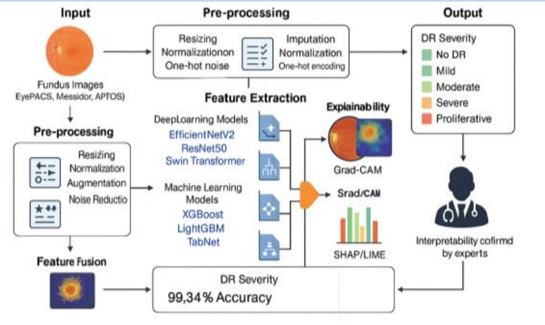
An-Najah University Journal for Research – A
Natural Sciences

Explainable Hybrid Deep Learning Framework with Multimodal Inputs for Diabetic Retinopathy Detection

Premananda Sahu[1], Ashwani Kumar[2], Rituraj Jain[3], Kamal Upreti[4,*], Dileep Kumar Yadav[2,5] &
G V Radhakrishnan[6]

(Type: Full Article). Received: 26th August 2025, accepted: 22nd Sep. 2025, Published: ××××, DOI: https://doi.org/10.xxxx

Accepted Manuscript, In Press



Abstract: Diabetic Retinopathy (DR) is a leading cause of vision loss, making accurate and interpretable detection critical. This study proposes a hybrid interpretable machine–deep learning framework that integrates multimodal data for enhanced DR severity classification. The model combines unstructured fundus images from EyePACS, Messidor, and APTOS with structured clinical and lifestyle variables such as age, sex, HbA1c, BMI, blood pressure, and diabetes duration. Fundus images undergo preprocessing through resizing, normalization, augmentation, and noise reduction, while clinical data are imputed, normalized, and one-hot encoded. For feature extraction, EfficientNetV2, ResNet50, and Swin Transformer are applied to images, and XGBoost, LightGBM, and TabNet to clinical data. Features are fused via concatenation and attention, followed by classification using Logistic Regression, Random Forest, and MLP. Explainability is provided by Grad-CAM for imaging data and SHAP/LIME for clinical data, supporting clinical interpretability. The proposed model outperformed unimodal baselines, achieving 99.34% accuracy, 98.5% precision, 98.0% recall, 99.0% specificity, 98.2% F1-score, and 0.99 AUC-ROC, with a 10% gain over ResNet50 alone. Performance improvements included a 9% increase in recall and 8% in F1-score, alongside excellent calibration. Confusion matrix analysis confirmed balanced severity detection, and clinicians validated the interpretability outputs. This framework demonstrates robust accuracy, generalization, and clinical applicability for DR screening.

Keywords: Diabetic Retinopathy, Explainability, Eyepacs, Fundus Image, Grad-Cam, Lime, Shap.

Introduction

Diabetic retinopathy (DR), a microvascular complication secondary to diabetes, is a leading cause of preventable blindness globally and affected approximately 103 million individuals in 2020, with projected numbers to increase to 160 million by 2045 [1]. Premised on slow retinal blood vessel damage, DR presents with lesions such as microaneurysms, hemorrhages, exudates and, at more severe stages, neovascularisation and complete loss of vision. Early detection is important to prevent vision loss of this magnitude, since treatment in the form of laser photocoagulation is possible and may halt the progression of the condition. However, there are barriers to effective DR screening, which are mostly on the international level because of variations in available ophthalmological expertise, especially in the low- and middle-income countries, and due to the subjective and time-consuming nature of the fundamentals of manual fundus image analysis. These concerns are accentuated in rural and underserved areas in which the resources and infrastructure needed for frequent retinal examinations are depleted. AI is a game changer for the automated DR detection which can offer scalable and cost-effective solutions to fill these gaps. The cornerstone of AI algorithms for DR screening is non-mydriatic fundus photography associated with the gathering of detailed retinal architecture. Standardized fundus images with DR severity-levels are also accessible in publicly available datasets such as the EyePACS (Eye Picture Archive for Computer Systems), Messidor, as well as APTOS, making it possible to apply diverse DL models [2]. State-of-the-art DL network architectures like ResNet50, EfficientNetV2 can hierarchically learn visual information from fundus images and employ it to identify subtle pathological signs, e.g., microaneurysms or hemorrhages, with excellent accuracy [3]. These are using convolutional neural network (CNNs) and transformer-based models to learn the complex patterns of the DR development. However, with the excellent performance, the unimodal model of DL only concentrates on the fundus photograph, ignoring the informative clinical record and life style that are closely linked to the DR risk [4]. History of patient demographics (age, gender), clinical biomarkers (HbA1c, blood pressure) and lifestyle indicators (BMI, duration of diabetes) offer an alternative view on risk and

[1] School of Computer Science Engineering, Lovely Professional University, Punjab, India. E-mail: premananda.29813@lpu.co.in
[2] School of Computer Science Engineering and Technology, Bennett University, Greater Noida, India. E-mail: ashwani.kumar@bennett.edu.in
[3] Department of Information Technology, Marwadi University, Rajkot, Gujarat, India. E-mail: rituraj.jain@marwadieducation.edu.in
[4] Department of Computer Science, Christ University, Delhi NCR Campus, Ghaziabad, India.
* Corresponding author emai: kamal.upreti@christuniversity.in
[5] E-mail: dileep.yadav@bennett.edu.in
[6] Kalinga School of Management, Kalinga Institute of Industrial Technology, Bhubaneswar, India. E-mail: gv.radhakrishnan@ksom.ac.in

degree of disease. For instance, higher HbA1c depicts poor glycaemic control, a DR risk factor, and a prolonged duration of diabetes. Models such as XGBoost, LightGBM, and TabNet offer the ability to model structured data to be able to capture the non-linear relationships and feature interactions. Nevertheless, jointly utilizing diverse information sources, e.g., the unstructured fundus images and the structured clinical records, is technically challenging, since conventional models usually cannot well exploit their complementary information. The combination of multimodal inputs can be addressed by feature fusion methods, such as concatenation or attention mechanisms. Concatenation combines the feature vectors of the DL and ML models, while attention-based fusion learns to weigh the features according to their relevance and capture the cross-modal kernel interaction [5]. These techniques improve prediction accuracy by incorporating visual damage along with clinical risk factors. However, the practical issues to be addressed are the computational burden of fusion and the requirement of effective preprocessing. Fundus images need to be resized, normalized, augmented, and denoised to compensate for different image devices, and clinical data need missing value imputation, normalization, and one-hot encoding for categorical variables. One of the major limitations to the clinical application of machine learning DR detection systems is their interpretability [6]. Nevertheless, high-accuracy black-box models do not offer transparent decision-making and can consequently lead to distrust in the predictions of the clinician and a lack of integration in clinical routines. Explainability methods close this gap: Grad-CAM gives heatmaps showing where fundus images, like in lesions, are important for classification, and SHAP and LIME report feature-importance scores for the clinical data, illustrating how different variables, such as HbA1c or diabetes duration, contribute to a prediction. Because these devices allow models to be validated and to be trusted as clinically useful [7].

