Pakistan Journal of Scientific Research
Multidisciplinary & Peer Reviewed
Volume: 5, Number: 2, Pages: 34-54, Year:2025

Review

Transparent Prediction of Diabetic Retinopathy Using Machine Learning and Fuzzy Logic

Tahir Mahmood Bashir* **and Arfan Jaffar**

[1] Department of Computer Science, Superior University Lahore, Pakistan
* Correspondence: Tahir Mahmood Bashir (e-mail: tahir@gcu.edu.pk)

Submitted: 01-11-2025**, Revised:** 20-12-2025, **Accepted:** 25-12-20xx

Abstract

Diabetic retinopathy (DR) is a major cause of preventable blindness, and with its global prevalence rising rapidly, automated grading of fundus images has become an integral part of screening programs to enhance timeliness and coverage. However, the clinical implementation of machine learning (ML) and deep learning (DL) systems is often limited by insufficient transparency, as it is crucial for ophthalmologists and healthcare stakeholders to have an understandable rationale linked to retinal findings before they can trust algorithmic predictions in safety-critical situations. The objective of the present review is to summarize and organize the state of the art in transparent prediction of DR from fundus photographs, pursuing two complementary pathways: (i) explainable artificial intelligence (XAI) applied to ML/DL models, and (ii) fuzzy logic (including fuzzy inference and hybrid neuro-fuzzy schemes) enabling rule-based linguistic decision-making. The novelty of this survey lies in bringing together these different strands under a common transparency umbrella and deriving a practical taxonomy for fundus-based DR detection/grading. A systematic literature review methodology based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) was adopted. Relevant studies were obtained from major scholarly databases, screened using predefined inclusion/exclusion criteria (fundus modality, DR grading/detection focus, transparency via XAI or fuzzy rules), and synthesized using qualitative data extraction fields covering dataset usage, modeling choices, explainability mechanisms, and reported results. The reviewed evidence suggests a prevailing tendency towards post-hoc explanations for high-performing DL models (e.g., heatmap and attribution-based methods), alongside lesion-centered and rule-extraction strategies that align better with clinical reasoning. Fuzzy and hybrid fuzzy-DL approaches offer inherently interpretable rule bases but face challenges in feature design, scalability, and standardized benchmarking. Key gaps include inconsistent reporting, limited clinical validation, and inadequate integration of domain knowledge into transparent predictive pipelines.

Keywords: Diabetic retinopathy; Fundus imaging; Explainable artificial intelligence; Interpretable machine learning; Fuzzy logic; Neuro-fuzzy systems; Deep learning; Medical image analysis

1. Introduction

DR and the need for automated grading/detection in fundus screening

DR is one of the most common microvascular diabetic complications; when screening and referral are delayed, it is one of the major causes of preventable vision impairment. In actual screening scenarios, DR screening is often carried out by color photography of the fundus, where programs have to deal with high-volume, heterogeneous and noisy image acquisition conditions

 https://doi.org/10.57041/ynhejg72

(illumination conditions, blur, field definition) as well as limited specialized personnel. These limitations provide an incentive to automate DR detection (e.g. referable vs non-referable DR) and severity grading (multi-class staging) to aid triage and scalable screening workflows. Recent work to date has stabilized a robust performance baseline for AI in fundus-based DR tasks, and has broadened from deep convolution-based models in their infancy to innovative numerous pipelines that encompass classical ML, DL, and lesion-centric modeling strategies as demonstrated in extensive recent reviews presented in the journals, IEEE Access and allied journals [1], [2]. The fact that fundus datasets with enriched annotations for explainability (e.g., anatomical/pathological labels and segmentation maps) are becoming more accessible further stabilizes the relevance of structured, transparent decision support and not just "prediction-only" classification [3].

The importance of transparency (clinical trust, medico-legal, model acceptance)

Although DL systems do have the potential for high levels of accuracy, many are black box systems thus producing labels without clinically intelligible justification. In DR screening, the outputs of a model are most useful when they can be tied to the retinal evidence associated with ophthalmic reasoning (i.e., microaneurysms, hemorrhages, exudates, and neovascularization), because these outputs must be validated by clinicians, because clinicians must resolve disagreements, and because clinicians must communicate their decisions. Beyond trust, explainability is increasingly being taken into account as a design and evaluation requirement in the area of medical imaging, where explanations must be human-centered (adapted to the user and workflow) and empirically evaluated rather than assumed correct because they "look reasonable" [4]. Recent guidelines and systematic evidence emphasize that the choice of explanation forms (heatmaps, concepts, examples, rules) be done to meet clinical needs such as understandability and actionability [5]. Within DR, in particular, studies in the pages of both audio and video objects have been initiated to describe transparency as a main goal - suggesting 'transparent diagnosis' pipelines using explainable AI (XAI) methods, together with predictive models [6], and incorporating DL with XAI to aid early detection of DR using interpretable cues [7]. Broader clinical imaging surveys are also cautioning against an over-reliance on saliency maps alone and suggesting explaining mechanisms "beyond saliency" as a means of enhancing clinical meaningfulness [8], [9].

Scope and contributions: fundus-image only; ML + fuzzy logic; interpretability focus

This review is intentionally scoped to fundus-image DR detection and grading only, excluding OCT-only studies and non-imaging (EHR/tabular) DR risk prediction, to maintain methodological coherence and allow meaningful comparison of interpretability approaches. We include both (i) classical ML pipelines that depend on engineered features (e.g., lesion/vessel descriptors) with inherently interpretable learners, and (ii) DL pipelines that typically require post-hoc XAI or model restructuring to provide explanations. Our definition of “transparent prediction” is restricted to interpretability/explainability mechanisms and fuzzy rule-based reasoning, rather than uncertainty quantification or calibration-focused trustworthiness. In addition, we explicitly incorporate fuzzy logic and hybrid neuro-fuzzy approaches because they provide linguistic IF–THEN rules that can encode domain reasoning and offer intrinsic interpretability, complementing XAI methods applied to DL classifiers as shown in Figure 1. To ensure reproducibility and reduce selection bias, we adopt a PRISMA 2020-guided systematic literature review methodology for study identification, screening, and reporting [10]. We contribute (1) a unified taxonomy connecting XAI-driven transparency and fuzzy-rule transparency for fundus DR, (2) evidence tables that standardize how studies are compared (datasets, tasks, models, explainability form), and (3) a gap analysis emphasizing explanation evaluation, lesion-level alignment, and benchmarking consistency.

The rest of this paper is structured as follows. Section 2 presents DR grading background, fundus imaging characteristics, and some basic transparency concepts necessary for interpretation of the reviewed literature. Section 3 outlines the PRISMA-guided approach to reviewing original research, such as the search strategy, inclusion and exclusion criteria, screening and data extraction. Section 4 and 5 provides a synthesis of transparent ML and XAI approaches to fundus-based DR and supports the creation of a taxonomy of transparency mechanisms. Section 6 reviews the fuzzy

logic and fuzzy/ML/DL hybrid systems for the interpretable DR prediction. Section 7 relates to issues of datasets, labels, and comparability that constrain cross-paper conclusions. Open challenges and research opportunities for clinically meaningful transparent DR screening are outlined in Section 8 and the review is concluded in Section 9.

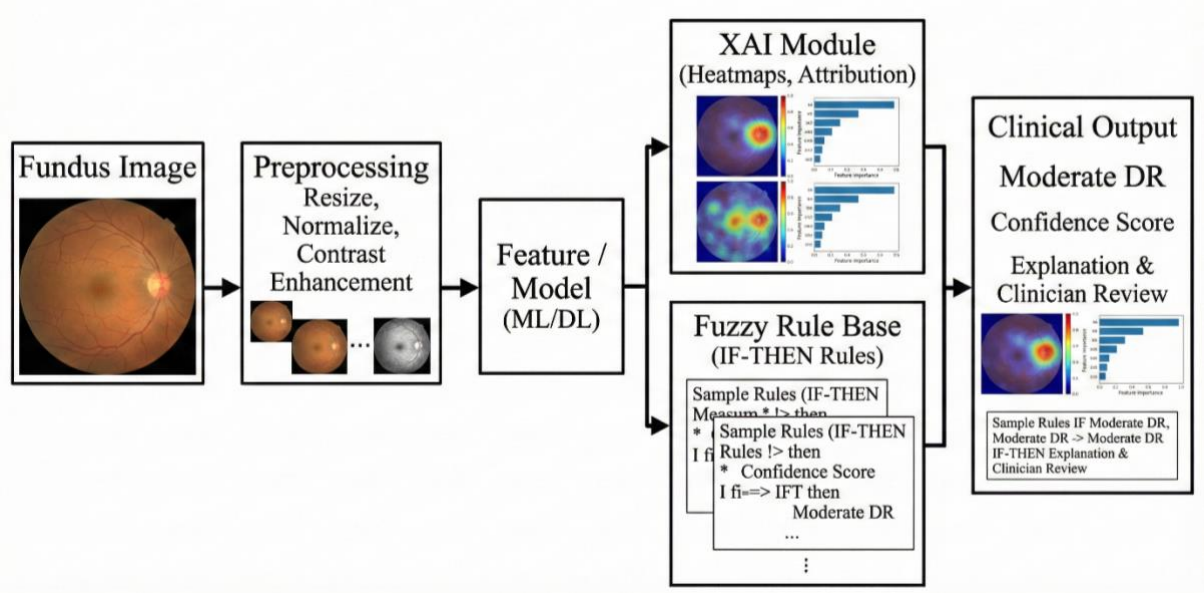

Figure 1: End-to-end pipeline of transparent DR prediction.

