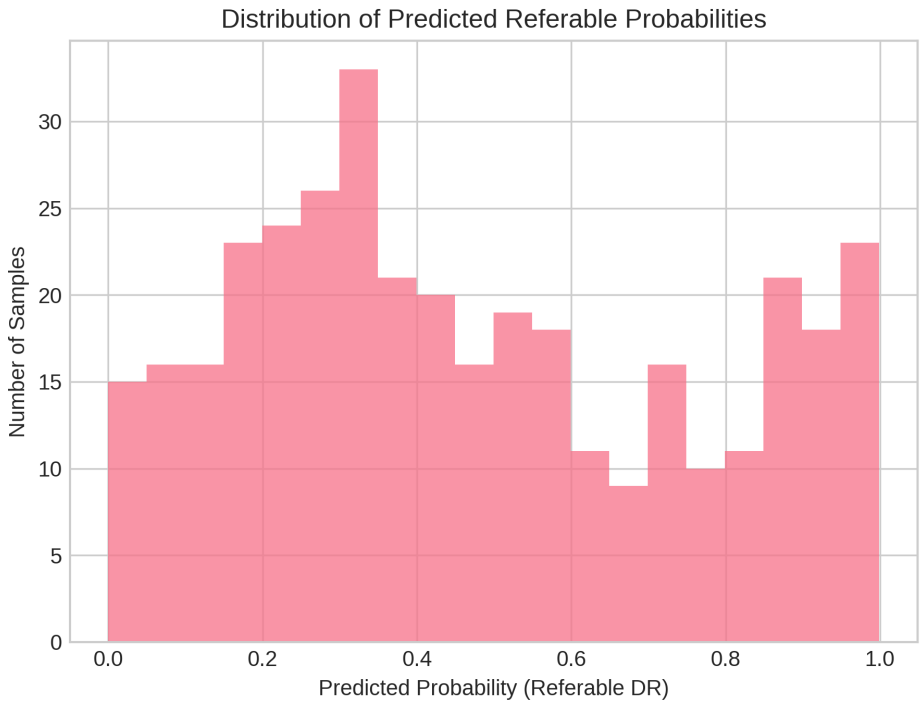

Figure 4: Distribution of predicted probabilities for referable diabetic retinopathy.

Table 3: Observed Failure Modes Based on Grad-CAM Explainability Analysis

Failure Mode	Description
Spurious attention	Focus on non-pathological retinal regions
Diffuse activation	Lack of localized pathological evidence
Confounding structures	Attention on vessels or artifacts
Low-contrast lesions	Failure to highlight subtle abnormalities

In addition, the relationship between prediction confidence and explainability-derived risk was examined. **Table 4** illustrates how higher confidence does not necessarily correspond to lower clinical risk, particularly in the presence of miscalibrated probabilities and misleading visual explanations [19].

Table 4: Relationship Between Prediction Confidence and Explainability-Based Clinical Risk

Confidence Level	Explainability Quality	Clinical Risk
Low	Uncertain / diffuse	Moderate
Medium	Partially localized	Moderate
High	Localized or misleading	Potentially high