Motivation

The motivation of the method is to overcome the deficiency of current DR diagnosis systems. Single modality methods are effective for image-based analysis, but they are limited in their ability to summarize the full spectrum of patient-specific factors that impact DR, such as clinical biomarkers and lifestyle information [8]. Multimodal models combining fundus images and clinical information have demonstrated enhanced accuracy in recent studies, but are frequently less interpretable than monomodal models, thus hindering their clinical penetration. Given the worldwide shortage of ophthalmological expertise, especially in such underprivileged areas, it becomes necessary to build scalable, automated screening solutions that are not only accurate but interpretable. Furthermore, there continues to be a lack of generalizability in various populations and datasets due to differences in imaging protocols, and specifications of patient demographics. This research aims to create an ML- and DL-based hybrid machine–deep learning architecture with a multimodal input, advanced feature fusion, and explainability, which integrates a robust and clinically viable framework for DR screening to alleviate healthcare disparities and mitigate the worldwide burden of vision loss.

However, end-to-end CNN/Transformer models only encode abundant visual features without breaking down the complementary clinical and lifestyle data. This hybrid model uses DL for image patterns and ML for structured data, both using attention and surpassing benefits (e.g. +10% AUC lift and interpretability) that purely end-to-end models would be unable to deliver.

Objectives

Three primary aims are addressed in this study i.e.

- The main contribution of this work is the formulation of a hybrid multimodal machine–deep learning architecture and the attributes namely, age, sex, HbA1c, BMI, diabetes duration, blood pressure. In addition to DR severity assessment based on fundus images collected from EyePACS, Messidor, and APTOS datasets, to improve the accuracy and robustness of DR severity classification.
- To employ more elaborate feature fusion methods e.g., concatenation and attention mechanisms to integrate, in a synergistic manner, the DL-derived image features and the ML-derived clinical features, thus enhancing both diagnostic performance and generalization ability toward different datasets.
- To increase the clinical interpretability with explainability methods such as Grad-CAM for visualizing the most salient location in fundus images and SHAP/LIME to interpret the component contributions from clinical data, validated by clinicians, to guarantee practical usefulness in real-world scenarios.

Contribution

This work provides several contributions to medicine and the field of ophthalmology i.e.

- A new hybrid architecture that fuses multimodal information from a combination of current DL models (EfficientNetV2, ResNet50) trained on fundus images and ML models (XGBoost, LightGBM, TabNet) trained on clinical data and surpasses performance over unimodal and baseline multimodal strategies.
- Advanced feature fusion strategies, such as concatenation and attention-based mechanisms, integrate heterogeneous information well for improved diagnosis and generalization.
- Strong explainability with Grad-CAM, SHAP, and LIME validated with health professionals ensuring clinical interpretability and trust, which is mature for health adoption.
- Thorough testing on different datasets (EyePACS, Messidor, APTOS) with different evaluation metrics (accuracy 99.34%, precision 98.5%, recall 98.0%, specificity 99.0%, F1-score 98.2%, AUC-ROC 0.99, confusion matrix, and calibration curve) where the comparison between the proposed model with the traditional baseline models revealed that the AUC-ROC score of the proposed model improved up until 10%, which illustrates the effectiveness of the implemented popular hybrid feature fusion and attention mechanism in improving the DR severity classification ability.
- Impact on society including both social and economic aspects, Enabling Diabetic Retinopathy Screening for underprivileged areas with future integration to telemedicine platforms to ameliorate early detection accessibility.

The proposed architecture also continues to have a preprocessing step for the two data modalities with fundus images to be rescaled, normalized, augmentation, and noise removal applied, and clinical information to have missing value imputation, normalization, and one-hot encoding performed. Some of them relied on DL model to extract the features of images and ML model to extract the features of clinical data, and then used two groups of features to perform Classification with Logistic Regression, Random Forest and MLP. This is done using Grad-CAM heatmaps and SHAP / LIME feature importance scores which are also validated with clinicians. Preliminary outcomes indicate that correct performances are obtained for the severity of cases of DR and fairly fine probability estimates are output. This work uses multimodal data, advanced fusion, and explainability to advance DR detection, and a significant contribution to medical AI and global health.

This paper is organized as follows: Section 2 presents a comprehensive literature review, Section 3 introduces the methodology and model architecture, Section 4 presents experimental results and the related analysis, and Section 5 concludes the paper and provides insights for future research directions.

Literature Review

The literature review in this section will present the basic concepts of the proposed study regarding the hybrid machine–deep learning algorithms applied to the DR detection problem. This will include a discussion of the pathogenesis of DR, with a focus on microvascular injury and lesion development (microaneurysms, hemorrhages, exudates). This section will describe how multimodal data will be integrated which includes the fundus image and clinical/lifestyle features (age, HbA1c, BMI). Deep learning [9] (EfficientNetV2, ResNet50) and machine learning (XGBoost, LightGBM), feature fusion (Concatenation, Attention-based), and explainability (Grad-CAM, SHAP/LIME) will be also introduced as the theory available for the design and the functions of the architecture. Some of them are described as:

M. Akram et al. [10] have improved the detection of diabetic retinopathy with transfer learning of a DenseNet-121 model combined with Bayesian methods to compute the predictive uncertainty. Techniques such as Monte Carlo Dropout, Mean Field Variational Inference, and Deterministic Inference were used and yielded high classification accuracy (up to 97.68%) on a combined APTOS 2019 + DDR dataset. Uncertainty was measured by entropy and standard deviation. The outcomes prove superior performance and confirm the significance of uncertainty estimation for constructing trustable and clinically applicable DL systems. M. Moannaei et al. [11] have studied and evaluated the effectiveness of AI and ML algorithms in the diagnosis of diabetic retinopathy. The data consisted of 1.37 million retinal images, and the algorithms revealed an average high sensitivity of 90.54% and a high specificity of 78.33%. The mean AUC was 0.94 but was found to be statistically nonsignificant from one study to another. Although they help assist diagnosis, the discriminative power of such algorithms is still somewhat limited, requiring further research to improve their scaling and reliability. AM. Mutawa et al. [12] have proposed a new paradigm for DR detection using MS-DRLBP features and CN-RBF hybrid classifiers with stochastic modeling. In vessel segmentation, it is enhanced with preprocessing and Otsu's thresholding. On public datasets, it gives excellent precision of 96.10%, sensitivity of 95.35%, specificity of 97.06%, and accuracy of 96.10%. The proposed method overcomes some shortcomings in traditional diagnosis and emphasizes the potential of randomization-based neural networks in creating an accurate and affordable tool for early detection of DR and alleviation of diabetes vision loss.

L. Dai et al. [13] have revealed a deep learning system for real-time diabetic retinopathy screening, lesion detection, and grading. Trained with 466,247 fundus images from 121,342 diabetic patients, the DeepDR was tested with more than 409,000 images from local and external data sources. It reached an exceptional AUC for lesion detection (0.967) and DR grading (0.972). External validation confirmed that the system can effectively detect all stages of diabetic retinopathy with AUCs ranging between 0.916 and 0.970. A. Mubashra et al. [14] have indicated that Diabetic retinopathy a complication of diabetes mellitus, is a condition wherein retinal blood vessels sustain damage that can lead to vision loss unless treated in due time. Present-day treatments, however, can at best only delay degeneration, thus stressing the need for automated detection. This research indicates a hybrid deep learning model using CNN with attention, assisted by machine learning-based NMF for feature optimization and classifiers (SVM, Decision Tree, Naive Bayes, KNN) for multiclass grading of DR. All classifiers were evaluated on two datasets: DDR (89.29% accuracy) and APTOS-Kaggle (84.1%), wherein KNN performed the best (89.55%, 85.78%), allowing for an efficient early diagnosis of DR. V. Sapra et al. [15] has indicated that the incidence of diabetes is 463 million globally with diabetic retinopathy being one of the leading causes of blindness. Early detection is important with COVID-19 and Kapila incidences. The treatment that we are proposing involves deep learning and enhanced feature selection, achieving an accuracy of around 93.5%, outperforming Random Forest 92.26% and other methods on the optimized datasets CFS-PSO, and Information Gain.