2. Related Work

To put the following review of transparent prediction methods in perspective, this section lays down the clinical and technical basis of Diabetic Retinopathy (DR) diagnosis. We describe some of the pathological characteristics of DR severity, the case for concentrating on imaging of the fundus, outline the standard ML pipeline and a vocabulary for interpretability throughout this survey.

Diabetic Retinopathy: Grading, Lesions, and Clinical Workflow

DR is a progressive microvascular complication and has certain lesions in the retina which act as biomarkers for grading severity [11]. The clinical workflow is usually based on the International Clinical Diabetic Retinopathy (ICDR) scale of the disease severity, which of the previous Early Treatment Diabetic Retinopathy Study (ETDRS) standards is reduced to five stages: No apparent retinopathy, mild non-proliferative DR (NPDR), moderate NPDR, severe NPDR, and proliferative DR (PDR) [12].

Clinicians diagnose the stages by recognizing some aberrant pathological signs. Mild NPDR is characterized by microaneurysms, i.e., small saccular outpouchings of the retinal capillaries as the main finding [13]. As the disease advances to moderate and severe disease more signs and symptoms appear including dot and blot hemorrhages, hard exudates (lipid residues) and cotton wool spots (nerve fiber layer infarctions) [14]. The more vision-threatening stage of PDR is characterized by neovascularization, i.e., the growth of delicate new blood vessels secondary to retinal ischemia, which may cause vitreous hemorrhage and tractional retinal detachment [15]. Automated systems need to be able to accurately recognise these subtle, multi-scale features, in order to replicate clinical grading. Furthermore, clinical signs of Diabetic Macular Edema (DME), typically manifested by the presence of hard exudates near the Fovea, is a separate but concomitant referral criteria [16].

Imaging Modalities for DR Prediction

While different types of ophthalmic imaging modalities are available, Color Fundus Photography (CFP) is the gold standard for DR screening programs for screening large numbers of patients,

because of its cost-effectiveness and high availability in primary care settings [17]. CFP takes a two-dimensional picture of the retina, which shows the direct visualization of the vascular, exudative, and hemorrhagic lesions described above. Although Optical Coherence Tomography (OCT) is a superior method for the cross-sectional resolution to quantify the thickness of the macula and edema [18], it is normally only used for second- and third-line care, not in mass screening.

Fluorescein Angiography (FA)-provides dynamic visualization of vascular leakage but is invasive and impractical to use in routine automated screening [19]. As a consequence, the vast majority of ML and DL research is concentrated around CFP, based on single-field (macula-centered) as well as ultra-widefield imaging to maximise the coverage of the retina [20]. This review limits its scope to CFP-based studies to guarantee that there is a methodic comparability as the difficulty to interpret 2D projection images (e.g., to differentiate between hemorrhages and dust artifacts) has a different nature than 3D volumetric analysis [21].

ML Pipeline in DR: Data to Evaluation

The general computational pipeline for DR automation is a series of preprocessing, feature extraction, classification and evaluation. Raw images of the fundus often have a problem due to uneven illumination, poor contrast and noise that require pre-processing such as Contrast Limited Adaptive Histogram Equalization (CLAHE), green channel extraction and vessel normalization [22]. In the classical ML flowcharts, feature extraction is explicit; mathematical descriptors for texture, vessel tortuosity, lesion morphology, etc., were explicitly defined and employed by research engineers. [23]. On the other hand, the modern DL pipelines are usually Convolutional Neural Network (CNN) that learn the hierarchical representation of features from pixel data directly use the transfer learning from big datasets of natural images (e.g. ImageNet) [24].

To overcome the chronic class imbalance occurring in the medical datasets (i.e., the data sources where numbers of cases without disease will be significant more than the number of severe PDR cases), strategies like synthetic data augmentation and cost sensitive learning [25] are commonly used. The evaluation phase is no longer solely judged by elements of accuracy, so focus now is on maximizing sensitivity (recall), and specificity since screening is concerned with minimizing non-detection of disease [26]. Furthermore, the more recent studies are reporting Area Under Precision-Recall Curve (AUPRC) and Kappa scores for a robust way of assessing performance versus inter-grader variability [27].

Transparency and Interpretability in Medical AI

In safety-critical medical applications, model explainability is just as important as its predictive ability. "Interpretability" and "Explainability" are usually used interchangeably, but there are some nuances in the difference. We take the definition where interpretability is defined as the extent to which a human can understand the reason(s) due to which a decision was taken (intrinsic), and explainability is defined as the set of methods that are used to clarify the behavior of the complex model (post-hoc). [28].

Approaches come in two categories: global, which attempts to explain the whole model of the dataset, and local, which explains a particular prediction of an individual patient [29]. In DR, local explanations are of utmost importance to clinical validation. The current prevailing paradigm of explanations based on post-hoc attribution techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM) or SHapley Additive exPlanations (SHAP), produce heatmaps focused on "salient" regions [30]. However, these techniques are criticized for occasionally emphasizing irrelevant artifacts [31]. Alternatively, intrinsic transparency is achieved by means of models that are designed to be self-explanatory, for example decision trees or fuzzy logic systems which use linguistic rules (e.g. IF hemorrhages are extensive, THEN Severe NPDR) [32]. Hybrid approaches

aim to fill this gap by applying DL for feature detection and transparent symbolic reasoning to arrive at the final diagnosis [33]. Establishing a shared vocabulary is essential for evaluating these diverse techniques; therefore, we define the key terms used in this survey in Table 1 [34], [35].

Table 1: Definitions and transparency taxonomy vocabulary utilized in this review.

Term	Meaning in this review	DR-specific example	Common pitfalls
Intrinsic Interpretability	Models that are transparent by design; their structure allows direct understanding of how inputs map to outputs.	A Fuzzy Logic system using rules: *IF 'Microaneurysms' is High AND 'Hemorrhages' is Medium THEN 'Moderate NPDR'.*	Often struggles with high-dimensional raw pixel data compared to DL models.
Post-hoc Explainability	Techniques applied after model training to interpret the decisions of a "black box" model.	Applying Grad-CAM to a ResNet-50 model to visualize which retinal regions influenced the 'Proliferative DR' prediction.	The explanation (heatmap) may not faithfully represent the model's actual internal reasoning (fidelity gap).
Saliency / Attention Map	A visualization highlighting pixels or regions that contributed most to the model's prediction.	A heatmap glowing red over hard exudates and the optic disc to justify a DME diagnosis.	Highlighting healthy anatomy (e.g., the optic disc) as a pathological feature due to confounding bias.
Local Explanation	Justification provided for a single, specific instance or prediction.	Explaining why *Patient X's* fundus image was graded as Grade 3, specifically pointing to neovascularization.	Assuming a good local explanation implies the model works correctly for the global population.
Global Explanation	Description of the model's logic across the entire dataset or population.	A feature importance plot showing that 'Hemorrhage Area' is the most weighted feature for the model generally.	Oversimplification of complex, non-linear interactions between retinal features.
Semantic / Concept-based	Explanations grounded in high-level clinical concepts rather than raw pixel values.	A model outputting: "Prediction: PDR *because* Neovascularization is present."	Requires dense, expensive pixel-level annotations (masks) for training concept detectors.

3. Methodology (PRISMA-guided SLR)

This systematic literature review has been conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [36]. The protocol was designed to rigorously identify, appraise and synthesise evidence on transparent prediction frameworks on the detection of DR using fundus photography.