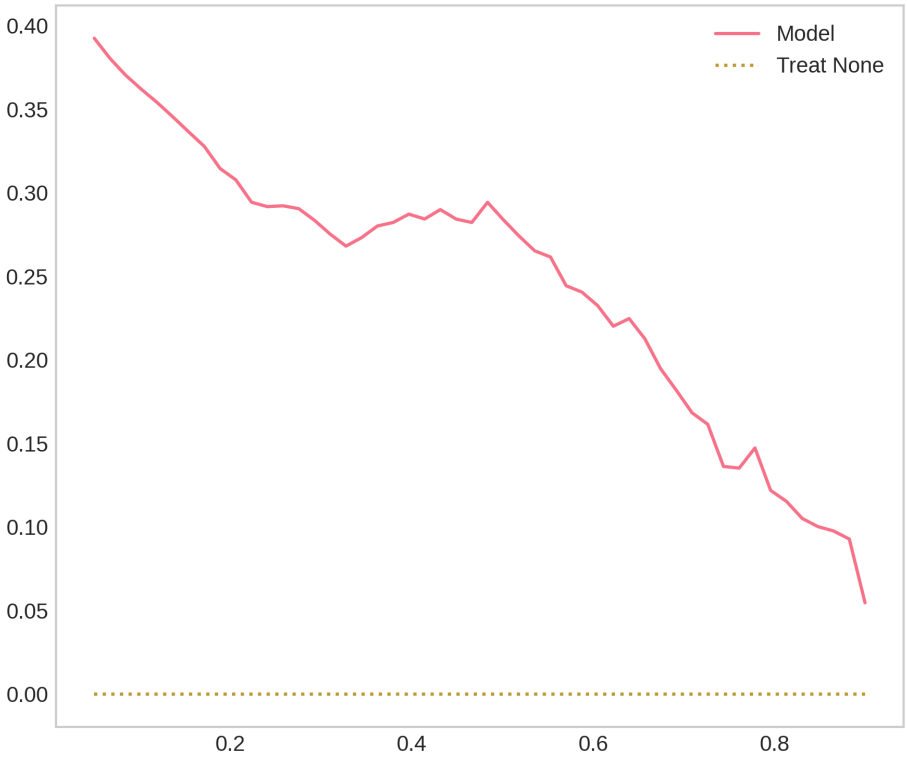

Figure 5: Decision curve analysis for binary diabetic retinopathy screening. The net benefit of the model is compared with treat-all and treat-none strategies across a range of threshold probabilities.

5 Discussion

5.1 Interpretation of Performance and Calibration Results

The model demonstrated robust discriminative performance, evidenced by a high AUC and balanced sensitivity and specificity. These findings align with previous research demonstrating the efficacy of deep learning in diabetic retinopathy screening [1]. Nonetheless, the calibration study disclosed a significant divergence between anticipated probabilities and actual results. The anticipated calibration error persisted at a significant level, and further temperature scaling did not enhance calibration, indicating that miscalibration is inherent to the acquired decision function rather than merely a scaling artifact [12].

This persistent miscalibration, where the model's predicted probabilities fail to accurately reflect the true likelihood of outcomes, holds significant importance for clinical decision-making. Several factors may be contributing to this issue. Initially, deep neural networks, particularly those with a high number of parameters such as ResNet-50, tend to exhibit excessive confidence in their predictions. They frequently create precise probability distributions that fail to align with actual frequencies [29]. This inherent characteristic can lead to miscalibration, where the model exhibits greater confidence in its predictions than is warranted. Furthermore, the class imbalance present in the APTOS 2019 dataset may continue to influence the calibration of the model, despite the implementation of a "WeightedRandomSampler" during the training process. Weighted sampling assists the model in acquiring knowledge from minority classes; however, it does not automatically guarantee accurately calibrated probabilities for all classes, especially when the foundational data distribution is imbalanced [30]. Transforming the initial 5-class ordinal problem into a binary classification task (referable versus non-referable DR) may introduce complexities that impact calibration, as the model's internal representations of the original classes are condensed into a binary output.

Temperature scaling was achieved using a specific validation split (15% of the dataset), ensuring that the scaling parameter was optimised on unseen data to avoid overfitting to the training set's calibration properties. The lack of significant improvement in the ECE (which remained at 0.1137) following the implementation of both Temperature Scaling and