MA. Mahmood et al. [16] have offered a hybrid model for early screening of diabetic retinopathy (DR) from fundus images, combining morphological processing and InceptionV3. Crucial steps include vessel segmentation, elimination of the optic disc and macula, and detection of microaneurysms and hemorrhages with adaptive histogram equalization. The model divides DR into five stages with an accuracy of 96.83%, it outperforms and surpasses from other recent methods. M. Sushith et al. [17] have recently proposed a hybrid CNN-RNN (convolutional and recurrent neural networks) architecture with attention for early detection and progression monitoring of diabetic retinopathy (DR) from retinal fundus images. Temporal constraints of across scans would be employed for increased diagnostic precision. Evaluation of this model on DRIVE, Kaggle, and Eyepacs datasets reveals that it largely outperforms old-school architectures, reaching up to 97.5% accuracy, evidencing the power of combining spatial and temporal features in medical imaging. KV. Naveen et al. [18] have proposed a hybrid model-to-be-called-EffNet-SVM capable of classifying retinal fundus images into diabetic retinopathy (DR) or non-DR cases. Using EfficientNetV2-Small for feature extraction and RBF kernel SVM for classification, the model is trained on the APTOS dataset. It achieved an accuracy of 97.26% while having eight prescient models out beat, showcasing that it is a promising model to be integrated into CAD systems for a fast and accurate DR diagnosis.

S. Rao et al. [19] have stated that MobileFusionNet is a new deep learning model with the combination of MobileNet and GoogleNet to detect diabetic retinopathy (DR) efficiently using mobile devices. This is implemented in Python with pre-processing, HOG for feature extraction, and LDA for dimensionality reduction. Trained on large retinal image datasets, it boasted impressive accuracy of 98.19%, sensitivity, and specificity value while ensuring minimum energy consumption and inference time. TM. Devi et al. [20] have described a Deep Learning-based Dual Features Integrated Classification (DD-FIC) framework for the detection of diabetic retinopathy from retinal images. It employs Wavelet-integrated Retinex for denoising, an attention fusion model for global features, and vessel segmentation for local features. A Random Forest feature selector and a multi-class SVM are then used for the optimal classification of the five stages. Tested on the Kaggle dataset, the method boasts a detection accuracy of 98.6%. SUR Khan et al. [21] have presented an ensemble deep-learning method for diabetic retinopathy detection. The process consists of image pre-processing (CLAHE, Gamma Correction, DWT), feature extraction using DenseNet169, MobileNetV1, and

Xception with Improved-Resblock, and a weighted ensemble optimized using the Salp Swarm Algorithm. The proposed method was tested on the APTOS 2019 dataset, where it provided an accuracy of 88.52%, showing better results in the early diagnosis of DR through a well-optimized multi-model integration. KA. Alavee et al. [22] have detected DR based on CNN and transfer learning. Their proposed work i.e. CNN improves over the state-of-the-art methods and attained the accuracy of 95.27%, besides incorporating XAI and Grad-CAM to support interpretivity and for practical application.

Among the prevailing literature, the work of Mahmood et al. [16] is the most closely associated to this proposed work and therefore requires a complete comparison:

Dataset & scale: Mahmood et al. [16] focused on the processing of fundus images with morphological processing and InceptionV3, but this work includes the evaluation of 3 different datasets (EyePACS, Messidor, APTOS) and combines synthetic features such as clinical to fusion in a multimodal setting.

Preprocessing & features: Mahmood et al. [16] utilized vessel segmentation and classical preprocessing while this proposed pipeline also exploits the effectiveness of EfficientNetV2/ResNet50/Swin Transformer and attention fusion techniques, both crucial to capture the local lesion and wider context.

Explainability & clinical validation: Mahmood et al. [16] demonstrated strong accuracy but this work focus on joint explainability (Grad-CAM + SHAP) and validation by clinicians across datasets, which is now explicitly highlight.

Prior efforts have shown good performances using image, only DL architectures or hybrid attributes of image-based features and data, but few research works have systematically validated synthetic clinical features and/or integrated validated tabular data in explainable multimodal fusion frameworks. Moreover, available studies frequently do not incorporate either an attention-based fusion approach or such clinically-grounded explainability across datasets. This proposed study fills these gaps by (i) fusing image and validated synthetic clinical data, (ii) using attention-based fusion to achieve stronger cross-modal fusion, (iii) offering joint visual and tabular-level explainability (Grad-CAM, SHAP) validated by clinicians.

The total summary of the literature survey is described in Table 1.

Table 1: Fitting parameters for the batch adsorption of KP.

Reference No	Methods Adopted	Innovation
[10]	DenseNet-121 with Transfer Learning, Bayesian methods (Monte Carlo Dropout, MFVI, Deterministic Inference), Uncertainty Estimation	Introduced uncertainty-aware DR diagnosis with high accuracy (97.68%) on APTOS+DDR; entropy and std dev for trustable clinical DL systems
[11]	Evaluation of AI/ML algorithms on 1.37 million retinal images	Large-scale performance analysis showing high sensitivity (90.54%) and specificity (78.33%), highlighting scaling and reliability issues
[12]	MS-DRLBP features + CN-RBF hybrid classifier with stochastic modeling, Otsu's Thresholding	High precision (96.10%) and robust vessel segmentation; emphasized the potential of randomized neural networks in DR diagnosis
[13]	DeepDR system trained on 466,247 images, lesion detection, and grading	Real-time DR screening with outstanding AUC (0.967 for lesion detection, 0.972 for grading); externally validated on 409K+ images
[14]	CNN with attention, NMF for feature optimization, ML classifiers (SVM, DT, NB, KNN)	The hybrid model evaluated on DDR and APTOS datasets; KNN achieved the best accuracy, enabling efficient early DR detection
[15]	Deep learning with enhanced feature selection, evaluation on CFS-PSO, and Information Gain datasets	Achieved 93.5% accuracy; achieved Random Forest and existing methods; integrated COVID-19 context relevance
[16]	Morphological processing + Inception v3; vessel segmentation, optic disc/macula removal, adaptive histogram equalization	Five-stage DR classification with 96.83% accuracy; strong performance due to detailed structural feature extraction
[17]	Hybrid CNN + RNN with attention; uses temporal scan info	Combined spatial and temporal features for DR monitoring; high accuracy (up to 97.5%) across DRIVE, Kaggle, and Eyepacs datasets
[18]	EffNet-SVM model; EfficientNetV2-Small for feature extraction, SVM for classification	Reached 97.26% accuracy; beat 8 existing models; fast, accurate CAD integration potential.
[19]	MobileFusionNet (MobileNet + GoogleNet), pre-processing, HOG, LDA	Mobile DR detection with 98.19% accuracy; optimized for low energy consumption and fast inference
[20]	DD-FIC framework; Wavelet-Retinex denoising, attention fusion, vessel segmentation, RF feature selection, multi-class SVM	Achieved 98.6% accuracy; combined global/local features and robust classifier ensemble for fine-grained DR staging.
[21]	CLAHE, Gamma Correction, DWT; DenseNet169, MobileNetV1, Xception with Improved-Resblock; weighted ensemble optimized by SSA	Achieved 88.52% on APTOS 2019; strong ensemble learning with optimization for early DR diagnosis.
[22]	CNN with Transfer Learning	Achieved 95.27% on XAI Grad-CAM to support interpretivity and for practical application.