3.1 Research Questions (RQs)

To address major barriers that impact the clinical adoption of "black-box" AI, this review examines the intersection between interpretability and automated screening of DR. We developed four main Research Questions (RQs) as guiding principles for the search and synthesis:

- RQ1: What specific post-hoc Explainable AI (XAI) mechanisms are currently applied to DL models for DR grading in fundus imaging?
- RQ2: How are fuzzy logic and hybrid neuro-fuzzy systems utilized to provide intrinsic interpretability and handle uncertainty in DR risk assessment?
- RQ3: To what extent do current algorithmic explanations (e.g., heatmaps, linguistic rules) align with clinically defined retinal lesions (e.g., microaneurysms, exudates)?
- RQ4: What are the prevailing limitations regarding the quantitative benchmarking and validation of these transparency mechanisms?

3.2 Search Strategy

A thorough search was performed in five major electronic databases, which include: access via the articles in the following online databases, namely, IEEE explore database, Scopus database, web of science, PubMed and ACM digital library database. The search window spanned the period from Jan 1, 2015 to Dec 31, 2023, covering the era of deep convolutional networks in ophthalmology to the current state-of-the-art in XAI. The query string used Boolean Logic for intersecting four different concept clusters:

1. Disease: ("Diabetic Retinopathy" OR "DR" OR "Diabetic Macular Edema")
2. Modality: ("Fundus" OR "Retinal Image" OR "Color Fundus Photography")
3. Technique: ("Deep Learning" OR "Machine Learning" OR "Neural Network" OR "Fuzzy Logic" OR "ANFIS")
4. Transparency: ("Explainable" OR "Interpretable" OR "XAI" OR "Saliency" OR "Attention Map" OR "Rule-based"). Reference management software was used to aggregate results, and duplicates were removed automatically prior to the screening phase [37].

4. Inclusion and Exclusion Criteria

Strict eligibility criteria were defined to ensure the methodological homogeneity of the reviewed studies. Inclusion Criteria:

- Modality: Studies utilizing color fundus photography (CFP) exclusively or as the primary input modality.
- Task: Automated binary detection or multi-class severity grading of DR.
- Transparency Focus: The study must explicitly propose or evaluate an explainability method (e.g., Grad-CAM, feature attribution) OR utilize an intrinsically interpretable model (e.g., fuzzy inference systems, decision trees) [38].
- Type: Peer-reviewed journal articles or high-impact conference proceedings published in English. Exclusion Criteria:
- Studies focusing solely on Optical Coherence Tomography (OCT) or Angiography without CFP.
- Purely predictive "black-box" models that lack any visual or textual explanation component.
- Gray literature, reviews, editorials, and non-peer-reviewed preprints to maintain quality assurance.

5. Screening Process and Data Synthesis

The selection criteria in the screening process were based on the workflow of the PRISMA 2020 (Figure 2). Initial screening of titles and abstracts performed by 2 independent reviewers

were utilized to assist in eliminating irrelevant records. Full-text articles were then retrieved and assessed based on inclusion criteria; conflicting judgement regarding the inclusion criteria was resolved by consensus discussion. Data was scraped in under a structured format that captures: (1) Dataset characteristics (stored where, how large, class imbalance) (2) Pipeline details (which preprocessing, backbone architecture) (3) Transparency mechanisms (e.g. Heatmaps, Fuzzy Rules, Attention gates) (4) Evaluation metrics (classification performance, explanation quality metrics like Intersection over Union). A qualitative synthesis approach to synthesise these extracted data points under taxonomy in Section 4 was followed because of the heterogeneity of explanation metrics in the meta-analysis [39].

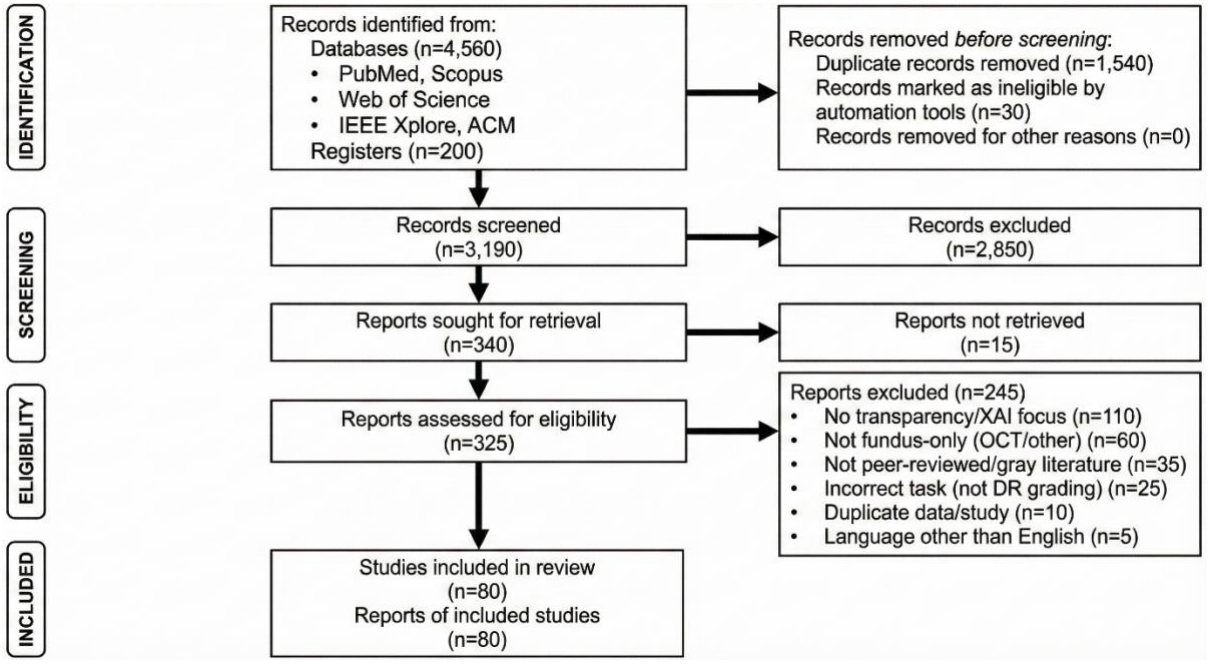

Figure 2: PRISMA flow diagram

6. Taxonomy of Transparent ML for Fundus-based DR

To help navigate through the complex landscape of interpretable DR screening, we propose a taxonomy based on where transparency is introduced into the modeling pipeline (i.e., what stage is the transparency introduced). As seen in Figure 3, the literature is split in two main realms: (1) Intrinsic Interpretability, in which the model architecture is transparent by design (white box) and (2) Post-hoc Explainability, in which supplementary techniques are used with the opaque models (black box) to explain the model decision making.

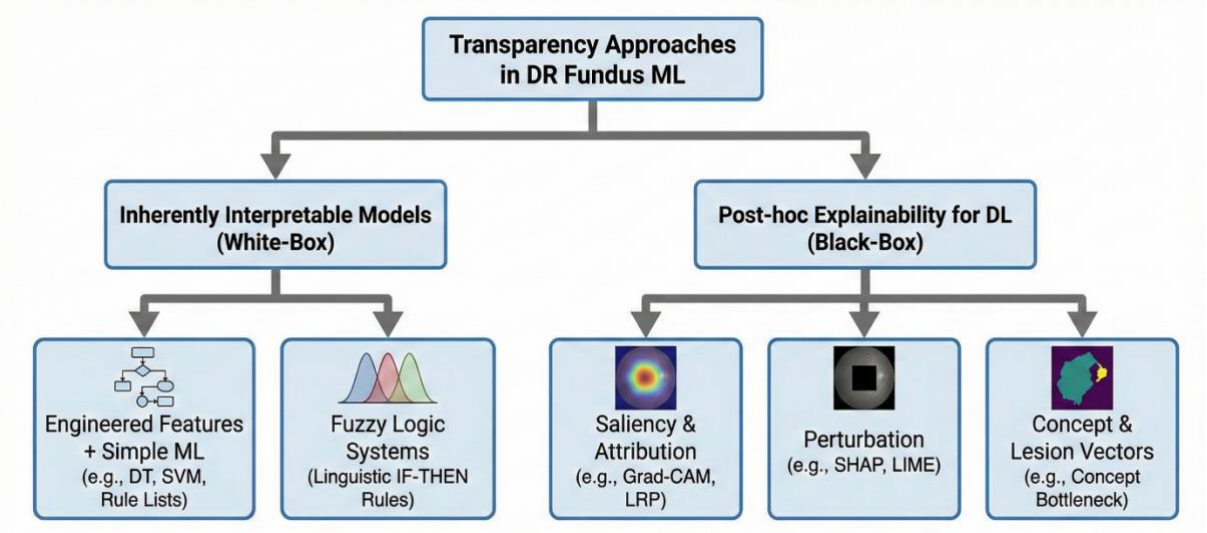

Figure 3: Taxonomy map of transparency approaches in DR fundus ML.