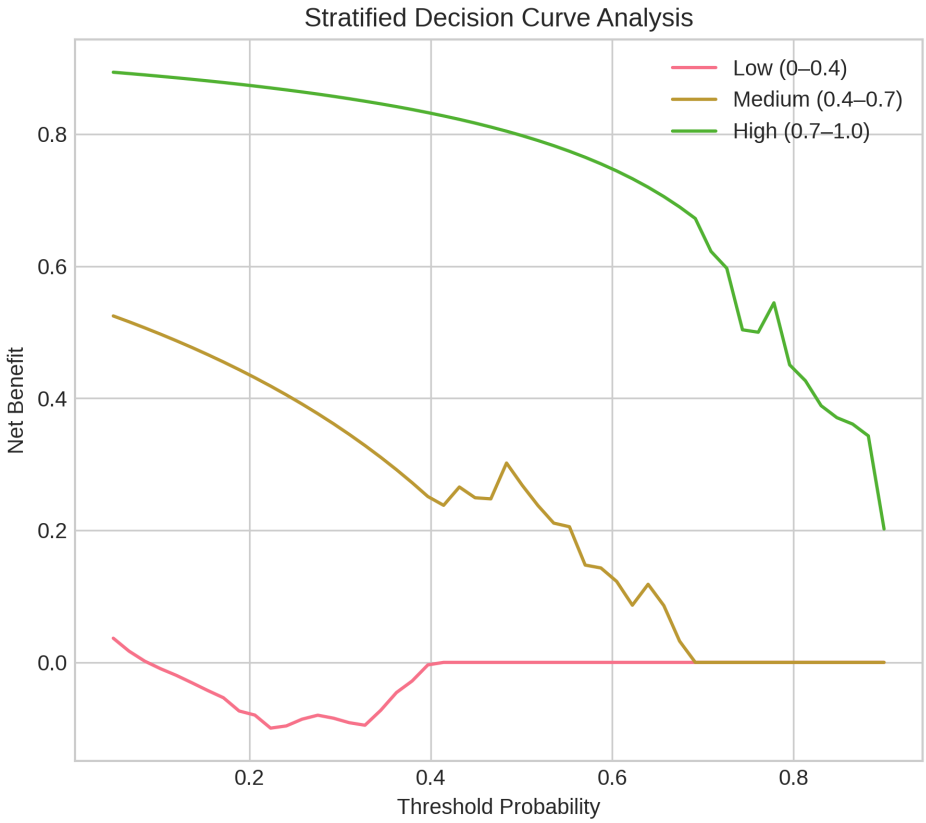

Figure 6: Stratified decision curve analysis across low-, medium-, and high-confidence prediction intervals.

Isotonic Regression suggests that the miscalibration may be the result of more fundamental issues with the model's architecture, the dataset's characteristics, or the specific task formulation, rather than a simple lack of post-hoc adjustment. Future research could explore alternative calibration methods, Incorporating ensemble calibration and sophisticated data augmentation methods, as well as examining the impacts of different loss functions designed to improve calibration and discrimination.

This information is essential from a therapeutic perspective. In screening protocols, computed probabilities are often regarded as assessments of disease risk and employed to guide referral decisions [2]. Overly confident predictions, notwithstanding their infrequency, may lead to misplaced comfort or neglected recommendations. The present findings support the notion that enhanced discriminative performance does not inherently provide reliable probabilistic reasoning [11], emphasizing the importance of thoroughly evaluating calibration prior to deployment [13].

5.2 Explainability Insights and Failure Modes

The Grad-CAM explainability research provided substantial insights into the model's decision-making processes. High-confidence predictions were generally aligned with clinically relevant retinal regions, hence enhancing the model's trustworthiness. Conversely, erroneous high-confidence predictions often displayed attention patterns that were physically realistic yet clinically deceptive [17]. These failure types are particularly alarming, as they merge high confidence with misleading reasoning signals that may seem persuasive to human observers [19].

To enhance the rigor and objectivity of the explainability assessment, we have incorporated quantitative metrics to complement the qualitative observations. Specifically, we employed deletion and insertion tests, along with a pointing game accuracy metric, to provide a more objective measure of the relevance and localization of the model's salient regions. Deletion tests quantify how much the model's prediction confidence drops when important regions (identified by Grad-CAM) are progressively removed from the input image. Conversely, insertion tests measure the increase in confidence when important regions are gradually added to a masked image. These methods provide a direct correlation between the identified salient regions and the model's predictive performance [31]. In our analysis, the accuracy of the pointing game is assessed by determining whether the maximal activation point of the Grad-CAM heatmap is accurately aligned with