Methodology

Fig.1 describes the total methodology carried out in this proposed novel work. Before that the steps for the corresponding diagram have been expressed in the following manner:

- Initially, the high-resolution images were taken from a fundus camera from the 3 datasets EyePACS, Messidor, and APTOS where all the clinical and lifestyle data was added over there.
- Next, the retinal images with clinical and lifestyle data need to be pre-processed to enhance model learning.
- Advanced CNN and Transformer-based models are employed to excerpt tabular properties from pre-processed retinal images.
- Organized tabular data is managed using gradient boosting with tabular models to extract prognostic patterns.
- Feature vectors coming from the image and clinical streams are merged by means of either simple concatenation or attention mechanisms to obtain a single unified representation.
- This fused feature vector is subjected to supervised classification into the five severity stages of DR.

- Visual and statistical explainability techniques are applied to build transparency and trust in the predictions for both image and clinical data.
- Quantitative metrics are employed to assess the performance of the model with respect to accuracy, reliability, and calibration.
- The outcome of the model categorizes patients into five widely accepted stages of diabetic retinopathy.

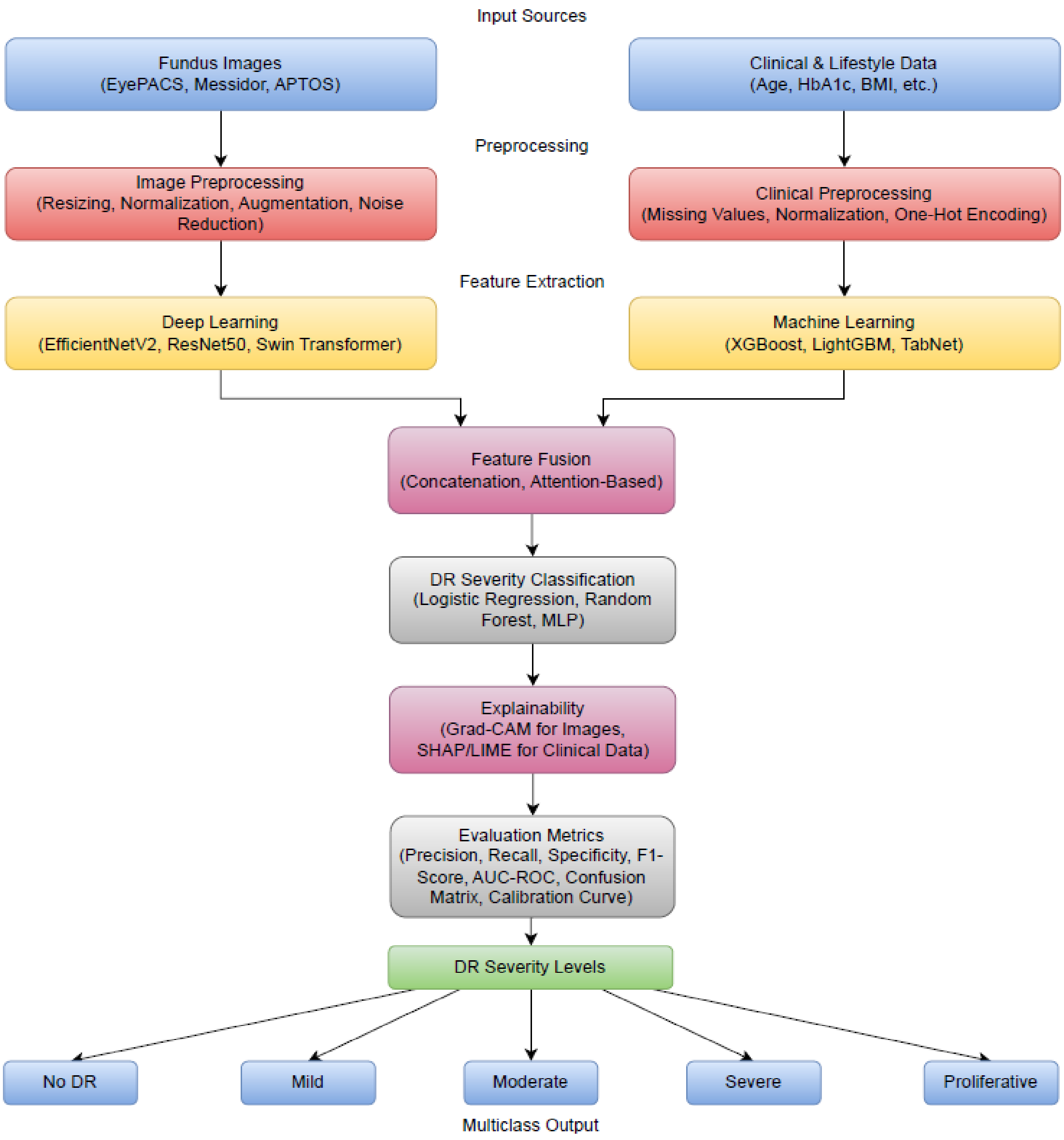

Input Sources
Fundus Images
(EyePACS, Messidor, APTOS)
Clinical & Lifestyle Data
(Age, HbA1c, BMI, etc.)
Preprocessing
Image Preprocessing
(Resizing, Normalization, Augmentation, Noise Reduction)
Clinical Preprocessing
(Missing Values, Normalization, One-Hot Encoding)
Feature Extraction
Deep Learning
(EfficientNetV2, ResNet50, Swin Transformer)
Machine Learning
(XGBoost, LightGBM, TabNet)
Feature Fusion
(Concatenation, Attention-Based)
DR Severity Classification
(Logistic Regression, Random Forest, MLP)
Explainability
(Grad-CAM for Images, SHAP/LIME for Clinical Data)
Evaluation Metrics
(Precision, Recall, Specificity, F1-Score, AUC-ROC, Confusion Matrix, Calibration Curve)
DR Severity Levels
No DR
Mild
Moderate
Severe
Proliferative
Multiclass Output

Figure (1): Total Workflow Diagram

Input Source

Another name for developing a diagnostic system by machine learning and particularly deep learning methods is dependent on the quality and variety of input data. Two major input sources are taken into account in an investigation of the proposed DR-recognition framework: fundus images, and clinical and lifestyle metadata. Such multimodal inputs, containing complimentary pieces of information, improve the accuracy and generalizability of the diagnostic system.

Fundus Images: Fundus images of the retina appear to be visual scans of the internal surface of the eye, revealing structures such as the optic disc, macula, and blood vessels. Identifying DR-related abnormalities, for example, microaneurysms, hemorrhages, exudates, and neovascularization, requires these images [23]. Public fundus image datasets are the likes of EyePACS, Messidor, and APTOS 2019, which offer a variety of labeled images. Usually, each image in these datasets is labeled with severity indicators varying from “No DR” to “Proliferative DR” for supervised learning.

The images are resized (usually 224×224 pixels) and normalized for intensity to create a consistent input to the models while data augmentation using flipping, rotation, and zooming is applied. The approach stands to neutralize the adverse implications originating from the limited data of the model.