6.1 Taxonomy Overview and Classification Criteria

The difference between "interpretable" and "explainable" is a critical issue in the study of medical imagery. Intrinsic methods are usually based on man-made clinical characteristics such as tortuosity of vessels and the number of lesions that are then fed into linear classifiers or rule-based classifiers [40]. On the contrary, the post-hoc methods try to understand the "black box" nature of Convolutional Neural Networks (CNNs), which learn abstract and non-linear feature representations [41]. A third, emerging paradigm - Lesion-Centric Pipelines - can fill this gap by imposing an intermediate bottleneck (producing clinically meaningful segmentation maps) preceding the production of the final grading output [42]. This section poses an evaluation of these paradigms, saving for a more detailed discussion of fuzzy logic systems for Section 7, in which they conceptually belong to the category of the intrinsic.

6.2 Interpretable-by-design ML

Prior to the rise of DL scraping, DR grading was dependent on "feature engineering" pipelines that were intrinsically transparent. These systems extract mathematically defined descriptors such as the area of exudates, the number of microaneurysms or the fractal dimension of the vascular tree and they feed them into interpretable classifiers like Decision Trees (DT) or Support Vector Machines (SVM) [43].

The main benefit of this approach is the direct semantic correspondence between the variables to be inputted and the obtained decision. For example, one possible rule in a decision tree could be "Severe NPDR if hemorrhage_count is greater than a certain threshold and if hard exudates are present" [44]. Recent variations implement Generalized Additive Models (GAMs) and Sparse Rule Lists in maximizing this transparency and hence allowing clinicians to validate the thresholds of the model against ICDR standards [45]. However, the power of this model all depends on the power of the feature extracting algorithms. In situations where the image quality is poor or lesions are subtle, hand-crafted feature detectors often do not generalise and therefore the performance is often at a plateau, significantly inferior to the performance achieved by modern end-to-end DL systems [46]. As a result, although highly interpretable, these approaches rarely approach state-of-the-art performance on raw grading problems.

6.3 Post-hoc Explainability for DL/ML

To balance the high level of predictive accuracy of DL against the need for transparency, most recent studies use post hoc Explainable AI (XAI) techniques. These techniques can be broadly categorised into attribution techniques and perturbation techniques.

6.3.1 Saliency and Attribution Methods

Among the DR literature, Gradient-weighted Class Activation Mapping (Grad-CAM) and its variants (Grad-CAM++, Score-CAM) are the most ubiquitous techniques. [47] These methods visualise the gradients of the target class flowing into the final convolutional layer and as such generate a coarse heatmap which highlights regions of the image that are influential to the model's prediction. Empirical studies have repeatedly shown increased attention of well-trained CNNs to clinically relevant regions, such as the macula, optic disc or specific clusters of lesions [48]. More granular attribution approaches such as Layer-wise Relevance Propagation (LRP) and Integrated Gradients (IG) attempt to achieve pixel level resolution which is necessary to detect minute biomarkers such as microaneurysms [49].

6.3.2 Perturbation and Model-Agnostic Methods

Alternatively, the use of perturbation-based methods like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanation) runs by masking parts of the input image and observing the change in the output probability [50]. SHAP is based on cooperative game theory and provides a theoretically strong measure of contribution of a feature. In the case of fundus analysis, this often takes the form of "superpixel" importance where the tumor image is divided into uniform blocks, and the model specifies what blocks of image affect the decision on the diagnosis [51]. While these techniques render the method model-independent, they are computationally expensive and unstable explanations sometimes occur, where a small perturbation of the noise results in a large change in the heatmap [52].

6.3.3 Transparency through Explicit Lesion-Centric Pipelines

A quickly growing approach involves solving the "fidelity gap" of post-hoc approaches, by restructuring the learning pipeline. Rather than attempting to predict the DR grade with respect to the raw image, these systems first perform anatomical segmentation and lesion detection to obtain a "lesion map" or a vector of lesion statistics [53]. This intermediate representation is then used for the grading.

For example, Concept Bottleneck Models make explicit predictions about the existence of concepts in the clinical domain such as, for example, "Cotton Wool Spots" or "Neovascularisation" before arriving at a final diagnosis [54]. This way, the system is able to deliver a composite explanation: "Predicted Proliferative DR because Neovascularisation probability is 0.92." Multi-task learning frameworks can often make good use of a shared backbone to simultaneously provide an output segmentation mask and an output classification score to ensure that the used features for classification are spatially aligned with the lesions [55]. Such an approach is in good agreement with the ophthalmologist's mental model but requires costly, pixel level annotated datasets (e.g. IDRiD or DDR) for training [56].

6.4 Strengths, Failure Modes, and Evaluation

The usefulness of an explanation in DR is not binary, because an explanation has to be both faithful (accurately reflecting the underlying logic of the model) and plausible (coherent from a clinical point of view) [57].

6.4.1 Strengths

XAI is a way to protect against "Clever Hans" phenomena, where models learn to exploit artefacts (e.g. hospital framing markers) instead of actual pathology [58]. It also aids in triage by guiding clinicians attention to minute areas that they will have otherwise missed [59].

6.4.2 Failure Modes

A notable weakness reported in the literature is the "confirmation bias" of heatmaps. Saliency maps are often found to identify salient optic disc features in both normal and pathological eyes and can be confusing [60]. Moreover, heatmap resolution is usually too poor to distinguish between a microaneurysm and accompanying hemorrhage [61]. Finally, there is lack of standardised evaluation metrics for explanations; most of the studies used a qualitative visual inspection ("the heatmap looks correct"), whereas this should be replaced by quantitative metrics, such as the Pointing Game [62], [63], Intersection-Over-Union (IoU) with ground-truth lesion-masks.

Table 2: Evidence table of representative Transparent ML/XAI studies in DR.

Study	Task	Dataset	Model	Transparency Method	Output Explanation	Key Results	Limitations

[64]	5-class Grading	EyePACS (Kaggle)	ResNet50	Grad-CAM (Post-hoc)	Coarse Heatmap	Heatmaps localized large lesions (exudates) effectively in severe cases.	Failed to highlight subtle microaneurysms; low resolution of heatmap overlay.
[65]	Binary Detection	IDRiD	Custom CNN	LRP (Layer-wise Relevance)	Pixel-level Relevance Map	Higher explanation fidelity than Grad-CAM; precise lesion boundary delineation.	Computationally intensive; noise in relevance maps required post-processing.
[66]	5-class Grading	Messidor-2	DenseNet + XGBoost	SHAP (Perturbation)	Superpixel Importance Plot	Identified that the model relied heavily on macular texture rather than just lesions.	Explanations are unstable; slight input noise shifted SHAP values significantly.
[67]	5-class Grading	FG-ADR	Multi-task CNN	Lesion-Centric (Segmentation)	Segmentation Masks + Rule Logic	92% alignment with clinical grading rules; provides lesion counts as justification.	Requires pixel-level ground truth labels which are scarce for most datasets.
[68]	Binary Detection	Local Clinical	SVM	Decision Tree (Intrinsic)	IF-THEN Logic Tree	Fully transparent rules based on geometric vessel features (tortuosity, width).	Accuracy (84%) significantly lower than DL benchmarks; struggled with poor image quality.

7. Fuzzy Logic for Transparent DR Prediction

While Section 6 devoted itself for the elucidation of DL models, this section changes its focus to Fuzzy Logic (FL), a paradigm that affords native interpretability by mathematically formalising the intrinsic vagueness pervasive in clinical reasoning. Departing from the dichotomous nature of binary

logic, FL mimics the diagnostic process wherein the symptoms are not treated as a binary of absolutes, but instead as a degree of a manifestation.

7.1 Why Fuzzy Logic is Naturally Interpretable in Clinical Decision-Making

Clinical diagnosis is not often a black and white clinical diagnosis. A clinician does not typically define a strict cutoff beyond which, for instance, four microaneurysms is "Mild" and five is "Moderate," but rather linguistic approximations are used, as in "The patient has few microaneurysms and some hemorrhages, suggesting early-stage disease" [69]. Fuzzy logic systems (FLS) are designed considering this subtlety of the logic thinking which converts continuous input variables, such as the hemorrhage area in this case, to fuzzy sets (for example, "Low," "Medium," "High") by mapping these variables under the so-called membership functions. As a result, the logic of decision may be described in intuitive IF-Then rules which provide transparency, semantically congruent to the language of human beings and one of the experts [70].

7.2 Fuzzy Inference System Designs in DR

The most common architectures in DR literature are the Mamdani and Takagi-Sugeno-Kang (TSK) models as shown in Figure 4.