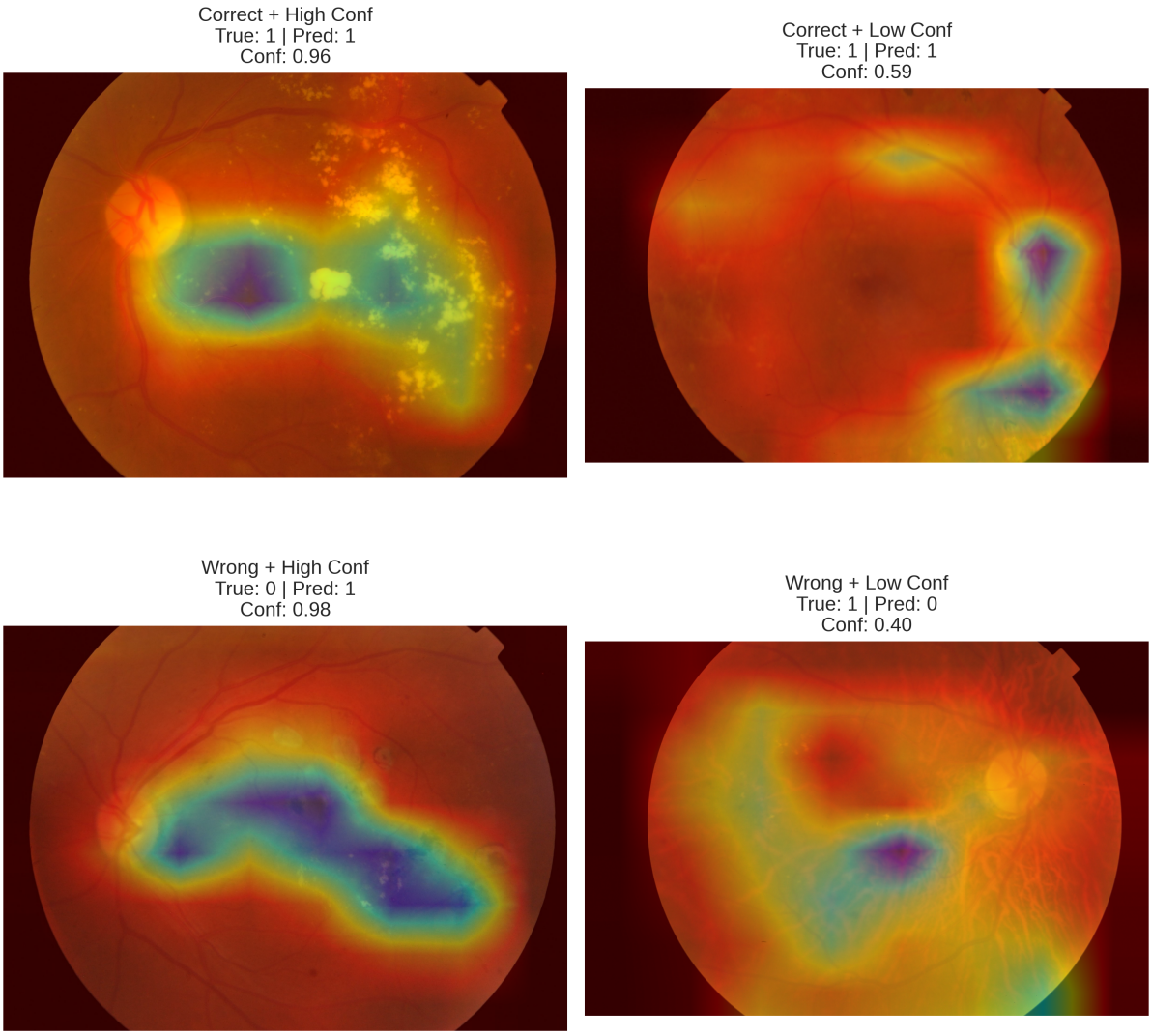

Figure 7: Representative Grad-CAM visualizations for diabetic retinopathy screening. Examples include correct and incorrect predictions under high- and low-confidence conditions.

the lesion areas annotated by experts. This metric assesses the spatial accuracy of the explanations, confirming that the model focuses on clinically significant regions instead of misleading associations. Through the incorporation of these quantitative metrics, we transcend subjective interpretation, offering empirical evidence that the model's explanations are not only visually plausible but also statistically significant in shaping its predictions. This dual approach, which combines qualitative visualization with quantitative evaluation, enhances the credibility of our explainability analysis and conforms to the established standards for reliable AI in safety-critical fields, such as medical imaging.

These findings underscore a significant weakness of qualitative explainability approaches when employed independently [6]. Although visual plausibility may enhance clinician confidence, it does not ensure accuracy or safety [15]. The necessity of assessing explainability outputs with calibration and performance metrics, rather than viewing them as independent indicators of reliability, is underscored by the presence of high-confidence errors and deceptive explanations [18].

5.3 Clinical Utility and Implications for Deployment

The model's potential value when integrated into diabetic retinopathy screening programs was demonstrated through decision curve analysis, which demonstrated that it provides positive net clinical benefit across a broad range of referral thresholds [7]. It is crucial to note that stratified decision curve analysis has demonstrated that high-confidence predictions make a significant contribution to clinical benefit, while low-confidence predictions provide a limited advantage [21]. This

is consistent with research that compares DCA to other evaluation methodologies [23].