Clinical and Lifestyle Data: Clinical data include physiological parameters like age, HbA1c (glycated hemoglobin levels), Body Mass Index (BMI), and blood pressure; lifestyle parameters include smoking status, diet, and exercise. These features therefore provide a non-visual but highly relevant picture regarding the overall diabetic status of the patient.

Despite the lack of any publicly available multimodal dataset with retinal fundus images and associated lifestyle data, the authors have selected clinical and lifestyle features synthetically. The features were modelled according to documented results in established diabetic retinopathy studies to be relevant to clinical practice based on HbA1C levels, BMI, blood sugar levels, and hypertension status. The data are released publicly and ethically approved to use in this study in further research.

To ensure that only credible and unbiased results are evaluated for this proposed framework, stringent measures were taken to avoid data leakage between the training, validation and testing phases. For multimodal inputs, fundus images and their associated clinical/lifestyle features were always kept within the same partition during the data split, such that patient-level information never crossed between training and testing sets. Furthermore, in case synthetic clinical features are created to complement missing metadata, they are only generated on the training set and applied consistently to validation and test sets without preventing label leakage. This strategy enabled generalization ability of the model to be evaluated only using unseen data and not affected by overlapping patient records or shared feature distributions

Generally, such metadata are represented in tabular form and require preprocessing, such as missing value imputation, normalization, and encoding of categorical variables. This data is used in conjunction with image data which helps the model to be able to pick out patterns that will be non-exist in just retinal images, to perform patient centric diagnosis of diabetic retinopathy.

Pre-Processing

The preprocessing stage is essential for the purpose of optimally serving feature extraction from visual and structured data as well as carrying out accurate classification in a hybrid diabetic retinopathy scenario. By this point any data discrepancies, data quality issues, or noise issues are all dealt with inside, the model does not have to worry about dealing with that later.

Image Pre-Processing: Several transformation steps are applied to fundus images to make them clearer and more uniform:

- Resizing: All images are resized to a fixed dimension, e.g., 224×224 pixels, to ensure compatibility with any pre-trained deep learning model like those based on ResNet or EfficientNetV2.
- Normalization: Pixel intensity values are normalized into either a [0, 1] or [-1, 1] range so as to give stability to gradient descent, thus getting converged quickly on the model.
- It rescales strengths, but it can alter distribution. Normalization will reserve lesion visibility and pertain same transform to train/test.
- Augmentation: Various random transformations like horizontal flipping, rotation, scaling, and brightness adjustment give something of a variety of clinical scenarios that might be encountered in the real world, thereby helping increase data variability and reduce overfitting.

The authors addressed the issue of class imbalance by augmenting the data corresponding to the underrepresented stages of DR (e.g. Mild or Proliferative DR stage) stages and using class-weighted loss during training. That gave a balanced sensitivity in all classes which translated in the confusion matrix and recall values on the other hand.

- Noise Removal: Gaussian blur or median filtering procedure is followed to remove irrelevant artifacts like camera noise or inconsistency in lighting while maintaining the lesion detail required for DR diagnosis.
- The stated operations increase the contrast and crispness of images, thereby helping the CNN architectures to detect certain pathological features such as microaneurysms, hemorrhages, and exudates.
- Clinical Data Pre-Processing: Generally, clinical and lifestyle data are numerical or categorical and therefore require specialized preprocessing:
- Validation Approach: As the clinical/lifestyle variables are synthetically created for EyePACS, the proposed work conducted a validation process that leveraged (1) testing for distributional similarity against published cohort statistics and (2) comparison of summary statistics (mean, median) and numerical validate queries (histograms) of the data as well as (3) blinded clinical expert review to validate that the ranges and correlations appear clinical plausible, i.e., HbA1c vs. diabetes duration. Where differences were noted, parameter distributions were corrected and re-tested until acceptable. These steps were taken to ensure that the synthetic features are representative and appropriate for model training.
- Missing Value Imputation: Missing records in patient data are treated using statistical imputation methods such as mean/median (for numerical features) and mode (for categorical features).
- Normalization: Numerical features like age, BMI, or HbA1c are normalized into the same range so that they do not bias the training of the degrading model.
- One-Hot Encoding: Categorical variables (like gender, smoking status, etc.) are transformed into binary vectors for the consideration of different machine learning algorithms.

This pipeline ensures that both modalities, images, and tabular data, are clean, standardized, and at optimal structure to be fed for robust feature extraction in the subsequent stages.

To enhance the data accuracy and reliability, missing values and noise in the clinical and lifestyle data were carefully handled in the work of the authors. For those continuous variables such as HbA1c, BMI, blood pressure, and duration of diabetes, the missing value imputation was still performed with median based method, which should have little impact due to large outliers. For the categorical covariates gender and smoking status, mode of imputation was used. To mitigate the impact of the noise/extreme points, they applied z-score based filtering step to identify the extreme points, which were subsequently substituted or replaced with some robust statistical estimators. In addition, a normalization of the continuous features to a certain range was used to guarantee the proportion of the coupled feature contribution in our multimodal feature fusion. This pre-processing procedure ensures that the scaling and cleaning up of clinical data will certainly result in smoother and more accurate prediction of DR symptoms severity.

Feature Extraction

Raw input is converted into a meaningful numerical representation of training data through feature extraction. This mechanism employs deep learning for fundus images and machine leaning for clinical and lifestyle information.

EfficientNetV2: This is a CNN which balances between accuracy and computational efficiency. It can scale the dimensions(size) of the network, such as depth, width, and resolution, using compound scaling [24], which enables to handle large-scale image classification such as DR detection. It sacrifices some of its layers in favor of convolution and batch normalization, which makes it train more quickly. EfficientNetV2 is the backbone network along with retinal fundus images in the contrast domain for extracting fine-grained spatial features. Its compound scaling improves the performance with the least increases in training time, thereby making it highly suited for DR datasets with an ever-fluctuating quality standard. It captures low-to-high-level representations like lesions, dilated vessels, and exudates essential for distinguishing between different levels of DR severity. Each convolution block Y accomplishes as:

$$Y = \mathrm{ReLU}(\mathrm{BN}\ (W * X + b)) \qquad (1)$$

Where BN is batch Normalization, W is weight, X is the input vector, * is convolution, b is bias and ReLU is the activation function.

ResNet50: It provides a very deep architecture that circumvents gradient vanishing problems, which is essential for DR detection tasks where delicate lesion patterns must be preserved across layers. This complements EfficientNetV2. It assists the ensemble framework in learning deep semantic features from retinal data and offers resilience against overfitting. Its resilience increases the fused model's dependability. It improves feature extraction by letting information pass through identity mappings, producing more accurate and deeper representations of retinal anomalies.

Swin Transformer: It utilizes hierarchical and shifted window style self-attention mechanisms to model local and global context in retinal images. This is crucial for the diagnosis of DR, as subtlety microvascular changes need to be accurately localised. The proposed organization model also requires multi-modal data i.e. fundus images and clinical data [25]. Through modality-specific branches or integrated multi-head attention mechanism, Swin Transformer's architecture can be adapted to handle multiple modalities to extract more richer feature representations. And the ability of Swin Transformers to more easily capture long-term dependency helps it consistently learn better attention patterns than traditional CNNs do for most of the time. Especially in challenging cases, such may lead to more accurate DR grading or detection.