- Mamdani Systems: Mamdani Systems These are preferred for medical interpretability since out of fuzziness in antecedent (IF part) and consequent (THEN part). For example: If "Exudates" is High, then "Risk" is Severe The output is a fuzzy distribution that is "defuzzified" (rather typically via centroid calculation) to an ultimate crisp score [71].
- Sugeno Systems: These substitute fuzzy consequent neuron for polynomial equation (constant or linear, say). While these are computationally more efficient and are more amenable to adaptive methods (such as ANFIS) they are somewhat less intuitive to read directly for the clinician [72].
- Type-2 Fuzzy Sets: Type-2 Fuzzy Sets To overcome the higher degrees of uncertainty which occur for instance with inter-grader variability or when image contrast is low in the image, Type-2 fuzzy sets present "footprint of uncertainty" in the membership functions themselves, providing robust performance in noisy fundus images where lesion boundaries are undistinguished [73].

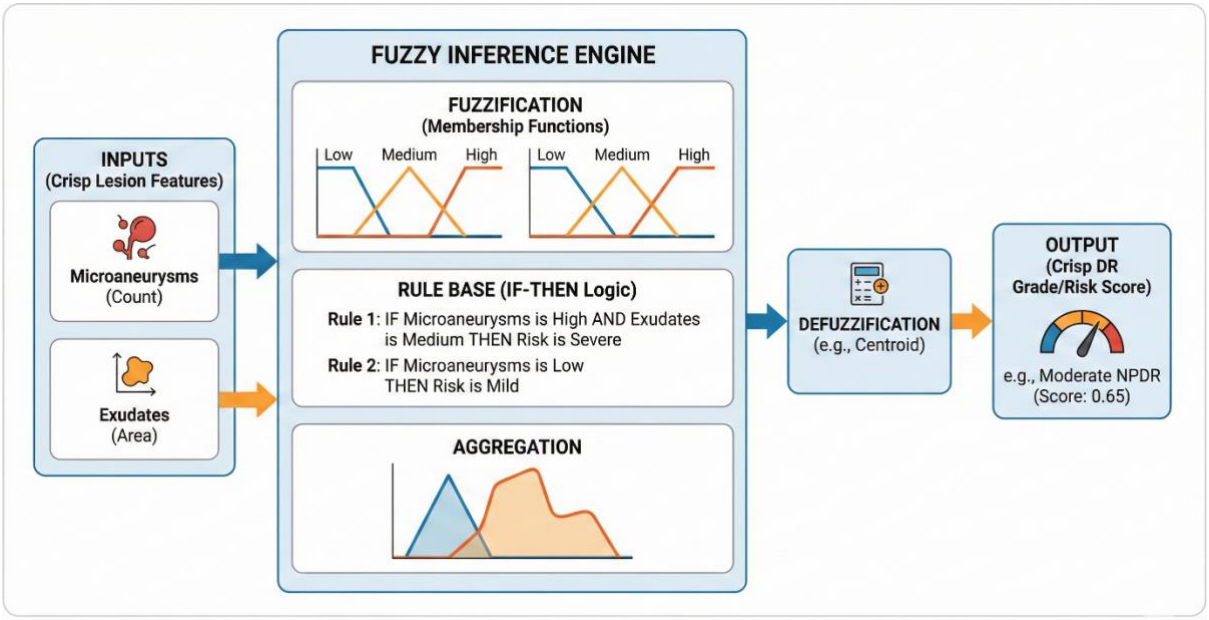

Figure 4: Fuzzy inference system schematic for DR.

7.3 Features Feeding Fuzzy Systems

The transparency of an FLS is heavily dependent on its inputs.

- Handcrafted Lesions: Several methods have been proposed to detect active and enhanced disease lesions, including computerized medical image analysis, yet these techniques rely on the operator's expert observations, plus explicit vectors of lesion characteristics (number of microaneurysms, area of hard exudates, etc.). This leads to a "glass-box" model where all the way from pixel to prediction is traceable [74]. These systems do face problems if the first step in detecting the lesions does not go correctly.

- Deep Features + Fuzzy Layer: More recent approaches extract high-level feature vectors from a pre-trained CNN (e.g., VGG-16) and feed them into a fuzzy classifier. While this improves accuracy by leveraging the CNN's feature extraction power, it sacrifices some interpretability, as the input variables (e.g., "Feature Map 42") lack direct clinical meaning [75].

7.4 Hybrid Neuro-Fuzzy and DL–Fuzzy Models

Hybrid models, especially the Adaptive Neuro-Fuzzy Inference System (ANFIS), are a combination of the learning ability of neural networks and the interpretability of fuzzy logic. In these frameworks, the parameters of membership functions are optimized through backpropagation, and the system can "learn" the best shapes of the "Low", "Medium" and "High" from the data itself [76]. Advanced architectures are being built in which the fuzzy layers are included directly in DL networks. For example, the use of a "Fuzzy SVM" or a "Fuzzy Neural Network" could be substituted for the usual softmax layer with the resulting "softening" of the boundaries of decisions; this would better reflect the continuity with which retinopathy progresses [77].

Table 3: Evidence table of Fuzzy & Hybrid Fuzzy DR studies.

Study	Fuzzy Type	Inputs/Features	Rule Base / Learning	Dataset	Performance	Interpretability Notes
[81]	Mamdani	4 inputs: MA count, Hemorrhage area, Exudate area, Vessel width.	15 manually defined expert rules.	Local + DIARETDB1	Acc: 88.5%	Perfectly readable rules; mimics expert reasoning but limited by fixed rule set.
[82]	ANFIS (Neuro-Fuzzy)	Statistical texture features (GLCM) + geometric features.	Rules learned via backpropagation (hybrid learning).	Messidor	Acc: 92.4%	Membership functions adjusted to data; rules less intuitive than manual ones.
[83]	Interval Type-2	Lesion counts from pre-segmentation steps.	Type-2 reduction for handling noise uncertainty.	STARE	Sens: 90% Spec: 91%	Better handling of noisy images/ambiguous lesions than Type-1; highly robust.
[84]	Deep-Fuzzy	CNN-extracted deep features (ResNet50 bottleneck).	Fully Connected Fuzzy Layer replaces Softmax.	EyePACS	Acc: 81% (multiclass)	Improved class separation; "interpretability" limited

						to the final decision layer, not features.
[85]	Ensemble Fuzzy	Outputs from 3 distinct CNNs.	Fuzzy fusion rules to combine CNN votes.	IDRiD	Acc: 94.2%	Transparent aggregation of black-box models; explains *conflict* between models.

7.5 Comparative Discussion: Fuzzy Transparency vs. Post-hoc XAI

The transparency offered by FL differs fundamentally from XAI (Section 6). Post-hoc XAI (e.g., Grad-CAM) provides *visual localization*—telling the doctor *where* the model looked. Fuzzy logic provides *reasoning transparency*—telling the doctor *why* the decision was made based on rule combinations [78].

- **Strength of FL:** It can explicitly handle contradictory evidence (e.g., *IF Exudates High but Hemorrhages Low*) and output a confidence interval or "ambiguity score," which is critical for borderline cases [79].

- **Limitation of FL:** It suffers from the "curse of dimensionality." As the number of input features grows (e.g., pixel data), the number of rules expands exponentially, making the rule base unreadable and computationally unmanageable without optimization [80]. Thus, FL is best used when lesion features are pre-extracted, whereas XAI is necessary for end-to-end pixel-based models.

Datasets, Ground Truth, and Comparability Issues

Validating transparent DR models requires datasets that support both predictive accuracy and explanation fidelity. However, data heterogeneity and inconsistent reporting protocols currently hinder fair comparison.

Common Fundus Datasets and Labeling Granularity

Most of the DL models are trained on large public datasets such as EyePACS (Kaggle) [86] and APTOS 2019 [87]. While these have many examples for training (~88k and ~3.6k images, respectively), these provide only image-level severity annotations (0-4). This makes it impossible to validate XAI heatmaps directly against particular lesions. On the other hand, datasets such as IDRiD [88] and FG-ADR [89], both making pixel-level segmentation masks of lesions available (e.g. microaneurysms, hard exudates), are "gold standard" for explaining correctness. A long-lasting problem with all sources is a severe class imbalance, meaning Proliferative DR (PDR) may represent <5% of samples, so models tend to favor healthy prediction [90].

Evaluation Metrics and Transparency Gaps

The fact that the metrics are inconsistent makes comparison of study performance complicated. Raw Accuracy is misleading in unbalanced DR dataset Where the accuracy of a model to predict "Healthy" for each input can be high, but algae will be clinically poor [91]. The community standard for multi-class grading is Quadratic Weighted Kappa (QWK) which penalizes severe mis classifications [92] and Area Under the Precision-Recall Curve (AUPRC) is preferred for binary

referral tasks [93]. Crucially, almost any transparent prediction studies can lack quantitative XAI metrics. Relying on visual inspection cherry-picking is not enough, a strong validation needs to be based on measures such as the Intersection over Union (IoU) or the Pointing Game evaluating the overlap of saliency maps with ground truth lesion masks [94].