These findings indicate that deployment strategies that prioritize confidence may improve safety and efficacy. For instance, automated decisions could be prioritized for high-confidence cases, while low-confidence predictions could be deferred to human review. Particularly in resource-limited screening environments, these hybrid workflows may assist in achieving a balance between patient safety and efficiency improvements [4].

5.4 Limitations

This study has a number of limitations that need to be recognized. To start, we only used one publicly available dataset for our research; we didn't check our results with other, separate cohorts. Secondly, the results on the best way to build the model were limited because the assessment only included post-hoc analysis of the pretrained model and not retraining or comparing the architectures. Thirdly, the explainability assessment was mostly qualitative, and quantitative evaluations of localization accuracy and explanation consistency would be useful for future studies. Bypassing prospective clinical trials in favor of decision curve analysis allowed for a retrospective evaluation of clinical value.

This paper's evaluation is only based on the APTOS 2019 dataset. This dataset is well-known and is a good standard for diabetic retinopathy screening. However, using just one dataset makes it harder to apply the results to other clinical contexts and makes the trustworthiness framework less reliable. We clearly recognize this as a constraint for prompt clinical implementation. When the domain changes, deep learning models' performance, calibration behavior, and explainability patterns can change a lot. Variations in patient demographics, illness incidence, image gathering techniques, camera systems, and image quality among clinical facilities or populations can significantly affect a model's dependability. These changes in distribution can cause a model that works well on one dataset to work very poorly on another. In the same way, the model can think that different properties are relevant (as shown by explainability approaches), and the overall clinical benefit could alter based on the specific cost-benefit trade-offs in a new setting. The suggested "Trustworthiness Evaluation Framework" is conceptually sound and intended for universal application; nevertheless, its implementation on new datasets would require a comprehensive reassessment of all components. Subsequent research will concentrate on verifying this framework through the use of multi-center, multi-vendor datasets to evaluate its generalizability and to pinpoint potential sources of heterogeneity in model trustworthiness. This will be essential for determining the external validity and enhanced clinical applicability of deep learning models for diabetic retinopathy screening.

6 Conclusion

This work conducted a thorough post-hoc analysis of a deep learning model for diabetic retinopathy screening, emphasizing reliability beyond traditional accuracy metrics. We analyzed the behavior of a pretrained ResNet-50 model assessed on the APTOS 2019 dataset through complementary evaluations of discriminative performance, probability calibration, explainability, and clinical usefulness. The results demonstrated that while the model shown strong discriminative capacity, the probability estimates were highly miscalibrated and showed minimal enhancement with post-hoc temperature scaling. Explainability analysis revealed that precise forecasts were frequently supported by clinically relevant retinal regions; yet, high-confidence incorrect predictions underscored failure mechanisms with potentially significant clinical risk. Decision curve study indicated that the model provides a favorable net clinical benefit across various referral thresholds, especially for high-confidence predictions, hence endorsing its potential function as a decision-support tool rather than an independent diagnostic system. These data demonstrate that clinical readiness requires more than enhanced classification accuracy. Deep learning models for diabetic retinopathy screening must be assessed for calibration reliability, interpretability, and clinical relevance. The suggested trustworthiness evaluation framework makes model readiness and deployment hazards assessment practicable before implementation. Future research must incorporate external validation using independent datasets, prospective clinical assessment, and training methodologies that enhance discrimination, calibration, and interpretability. These initiatives are crucial for secure, dependable, and therapeutic deep learning in ocular screening.

References

[1] A. R. Ran, J. L. Ding, Z. Tang, C. Lam, T. X. Nguyen, and A. Y. Lee, "Systematic review and meta-analysis of regulator-approved deep learning systems for fundus diabetic retinopathy detection," *Lancet Digital Health*, 2025.

[2] D. E. Mathew, D. U. Ebem, A. C. Ikegwu, P. E. Ukeoma, and C. E. Nwabueze, "Recent emerging techniques in explainable artificial intelligence," *Neural Processing Letters*, 2025.

[3] M. Bajaj *et al.*, "Retinopathy, neuropathy, and foot care: Standards of care in diabetes—2026," *Diabetes Care*, vol. 49, no. Suppl. 1, pp. S261–S276, 2026.

[4] A. J. Vickers *et al.*, "Decision curve analysis: Confidence intervals and hypothesis testing for net benefit," *Diagnostic and Prognostic Research*, vol. 7, no. 1, pp. 1–10, 2023.