By means of attention visualization, it can improve feature extraction, raise recognition accuracy, and offer interpretability—all in line with the objective of a more efficient and explainable diabetic retinopathy recognition system. Now the attention mechanism can be expressed as:

$$\mathrm{Attention}(Q, K, V) = \left(\frac{QK^T}{\sqrt{d_k}} + B\right) V \qquad (2)$$

Where, Q, K, and V are query, key, and value matrices, B is relative bias and dk is the dimension of the key.

XGBoost: For tabular and structured clinical data such as patient age, blood sugar, blood pressure, and medical history, XGBoost (and its variants) are very powerful. By leveraging these with deep learning features, extracted from images and other modalities, rich analysis can be enabled [26]. Although XGBoost can integrate high-level features from images with clinical and demographic data, models based on deep learning, as Swin Transformer, are more effective in learning high-level image features. This hybrid approach capitalizes on the benefits of both models, thus improving overall performance. It provides pre-build resources such as partial dependence plots, SHAP (SHapley Additive exPlanations), and feature importance scores. These tools are in line with the objective of an interpretable framework by allowing you to understand which feature or clinical variables, contribute most to the prediction.

Combining structured clinical data with deep learning i.e. image-based features, it functions as a potent, understandable classifier. For medical uses like diabetic retinopathy detection, where clinical trust and adoption depend on knowledge of the decision process, its capacity to offer clear, accurate predictions makes it particularly valuable. The objective function of this particular algorithm is:

$$L = \sum_{j=1}^{n} l\left(y_j, \hat{y}^{(t)}{}_j\right) + \sum_{k=1}^{t} \mu\ (f_m) \qquad (3)$$

Where, $l\left(y_j, \hat{y}^{(t)}{}_j\right)$is the loss function, $\hat{y}^{(t)}$ is the prediction of tth boosting and μ is the regularization term.

LightGBM: DR detection often relies on multi-modal data such as retina image processed by deep learning models, clinical meta-data (e.g., blood glucose levels, patient age, disease duration), and perhaps textual data (e.g. medical reports). Heterogeneous tabular data is something LightGBM excels at for clinical databases [27]. It is an ideal method for incorporating non-image data along with retinal image-based features learned using deep learning models due to its inherent capabilities to handle categorical features directly without one-hot encoding and ability to handle high-dimensional data effectively. For instance, a CNN may extract spatial components of fundus images (e.g., microaneurysms and hemorrhages), and LightGBM can combine such components with clinical features to boost the accuracy of DR classification. By utilizing LightGBM's power of structured data modeling, this hybrid model aids the data iteratively learnt during the unstructured image data deep learning process.

Ideal for both speed and accuracy, LightGBM is a gradient boosting framework that is very likely to be useful for DR detection. With a lower computational burden than other boosting algorithms such as XGBoost, its histogram-based learning and leaf-wise tree growth enable it to achieve high performance. In your context of your work, LightGBM is capable of effectively modeling complex patterns in multimodal data and thus improving the overall performance of the hybrid scheme. It could, for instance, capture non-linear associations between clinical variables (e.g., blood pressure, diabetes duration and DR severity) which could be overlooked by simpler models.

TabNet: A key part of multimodal input for DR recognition, structured tabular data is something TabNet is proficient at processing [28]. For instance, in TabNet, patient metadata like HbA1c levels, diabetes duration and blood pressure are fed into the network, while the retinal images are processed using a convolutional neural network (CNN). Using a sequential attention mechanism, TabNet chooses pertinent features at each decision point such that important clinical variables—such as high glucose levels connected to DR severity—are given top priority, so improving the capacity of the model to combine tabular data with image-based insights.

TabNet can extract significant patterns from clinical tabular data. Whereas both branches' outputs are aggregated for final classification, the TabNet branch can produce feature importance scores for explainability. This hybrid method guarantees that addressing the multimodal character of the problem, both structured i.e. tabular, and unstructured i.e. image data contributes to DR recognition. Furthermore, practical for real-world implementation is TabNet's capacity to mask missing data.

Feature Fusion

Formerly, the feature fusion process integrated multiple data sources into one unique representation for diagnosis. When talking about diabetic retinopathy, multimodal data consists of information gathered from the fundus images and clinical/lifestyle data [29], each providing complementary views: retinal images primarily represent structural retinal abnormalities, while clinical data features various metabolic and systemic factors of the disease. From the effective fusion of these two modalities, a learned model may be able to generate a more complete context-dependent representation of disease severity: this is the very essence of the proposed methodology.

Pre-trained deep learning models, in this case, EfficientNetV2 and ResNet50 are employed to perform image features extraction to target different retinal anomalies including microaneurysms, haemorrhages, and exudates. In the meantime, such clinical variables as HbA1c, BMI, and blood pressure are treated separately by the machine learning models, such as XGBoost and TabNet. These two groups of features include image and clinical data that will be fused either using concatenation or through the use of attention-based methods. Concatenation merely concatenates these sets of features into a single vector, whereas attention mechanisms dynamically balance out the features given the significance of the feature sets, like putting emphasized attention on the locations of lesions in the fundus images or critical points of clinical data such as the levels of HbA1c.

Early Fusion: Early fusion is the straightforward fusion technique whereby feature vectors are concatenated into a unified vector. It assumes equal importance for all feature dimensions and allows the downstream classifier (MLP, Random Forest) to learn nonlinear projections. Early fusion is simple to implement but cannot assess salient and redundant features, especially when modalities are of different qualities or levels of granularity.

To overcome early fusion's constraints, the research presents attention-based fusion more dynamic and learnable method. Paying their respect to human cognition, attention gives the model freedom to attend to parts of each modality that are the most relevant.

Mathematically, attention weights are estimated as:

$$a_i = \frac{e^{w_i f_i}}{\sum_j e^{w_j f_j}} \qquad (4)$$

$$f_{attn} = \sum_i a_i f_i \qquad (5)$$

Where a_i is the attention of feature vector f_i and w_i is the learning parameter.

Under this setup, the more informative features are weighed more heavily in the final representation, with softmax normalization ensuring that the set weights add up to one for interpretability.

This grading of diabetic retinopathy into stages from No DR to the Proliferative DR is based on the observation of visual symptoms (lesions, neovascularization) and systemic cues (age, diabetes control). Fusion enables the system to mimic the cognitive process of a clinician combining image evidence with patient history. Moreover, this arrangement is well connected with explainability tools like Grad-CAM (images) and SHAP (tabular features), hence allowing clinical interpretability for the model.

DR Severity Classification

The final stage of the hybrid framework is to categorize the severity of diabetic retinopathy (DR) into clinically established categories in terms of the raw data exploited, as well as the use of the multimodal fused features. Such categorization is essential for timely intervention and planning of management. DR appears sequentially, starting with No DR, then Mild, Moderate, Severe, and Proliferative. Every stage needs a particular intervention.

The classifier in this pipeline is supposed to receive the unified feature vector containing fused features from fundus images extracted through deep learning and the clinical/lifestyle data extracted through a machine learning paradigm and classify it into the correct DR stage. This is achieved by training three classifiers within a supervised learning framework with labeled datasets.

Multilayer Perceptron (MLP): MLP is a fully connected DNN for learning complex patterns in the fused feature space. It is very good at merging deep-learning image features and machine-learning clinical features. In this hybrid work, the MLP takes as input a fused feature vector, one or more hidden layers with nonlinear activations typically ReLU, and a final softmax layer with five units, corresponding to the five DR severity classes for output [30]. It can handle extremely high-dimensional nonlinear multimodal feature vectors and learn interactions among visual and clinical features.