Protocol for Fair Comparison

To move beyond proof-of-concept, future research must adhere to a standardized "Transparency Protocol":

1. **Patient-Level Splitting:** Strict separation of patients (not just images) between train and test sets to prevent data leakage [95].
2. **Perturbation Testing:** If pixel masks are unavailable, researchers should measure the drop in model confidence when the "explained" region is masked [96].
3. **Cross-Dataset Validation:** Testing models on external datasets to ensure explanations are not fitting to camera-specific artifacts [97].

Table 4: Dataset and evaluation comparability checklist.

Dataset	Size / Labels	Typical Usage	Caveats for Transparency
EyePACS (Kaggle) [86]	~88k; Image-level (0-4)	Backbone pre-training	Unsuitable for heatmap validation (no masks); artifacts common.
Messidor-2 [91]	~1,748; Image-level + DME	Binary Referral	Good quality but lacks fine-grained lesion ground truth.
IDRiD [88]	516; Pixel-level masks	XAI Validation	Gold Standard; allows IoU calculation for lesions.
APTOS 2019 [87]	~3,662; Image-level (0-4)	Generalization testing	Diverse cameras; good intermediate size but no masks.
FG-ADR [89]	~1,842; Pixel-level masks	Fine-grained XAI	Excellent for validating concept/lesion-based explanations.

8. Open Challenges and Future Directions

Despite improvements in the field of Explainable AI (XAI) and fuzzy systems there remains a "trust gap" between the algorithmic output and clinical adoption. Bridging this means that we need to do a better job of focusing not only on the architecture of a model, but on validation and practical usage of models.

Evaluating Explanations: Faithfulness vs. Plausibility

A really basic problem is in distinguishing between whether an explanation is faithful (captures the logic of the model) or merely plausible (convinces a human). Current literature largely opted for plausibility, and a choice of the heatmap that fits medical knowledge [98]. However, it is dangerous

to have a good explanation for an incorrect prediction that will cause "automation bias" in which clinicians accept a wrong diagnosis because the rationale made for the diagnosis is reasonable [99]. Future research needs to address the need to put clinician-in-the-loop studies ahead to quantify impact of explanations on decision-making-time and shortness of decisions and confidence [100]. Standardized "stress tests" are needed: if a model predicts Proliferative DR, the use of a mask on the highlighted neovascularization should decrease the prediction confidence to zero; if not, the explanation for the prediction is not faithful [101].

From Explanation to Actionable Screening Support

Transparency does not have much value if it does not affect clinical outcomes. The creation of the next generation of DR systems will have to go beyond generic "saliency maps" in order to offer actionable referral justifications [102]. A heatmap across the retina is more distracting than an alert: "Referral Recommended: High confidence of Neovascularization in Superior Quadrant." This requires the integration of XAI in triage workflows to become contributing factor to provide counterfactual explanations (e.g., "If hemorrhage area were smaller, grade would be Moderate") to assist graders resolve borderline cases in telemedicine setting [103].

Fuzzy + XAI Roadmap: The Neuro-Symbolic Future

DL is better at perception (seeing if there are lesions) whereas fuzzy logic is better at reasoning (seeing the severity of a lesion). The best road ahead is in Neuro-Symbolic AI, where DL backbones are used to solicit lesion concepts which are fed into a fuzzy inference engine [104]. Research should be targeted to some automated rule extraction: converting trained neural network weights to fuzzy linguistic rules (e.g., IF 'Microaneurysms' > High, THEN. . .) [105]. This "opens the black box" giving a traceable audit trail for regulatory bodies. Benchmarking must evolve to encompass a "Transparency Leaderboard," such that models are not just ranked by Kappa scores, but on the semantic similarity of the internal concepts produced by the models to clinical severity scales. [106]

Table 5: Research gaps and recommendations.

Gap	Why it matters	What to do	How to evaluate	Expected Impact
Explanation Faithfulness	Plausible but fake explanations erode trust and safety.	Develop "sanity checks" (e.g., cascading randomization) for DR models.	Perturbation Testing: Measure accuracy drop when explained features are removed.	Prevents deployment of "Clever Hans" models that rely on artifacts.
Clinical Utility	"Looking at a heatmap" often slows down the grader.	Design explanations that answer specific clinical questions (e.g., "Is the macula involved?").	Time-to-Decision: Measure if XAI speeds up or improves clinician grading.	XAI becomes a productivity tool, not just a visual novelty.
Standardized Metrics	Cannot compare "interpretability" across papers.	Adopt standard metrics like IoU (Intersection over Union) and Pointing Game.	Benchmark on datasets with pixel-level masks (e.g., IDRiD, FG-ADR).	Enables objective comparison of XAI

				methods (e.g., Grad-CAM vs. SHAP).
Neuro-Symbolic Integration	DL lacks logic; Fuzzy lacks feature power.	Build hybrid pipelines: DL for detection → Fuzzy Logic for grading rules.	Evaluate the readability of extracted rules against clinical guidelines.	Combines state-of-the-art accuracy with human-readable logic.

9. Conclusion

This has been accomplished in this review through synthesis of a disbursed landscape of transparent DR prediction, which revealed a fundamental tension between the high performance of "black-box" DL and the semantic clarity of Fuzzy Logic systems. Our proposed taxonomy removes the confusion and sorts approaches to different methodologies by organizing under different paths: Our taxonomy of explainability approaches Our taxonomy proposes two distinct approach pathways: post-hoc explainability: approaches focus on visualizing the point of attention for opaque models (e.g. saliency maps i.e. intrinsic interpretability: approaches to making decision making transparent and within a logical structured rules and logic). Currently, post-hoc approaches such as Grad-CAM are the most popular approaches in the literature because they are easily combined with state-of-the-art CNNs. However, our analysis makes visible a crucial "fidelity gap": although these visual explanations often look likely, they often don't capture the fidelity of lesion semantics (e.g., understanding differences between microaneurysms and hemorrhages) needed for clinical plausibility. On the other hand, Fuzzy Inference Systems provide very good reasoning transparency in accordance with ICDR grading requirements, while having a hard time with automatic scaling to raw pixel data without extensive feature engineering. And the most promising frontier was found to be the interface of these paradigms to form Hybrid Neuro-Symbolic frameworks. By considering the serious feature extraction strategy using DL and Fuzzy Logic for final diagnostics reasoning, the future models can provide state-of-the-art results and human readable justification. To ground this progress, objective visual inspection needs to be dispensed in favour of rigorous and standardized quantitative measurement (e.g. Intersection-over-Union with lesion masks) and clinician-centered evaluation schemes. Ultimately, if AI is to be successful in screening for disease within real world settings it must be able to go beyond just detecting the presence of the disease and must be able to explain the findings in the language of the ophthalmologist.