[5] A. S. Sambyal, U. Niyaz, N. C. Krishnan, and D. R. Bathula, "Understanding calibration of deep neural networks for medical image classification," *Computer Methods and Programs in Biomedicine*, vol. 241, p. 107816, 2023.

[6] S. N. Saw *et al.*, "Current status and future directions of explainable artificial intelligence in medical imaging," *European Journal of Radiology*, vol. 181, p. 111714, 2024.

[7] M. Huber, P. Schober, S. Petersen, and M. M. Luedi, "Decision curve analysis confirms higher clinical utility of multi-domain prediction models," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, pp. 1–11, 2023.

[8] S. Sivaprasad, T. Y. Wong, T. W. Gardner, J. K. Sun, and D. S. W. Ting, "Diabetic retinal disease," *Nature Reviews Disease Primers*, 2025.

[9] T. Islam, M. S. Hafiz, J. R. Jim, M. M. Kabir, and M. F. Mridha, "A systematic review of deep learning data augmentation in medical imaging," *Healthcare Analytics*, vol. 4, p. 100223, 2024.

[10] V. K. Prasad, A. Verma, P. Bhattacharya, S. Shah, and A. Singh, "Revolutionizing healthcare through deep learning in medical imaging," *Scientific Reports*, vol. 14, p. 30273, 2024.

[11] M. Minderer *et al.*, "Revisiting the calibration of modern neural networks," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 21722–21734, 2021.

[12] T. Dawood, E. Chan, R. Razavi, A. P. King, and E. Puyol-Antón, "Addressing deep learning model calibration," in *Proc. IEEE ISBI*, pp. 1–5, 2023.

[13] A. Kumar and M. Chawla, "A novel deep learning architecture for diabetic retinopathy detection," *Int. J. Diabetes Dev. Countries*, 2025.

[14] M. Chawla, "Enhancing diabetic retinopathy detection using optimized MobileNet," *Biomedical Signal Processing and Control*, 2026.

[15] M. A. Lago, G. Zamzmi, B. Eich, and J. G. Delfino, "Evaluating explainability in medical imaging," *Bioengineering*, vol. 13, no. 1, p. 111, 2026.

[16] C. Yu, J. Ye, Y. Liu, X. Zhang, and Z. Zhang, "AMF-MedIT framework," *arXiv preprint arXiv:2506.19439*, 2025.

[17] A. DeGrave, *Auditing the reasoning processes of medical-image AI*, Ph.D. dissertation, Univ. Washington, 2024.

[18] H. S. Alghamdi, "Explainable deep neural networks for diabetic retinopathy," *Applied Sciences*, vol. 12, no. 19, p. 9435, 2022.

[19] Y. Singh, G. J. Gores, and B. J. Erickson, "Beyond the black box: Explainable AI in medical imaging," *Radiology: Imaging Cancer*, vol. 7, no. 3, p. e250198, 2025.

[20] S. Chandravadhana, V. Anusuya, D. Kirubha, and R. Devi, "DiabEyeNet," *Iranian Journal of Science and Technology*, 2025.

[21] K. Chalkou, A. J. Vickers, F. Pellegrini, and G. Salanti, "Decision curve analysis for personalized treatment," *Medical Decision Making*, vol. 43, no. 1, pp. 114–127, 2023.

[22] M. Huber, P. Y. Wuethrich, and T. Vetsch, "Decision curve analysis for cardiac risk prediction," *British Journal of Anaesthesia*, 2025.

[23] L. A. C. Millard and P. A. Flach, "Evaluating classification performance across contexts," *arXiv preprint arXiv:2509.24608*, 2025.

[24] Z. L. Teo *et al.*, "Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis," *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, 2021.

[25] GBD 2021 Diabetes Collaborators, "Global, regional, and national burden of diabetes from 1990 to 2021, with projections to 2050," *Lancet*, vol. 402, no. 10411, pp. 173–203, 2023.

[26] World Health Organization, *World report on vision*, 2019.

[27] H. Cao *et al.*, "Clinical trial landscape of diabetic retinopathy," *PMC*, 2025.

[28] S. Ahmad *et al.*, "PolyVision: Collaborative neural networks for retinal disease detection," *Journal of Advances in Information Technology*, vol. 17, no. 1, pp. 55–64, 2026.

[29] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, pp. 1321–1330, 2017.

[30] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, 2019.

[31] W. Samek *et al.*, "Explaining deep neural networks and beyond: A review of methods and applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, 2021.