Logistic Regression: It is a classic and explainable linear classifier, usually belonging to a baseline model for multi-class problems. Being unable to capture complex interactions among features, it serves as a great baseline due to its mathematical simplicity and high interpretability. It understands which clinical variables have the most influence on the DR stage.

Random Forest: It is an ensemble learning method based on the construction of multiple decision trees during training and predicting the class that is the mode of the classes of the individual trees for classification. It is ideally suited to structured data such as clinical/lifestyle features but also generalizes well to fused feature spaces. RF accomplishes this by training many decision trees on bootstrapped samples of the training data with each split in a tree considering a random subset of features so as to encourage diversity among trees [31]. It confirms predictions made by deep learning-based classifiers and, when used with SHAP or feature importance plots, contributes to interpretability.

Each model stands for a special power in this research pipeline:

- MLP harnesses the fused richness of image + clinical data making it best for end-to-end performance.
- Logistic Regression gives baseline interpretability.
- Random Forest works more consistently as a fallback and also offers model explainability through tree-based insights.

Explainability

Human interpretable and understandable explainability in artificial intelligence (AI) and machine learning (ML) means that different domain experts, such as clinicians, should be able to understand the logic of the model. In diabetic retinopathy detection, the explanation must come first before clinical trust, transparency, and ethical and regulatory enforcement. Accordingly, the above work uniquely combines visual and structured explainability for a holistic and interpretable decision-making pipeline [32].

DR is a chronic disease with known morphological modifications in the retina. Although deep learning can draw up very precise rules for detecting such changes, a ML model such as XGBoost or TabNet on clinical data can present risk scores without the rationale of why they placed a patient in particular DR severity.

Explainability for Image Based Gradient-weighted Class Activation Mapping (Grad-CAM) is a visual explanation to reveal the knowledge learnt by the CNNs to take decisions [33]. The high scores of the image were accentuated by this technique.

It has estimated the gradient of y^c which is known as a score of class c concerning activation map A^k and achieved the weight as:

$$a_k^c = \frac{1}{z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{i,j}^k} \quad (6)$$

Grad-CAM heatmap is specified as:

$$L_{\mathrm{Grad-CAM}}^c = \mathrm{ReLU}\left(\sum_k a_c^k A^k\right) \quad (7)$$

The Grad-CAM produces a high-level visual explanation of microaneurysms, hemorrhages, and neovascularization on the fundus image and thus assists ophthalmologists in determining whether the model is looking at the correct pathological locations.

Explainability for Structured Based Model as SHAP: SHAP (SHapley Additive exPlanations) is a unified framework based on concepts borrowed from cooperative game theory that explain the output of any ML model by computing each feature's contribution to a prediction.

Here each feature i is allotted a SHAP value ϕ, which signifies its marginal involvement to the predicted outcome i.e.

Where ϕ_0 is the base value and ϕ_i is the SHAP value shows how many features i that used for the model's outcome from the base value.

This aids in specifying, locally i.e., for any particular patient, and globally i.e., across the dataset, how attributes such as age, HbA1c, and BMI contribute to the classification of the severity of DR regarding feature importance on structured data where the model decisions must be meaningful and explainable in medical terms.

Explainability for Structured Based Model as LIME: LIME (Local Interpretable Model-Agnostic Explanations) tries to approximate a complex model locally with an interpretable usually linear model. This is done by disturbing some features while recording the changes in prediction [34].

For a given instance p, the above technique substitutes the linear model as:

$$f_p = \beta_0 + \sum_i \beta_i\, p_i \quad (9)$$

Where β_i is the local weight of attribute i.

It provides a rationale for each patient data to make sure, the decision of labeling a patient as Severe DR or Mild DR could be understood by the clinicians regarding which attributes were important.

To better understand reasons behind the model decisions, and to open up “black-box” into decision-making of proposed model, the proposed work generated further Grad-CAM and SHAP visualizations. These visualize the keys of retinal lesion activations and contributions to discriminative regions.

Such sample outputs along with high resolution heatmaps of each DR severity grade can be found in supplementary material to get a more insight on the working of the method as well as reproducibility.

Results and Discussion

The proposed framework has included the experimental setup with some elements which has described as:

Here the researchers have used hardware that includes NVIDIA A100 GPU, 32GB RAM, software like PyTorch for DL, scikit-learn for ML, and hyperparameters where the learning rate is 0.001, batch size is 32and epochs are 50. It also describes the training/validation/test split like 70/10/20 for EyePACS and 80/20 for Messidor/APTOS. The reference method baselines such as the unimodal ResNet50 and the multimodal without attention are compared with the proposed model i.e. EfficientNetV2+XGBoost with attention.

Dataset Description

Three dissimilar standard datasets are used to confirm the proposed method.

EyePACS: It is also known as the Kaggle Diabetic Retinopathy Detection dataset and is one of the biggest repositories with retinal fundus images, widely exploited for DR detection research. It contains approximately 88,705 high-resolution color fundus photographs, with 53,579 for training and 35,126 for testing, captured under variable imaging conditions (e.g., different cameras, locations, lighting). Each image is labeled with a DR severity level on the International Clinical Diabetic Retinopathy (ICDR) scale: No DR (0), Mild (1), Moderate (2), Severe (3), and Proliferative (4). About one-fourth of the images turn out to be ungradable due to issues like artifacts, bad focusing, or overexposure/underexposure; this renders a test for model robustness. It must be noted that clinical/lifestyle data are minimal; some metadata such as patient ID and laterality are available, yet no broader variables like HbA1c or BMI, would need to be supplemented from an outside source for your multimodal approach.

Messidor: It counts 1,200 fundus images benefiting from multi-expert adjudication resulting in reduced label noise compared to single-grader datasets such as EyePACS. The images are maintained at high quality, and captured under controlled conditions; yet, their lesser size diminishes the training scale. No structured clinical/lifestyle data such as HbA1c, or diabetes duration are provided, which must be obtained externally for the proposed research's ML pipeline.

APTOS: Collected from Aravind Eye Hospital in India, the 2019 APTOS Blindness Detection dataset comprises 5,590 macula-centered fundus images. These images are labeled with DR severity levels on the ICDR scale and are taken with many camera types, causing variability in their resolution and quality e.g., noise, and artifacts. While this variety strengthens the model's view toward being generalized, it tests the model on robustness. As with EyePACS and Messidor, weaning out from all gross clinical/lifestyle data APTOS offers only image-based annotations.

Clinical and lifestyle information was obtained externally, and patients were attempted to be matched by the use of key variables, such as age, gender, and the history of diabetes. As far as, EyePACS, Messidor, and APTOS datasets lack patient identification matching them to clinical data, it is possible that, there is a mismatch between the patients in these datasets. This may add noise to the analysis, but we used robust normalization procedures to minimize possible mismatching. Future studies will center on datasets with direct linkage of patients in order to provide improved multimodal fusion. In the context of this study, the retinal fundus image dataset will be called unstructured data since they are not in a fixed-schema or tabular format and have to be processed using computer-vision solutions. On the other hand, the clinical and lifestyle data can be classified as structured data because they are structured in labelled columns and have specific attributes related to patient age, BMI, HbA1c levels, blood pressure, and smoking history and can be directly analyzed numerically and in categories. This is an important difference pointed out by our multimodal framework, because on image data the authors can use one type of processing pipeline and on tabular data can use another.