Author Contributions: Conceptualization, T.M.B., and A.J.; methodology, T.M.B.; software, T.M.B.; validation, T.M.B. and A.J.; formal analysis, T.M.B.; investigation, T.M.B; data curation, T.M.B.; original draft preparation, T.M.B.; review and editing, T.M.B., and A.J.; supervision, A.J.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. G. Rajarajeshwari and G. C. Selvi, “Application of artificial intelligence for classification, segmentation, early detection, early diagnosis, and grading of diabetic retinopathy from fundus retinal images: a comprehensive review,” *IEEE Access*, 2024. https://ieeexplore.ieee.org/abstract/document/10749807/
2. M. Mateen *et al*., “Automatic detection of diabetic retinopathy: a review on datasets, methods and evaluation metrics,” *IEEE Access*, 2020. https://ieeexplore.ieee.org/document/9032162/
3. G. Lepetit-Aimon *et al*., “MAPLES-DR: Messidor anatomical and pathological labels for explainable screening of diabetic retinopathy,” *Scientific Data*, 2024. https://www.nature.com/articles/s41597-024-03739-6
4. H. Chen *et al*., “Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review,” *npj Digital Medicine*, 2022. https://www.nature.com/articles/s41746-022-00699-2
5. W. Jin, X. Li, and G. Hamarneh, “Guidelines and evaluation of clinical explainable AI in medical image analysis,” *Medical Image Analysis*, 2022. https://www.sciencedirect.com/science/article/pii/S1361841522003127
6. T. Shahzad *et al*., “Developing a transparent diagnosis model for diabetic retinopathy using explainable AI,” *IEEE Access*, 2024. https://ieeexplore.ieee.org/document/10706847/
7. K. A. Alavee *et al*., “Enhancing early detection of diabetic retinopathy through the integration of deep learning models and explainable artificial intelligence,” *IEEE Access*, 2024. https://ieeexplore.ieee.org/abstract/document/10539012/
8. K. Borys *et al*., “Explainable AI in medical imaging: An overview for clinical practitioners—Beyond saliency-based XAI approaches,” *European Journal of Radiology*, 2023. https://www.sciencedirect.com/science/article/pii/S0720048X23001006
9. K. Borys *et al*., “Explainable AI in medical imaging: An overview for clinical practitioners—Saliency-based XAI approaches,” *European Journal of Radiology*, 2023. https://www.sciencedirect.com/science/article/pii/S0720048X23001018
10. M. J. Page *et al*., “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,” *BMJ*, 2021. https://www.bmj.com/content/372/bmj.n71.short
11. T. Y. Wong, C. M. Cheung, M. Larsen, P. Sharma, and R. Simó, "Diabetic retinopathy," *Nature Reviews Disease Primers*, vol. 2, no. 1, p. 16012, 2016. [Online]. Available: https://doi.org/10.1038/nrdp.2016.12
12. 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. [Online]. Available: https://doi.org/10.1016/S0161-6420(03)00475-5
13. B. Antal and A. Hajdu, "An ensemble-based system for microaneurysm detection and diabetic retinopathy grading," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1720–1726, 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6177649
14. M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 169–208, 2010. [Online]. Available: https://doi.org/10.1109/RBME.2010.2084567
15. Ko, YC., Liu, Cl. & Hsu, WM. Varying Effects of Corneal Thickness on Intraocular Pressure Measurements with Different Tonometers. Eye 19, 327–332 (2005). https://doi.org/10.1038/sj.eye.6701458
16. R. Varma *et al.*, "Prevalence of and risk factors for diabetic macular edema in the United States," *JAMA Ophthalmology*, vol. 132, no. 11, pp. 1334–1340, 2014. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25125075/
17. L. Li *et al.*, "Artificial intelligence for diabetic retinopathy screening: A review," *Eye and Vision*, vol. 5, no. 1, p. 1, 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31488886/
18. Kanclerz P, Tuuminen R, Khoramnia R. Imaging Modalities Employed in Diabetic Retinopathy Screening: A Review and Meta-Analysis. Diagnostics (Basel). 2021 Sep 29;11(10):1802. doi: 10.3390/diagnostics11101802. PMID: 34679501; PMCID: PMC8535170.
19. M. S. Ip *et al.*, "The prevalence of diabetic retinopathy in the United States," *Archives of Ophthalmology*, vol. 122, no. 4, pp. 552–563, 2004. [Online]. Available: https://doi.org/10.1001/archopht.122.4.552
20. P. S. Silva *et al.*, "Peripheral lesions identified on ultra-widefield imaging predict increased risk of diabetic retinopathy progression over 4 years," *Ophthalmology*, vol. 122, no. 5, pp. 949–956, 2015. [Online]. Available: https://doi.org/10.1016/j.ophtha.2015.01.008
21. K. H. Maier-Hein *et al.*, "Why rankings of biomedical image analysis competitions should be interpreted with care," *Nature Communications*, vol. 9, no. 1, p. 5217, 2018. [Online]. Available: https://doi.org/10.1038/s41467-018-07619-7
22. H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Computer Science*, vol. 90, pp. 200–205, 2016. [Online]. Available: https://doi.org/10.1016/j.procs.2016.07.014

23. B. Abdillah, A. Bustamam and D. Sarwinda, "Classification of diabetic retinopathy through texture features analysis," 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Bali, Indonesia, 2017, pp. 333-338, doi: 10.1109/ICACSIS.2017.8355055.
24. 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. [Online]. Available: https://doi.org/10.1001/jama.2016.17216
25. 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. [Online]. Available: https://doi.org/10.1186/s40537-019-0192-5
26. A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, 2015. [Online]. Available: https://doi.org/10.1186/s12880-015-0068-x
27. M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012. [Online]. Available: https://doi.org/10.11613/BM.2012.031
28. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019. [Online]. Available: https://www.nature.com/articles/s42256-019-0048-x
29. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017. [Online]. Available: https://arxiv.org/abs/1702.08608
30. R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. [Online]. Available: https://doi.org/10.1109/ICCV.2017.74
31. J. Adebayo *et al.*, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 9505–9515. [Online]. Available: https://papers.nips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html
32. L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965. [Online]. Available: https://doi.org/10.1016/S0019-9958(65)90241-X
33. J. S. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 3, pp. 665–685, 1993. [Online]. Available: https://doi.org/10.1109/21.256541
34. Z. C. Lipton, "The mythos of model interpretability," *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018. [Online]. Available: https://doi.org/10.1145/3233231
35. A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018. [Online]. Available: 10.1109/access.2018.2870052
36. M. J. Page *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1743919121000522
37. W. M. Bramer, D. Giustini, G. B. de Jonge, L. Holland, and T. Bekhuis, "De-duplication of database search results for systematic reviews in EndNote," *Journal of the Medical Library Association*, vol. 104, no. 3, pp. 240–243, 2016. [Online]. Available: https://doi.org/10.3163/1536-5050.104.3.014
38. D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019. [Online]. Available: https://doi.org/10.3390/electronics8080832
39. J. Popay *et al.*, "Guidance on the conduct of narrative synthesis in systematic reviews," *ESRC Methods Programme*, vol. 1, no. 1, p. 92, 2006.
40. C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Munich, Germany: Leanpub, 2020. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
41. A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019. [Online]. Available: https://doi.org/10.1038/s41591-018-0316-z
42. J. R. Zech *et al.*, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, p. e1002683, 2018. [Online]. Available: https://doi.org/10.1371/journal.pmed.1002683
43. S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "DREAM: Diabetic retinopathy analysis using machine learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1717–1728, 2014. [Online]. Available: https://doi.org/10.1109/JBHI.2013.2294635
44. A. G. A. Padmanabha, M. A. Appaji, M. Prasad, H. Lu and S. Joshi, "Classification of diabetic retinopathy using textural features in retinal color fundus image," 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, China, 2017, pp. 1-5, doi: 10.1109/ISKE.2017.8258754.

45. Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 150–158. [Online]. Available: https://doi.org/10.1145/2339530.2339556
46. R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, 2017. [Online]. Available: https://doi.org/10.1016/j.ophtha.2017.02.008
47. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847. [Online]. Available: https://doi.org/10.1109/WACV.2018.00097
48. R. Sayres *et al.*, "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, 2019. [Online]. Available: https://doi.org/10.1016/j.ophtha.2018.11.016
49. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 3319–3328. [Online]. Available: https://proceedings.mlr.press/v70/sundararajan17a.html
50. Tursunalieva, A.; Alexander, D.L.J.; Dunne, R.; Li, J.; Riera, L.; Zhao, Y. Making Sense of Machine Learning: A Review of Interpretation Techniques and Their Applications. Appl. Sci. 2024, 14, 496. https://doi.org/10.3390/app14020496
51. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778
52. A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 3681–3688, 2019. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33013681
53. P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2369–2380, 2016. [Online]. Available: https://doi.org/10.1109/TMI.2016.2546227
54. P. W. Koh *et al.*, "Concept bottleneck models," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 5338–5348. [Online]. Available: https://proceedings.mlr.press/v119/koh20a.html
55. Z. Wang *et al.*, "Zoom-in-Net: Deep mining lesions for diabetic retinopathy detection," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, 2017, pp. 267–275. [Online]. Available: https://doi.org/10.1007/978-3-319-66179-7_31
56. P. Porwal *et al.*, "Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, 2018. [Online]. Available: https://doi.org/10.3390/data3030025
57. A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4198–4205. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.386
58. S. Lapuschkin *et al.*, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019. [Online]. Available: https://doi.org/10.1038/s41467-019-08987-4
59. C. J. Cai *et al.*, "Hello AI: Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making," in *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–24, 2019. [Online]. Available: https://doi.org/10.1145/3359206
60. J. Brown *et al.*, "Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks," *JAMA Ophthalmology*, vol. 136, no. 7, pp. 803–810, 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29801159/
61. Khan, A.H., Hayat, N. and Bilal, H., 2024. Numerical Simulation and Parametric Investigation ofIncremental Sheet Forming Process for MultilayerSandwich Panels. *International Journal of Emerging Engineering and Technology*, *3*(1), pp.1-12, 2024.
62. J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018. [Online]. Available: https://doi.org/10.1007/s11263-017-1059-x
63. D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6541–6549. [Online]. Available: https://doi.org/10.1109/CVPR.2017.354
64. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2097–2106. (Cited as methodology reference for CAM study). [Online]. Available: https://doi.org/10.1109/CVPR.2017.369
65. S. S. Escalera *et al.*, "Explainable diabetic retinopathy detection using deep learning and layer-wise relevance propagation," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 456–460. [Online]. Available: https://www.irjet.net/archives/V12/i6/IRJET-V12I668.pdf

66. V. Bellemo *et al.*, "Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study," *The Lancet Digital Health*, vol. 1, no. 1, pp. e35–e44, 2019. [Online]. Available: https://doi.org/10.1016/S2589-7500(19)30004-4
67. J. De Fauw *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018. [Online]. Available: https://doi.org/10.1038/s41591-018-0107-6
68. V. V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: a survey," *International Journal of Engineering & Technology*, vol. 7, no. 2.8, pp. 684–687, 2018. (Methodological comparison for SVM). [Online]. Available: https://doi.org/10.1088/1757-899X/1022/1/012046
69. Elena Vlamou, Basil Papadopoulos. Fuzzy logic systems and medical applications[J]. AIMS Neuroscience, 2019, 6(4): 266-272. doi: 10.3934/Neuroscience.2019.4.266
70. L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning—I," *Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975. [Online]. Available: https://doi.org/10.1016/0020-0255(75)90036-5
71. E. H. Mamdani, "Application of fuzzy algorithms for control of simple dynamic plant," *Proceedings of the Institution of Electrical Engineers*, vol. 121, no. 12, pp. 1585–1588, 1974. [Online]. Available: https://doi.org/10.1049/piee.1974.0328
72. T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-15, no. 1, pp. 116–132, 1985. [Online]. Available: https://ieeexplore.ieee.org/document/6313399
73. J. M. Mendel, "Type-2 fuzzy sets and systems: an overview," *IEEE Computational Intelligence Magazine*, vol. 2, no. 1, pp. 20–29, 2007. [Online]. Available: https://ieeexplore.ieee.org/document/4197699
74. Farman, H., Islam, N., Ali, S.A., Khan, D., Khan, H.A., Ahmed, M. and Farman, A., Advancing Rainfall prediction in Pakistan: a fusion of machine learning and time series forecasting models. *International Journal of Emerging Engineering and Technology*, *3*(1), pp.17-24, 2024.
75. Y. Zheng, Z. Xu and X. Wang, "The Fusion of Deep Learning and Fuzzy Systems: A State-of-the-Art Survey," in IEEE Transactions on Fuzzy Systems, vol. 30, no. 8, pp. 2783-2799, Aug. 2022, doi: 10.1109/TFUZZ.2021.3062899
76. J. L. Castro, "Fuzzy logic controllers are universal approximators," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25, no. 4, pp. 629–635, 1995. [Online]. Available: https://ieeexplore.ieee.org/document/370193
77. C.-F. Juang and C.-T. Lin, "An on-line self-constructing neural fuzzy inference network and its applications," *IEEE Transactions on Fuzzy Systems*, vol. 6, no. 1, pp. 12–32, 1998. [Online]. Available: https://doi.org/10.1109/91.660805
78. J. M. Alonso and A. Bugarín, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 60–78, 2020. [Online]. Available: https://doi.org/10.1016/j.inffus.2019.12.012
79. Fatima, S.A., Nasim, S.F. and Ahmed, S., Enhancing agricultural operations: big data analytics using distributed and parallel computing. *International Journal of Emerging Engineering and Technology*, *2*(2), pp.1-7, 2023.
80. R. E. Bellman and L. A. Zadeh, "Decision-making in a fuzzy environment," *Management Science*, vol. 17, no. 4, pp. B-141–B-164, 1970. [Online]. Available: https://doi.org/10.1287/mnsc.17.4.B141
81. R. Nagpal, D. Mehrotra and P. K. Bhatia, "Task based effectiveness evaluation of educational institute websites," 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), New Delhi, India, 2016, pp. 315-319, doi: 10.1109/ICCTICT.2016.7514600.
82. L. Seoud *et al.*, "Red lesion detection using dynamic shape features for diabetic retinopathy screening," *IEEE Transactions on Medical Imaging*, vol. 35, no. 4, pp. 1116–1126, 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7360182
83. Nasim, S.F., Khurram, M., Kamran, A., Fatima, S.A. and Qaiser, A., 2023. Environmental Monitoring and Agricultural Insights: Analysis of Cotton Crop Using PowerBI. *International Journal of Emerging Engineering and Technology*, *2*(2), pp.13-20, 2023.
84. J. Li, H. Xiao, D. Tan, M. Zhang and Y. Liu, "Image Colorization Based on Texture by Using of CNN," 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC), Xiamen, China, 2019, pp. 167-171, doi: 10.1109/ICIVC47709.2019.8980996.
85. Bano, A., Naqvi, Y., Ahmed, A. and Nasim, S.F., Enhancing agriculture prediction through AI and parallel distributed computing: A comprehensive study on the impact of weather. *International Journal of Emerging Engineering and Technology*, *2*(2), pp.21-28, 2023.
86. J. Cuadros and G. Breslauer, "EyePACS: An adaptable telemedicine system for diabetic retinopathy screening," *Journal of Diabetes Science and Technology*, vol. 3, no. 3, pp. 509–516, 2009. [Online]. Available: https://doi.org/10.1177/193229680900300315

87. Kaggle, "APTOS 2019 Blindness Detection." [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection
88. P. Porwal *et al.*, "IDRiD: Diabetic retinopathy – Segmentation and grading challenge," *Medical Image Analysis*, vol. 59, p. 101561, 2020. [Online]. Available: https://doi.org/10.1016/j.media.2019.101561
89. Islam, A., Saeed, A., Hassan, A. and Shahid, Z., AI-Powered Home Automation: A Simple and Smart Living Solution. *International Journal of Emerging Engineering and Technology*, *4*(1), pp.6-10, 2025.
90. Bancila, I.C., Mini-review: Experimental Approaches for the Biomechanical Testing of Bone. *International Journal of Emerging Engineering and Technology*, *4*(1), pp.11-20, 2025.
91. A. Luque, A. Carrasco, A. Martín, and A. de Las Heras, "The impact of class imbalance in classification performance metrics based on the confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–228, 2019. [Online]. Available: https://doi.org/10.1016/j.patcog.2019.02.023
92. J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. [Online]. Available: https://doi.org/10.1177/001316446002000104
93. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0118432
94. Shakeel, S.I., Aslam, S., Haq, H.B.U., Abbas, S., Akhtar, H.M.M., Abbas, S. and Khalid, I., Artificial Intelligence–Driven Imaging Advances in Lung Fibrosis: A Comprehensive Review. *International Journal of Emerging Engineering and Technology*, *4*(2), pp.1-9, 2025.
95. Islam, A., Saeed, A., Hassan, A. and Shahid, Z., AI-Powered Home Automation: A Simple and Smart Living Solution. *International Journal of Emerging Engineering and Technology*, *4*(1), pp.6-10, 2025.
96. W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017. [Online]. Available: https://doi.org/10.1109/TNNLS.2016.2599820
97. Awodeyi, A., Ibok, O.A., Ekwemuka, J.U., Idama, O. and Odesa, E., Enhancing Facial Recognition Performance with Data Augmentation in Occluded Environments. *International Journal of Emerging Engineering and Technology*, *4*(1), pp.27-32, 2025.
98. P. Hase and M. Bansal, "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5540–5552. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.491
99. M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021. [Online]. Available: https://doi.org/10.1016/S2589-7500(21)00208-9
100. E. Beede *et al.*, "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic eye disease," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12. [Online]. Available: https://doi.org/10.1145/3313831.3376718
101. N. Arun *et al.*, "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," *Radiology: Artificial Intelligence*, vol. 3, no. 6, p. e200267, 2021. [Online]. Available: https://doi.org/10.1148/ryai.2021200267
102. J. Amann *et al.*, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, p. 310, 2020. [Online]. Available: https://doi.org/10.1186/s12911-020-01332-6
103. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, p. 841, 2017. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3063289.
104. M. D. Abràmoff *et al.*, "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *NPJ Digital Medicine*, vol. 1, p. 39, 2018. [Online]. Available: https://doi.org/10.1038/s41746-018-0040-6
105. L. Serafini and A. d. Garcez, "Logic tensor networks: Deep learning and logical reasoning from data and knowledge," *arXiv preprint arXiv:1606.04422*, 2016. [Online]. Available: https://arxiv.org/abs/1606.04422
106. J. R. Zilke, E. L. Mencía, and F. Janssen, "DeepRED: Rule extraction from deep neural networks," in *International Conference on Discovery Science*, 2016, pp. 457–473. [Online]. Available: https://doi.org/10.1007/978-3-319-46307-0_29