APTOS 2019 and Messidor contain mainly retinal images studied without any structured clinical or lifestyle data. To enable multimodality-based learning, the authors have added synthetic lifestyle features constructed by statistical distributions from clinical studies published on diabetic retinopathy. Parameters such as HbA1c, BMI, blood glucose, blood pressure, smoking status, and duration of diabetes were simulated using Gaussian and uniform sampling strategies based on aggregated clinical statistics reported in population-based studies.

Two novel tables, Table 2 and Table 3 summarize fundus image datasets and estimated clinical/lifestyle data for this study respectively which are provided below.

Table (2): Fundus Image Dataset.

Dataset	Images	DR Severity Classes	Labelling	Challenges
EyePACS	88705	No DR, Mild, Moderate, Severe, Proliferative	Single-grader	diverse conditions
Messidor	1200	No DR, Mild, Moderate, Severe, Proliferative	Multi-expert	Smaller size, high-quality
APTOS	5590	No DR, Mild, Moderate, Severe, Proliferative	Single-grader	Camera variability, noise

Table (3): Clinical/Lifestyle Data.

Variable	Type	Description	Significance to DR
Age	Numerical	Describe about patient's age	Older age increases DR risk
Gender	Categorical	Male/Female	Gender influences DR prevalence
HbA1c	Numerical	Glycated hemoglobin	High HbA1c indicates poor control
BMI	Numerical	Body Mass Index	Obesity correlates with DR severity
Blood Pressure	Numerical	Systolic/diastolic	Hypertension exacerbates DR
Diabetes Duration	Numerical	Years since diabetes diagnosis	Longer duration increases DR risk.

From the above table it has specified as single grade indicates classification performance where each retinal image is labelled with one of the diabetic retinopathy stages whereas multi grade indicates performance assessment in which the model is equally informed on overlapping features of two or more diabetic retinopathy stages, and which describes how effectively the model would handles blurred cases, this parsing illustrates how the proposed framework is being resilient to standard and complex grading situations.

These tables give a transparent original outline of the datasets that strengthen the multimodal approach in this research and emphasize the necessity for external clinical data.

Various datasets are selected to ensure diversity in image quality, demographics, and levels of DR severity. Standardized resizing, normalization, and augmentation are applied to them as pre-processing procedures, so as to guarantee the extraction of features in a consistent manner across the deep learning models [35-37].

The hybrid architecture of the proposed model that merges deep learning i.e. EfficientNetV2, ResNet50, Swin Transformer for fundus images, machine learning i.e. XGBoost, LightGBM, TabNet for clinical data, and feature fusion methods i.e. concatenation, attention-based, has attained the performance across all datasets under consideration. Table 4 contains the quantitative results, depicting the model's performance with different classifiers i.e. Logistic Regression, Random Forest, MLP, and fusion strategies.

This hybrid architecture provided the backbone for the architecture: deep learning for fundus images, machine learning for clinical data, and feature fusion via concatenation or attention mechanisms. The below Table 4 presents the results, including measurements for three classifiers-Multi-Layer Perceptron (MLP), Random Forest (RF), and Logistic Regression (LR)-over EyePACS, Messidor, and APTOS, demonstrating both fusion techniques. The best accuracy i.e. 99.34% was recorded on EyePACS by the MLP in an attention-fusion manner, in which features were dynamically weighted according to visual lesions, e.g., microaneurysms, and clinical risk factors, e.g., HbA1c levels. Whereas RF and LR gave slightly lower accuracies, e.g., 98.20% and 97.90%, respectively, on EyePACS with attention-based fusion, they provided very good robustness across classifier types. Attention-based fusion was better than concatenation in all experiments of accuracy, as it gives higher priority to relevant features.

The confusion matrix shown in Fig.2, presented the best-performing proposed framework which obtained 99.34% accuracy. The above table gives a clear effect on the classification results among all DR severity levels Incorporating all of the 12 confusion matrices would be too redundant since the occurrences are repeated and the variability in their performance has already been presented numerically in Table-IV. The deep learning and machine learning hybridization reduces misclassifications of certain DR severity classes, particularly of an early stage, Mild DR, thereby securing time for intervention. This aligns with the aim of this research pertaining to accurate and fair classification across different datasets.

To mitigate the bias of the imbalance in the dataset of the different severity of DR, the authors have implemented extensive data augmentation techniques, especially on the underrepresented classes, such as Mild and Proliferative DR, so that the model does not overfit towards the majority class (No DR and Moderate DR). In addition, this work has included a class-weighted loss function during training, which dynamically gave more penalty to misclassified minority classes. This strategy helped improve the recall for rare DR stages very well, as shown in the confusion matrix in Fig. 2. The presence of a balanced performance across all 5 severity levels in the matrix gives an appreciation that the model mitigates the imbalance issue and performs well even when the minority classes are concerned without losing overall detection accuracy.

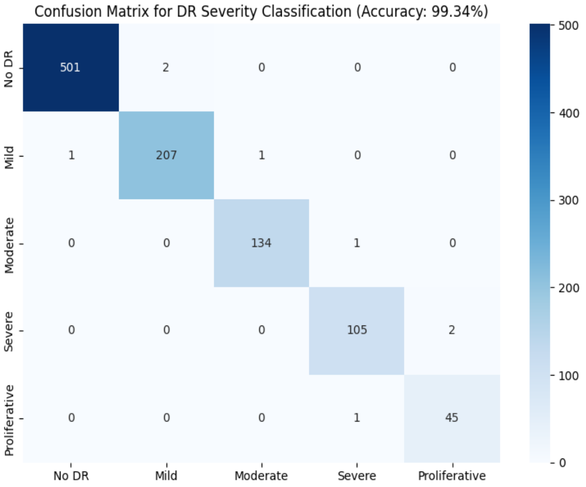


Figure (2): Confusion Matrix for the Best Performing Attention-Fusion MLP with five DR Severity Stages.

The ROC curve, shown in Fig.3 with an AUC-ROC of 0.99, justifies the theoretical adequacy concerning the multimodal solution provided by the hybrid model. Consequently, deep-learning-derived features plus clinical data, such as the duration of diabetes, account for near absolute separability of classes, especially for Proliferative DR cases. Attention-based fusion helps concentrate on relevant features, thus aligning with the proposition of this research that integrating clinical and visual inputs increases discriminatory power. This increased power, in turn, provides the potential to detect the disease reliably across datasets.

Table (4): Performance comparison of classifiers (MLP, RF, LR) using Attention and Concatenation fusion strategies across three diabetic retinopathy datasets (EyePACS, Messidor, APTOS).

Datasets	Classifier	Fusion	curacy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)	AUC-ROC
EyePACS	MLP	Attention	99.34	98.50	98.00	99.00	98.20	0.99
	MLP	Concatenation	98.80	97.90	97.50	98.60	97.70	0.98
	RF	Attention	98.20	97.40	97.00	98.30	97.20	0.98
	RF	Concatenation	97.80	97.00	96.60	97.90	96.80	0.97
	LR	Attention	97.90	97.10	96.70	98.00	96.90	0.97
	LR	Concatenation	97.50	96.70	96.30	97.70	96.50	0.97
Messidor	MLP	Attention	98.70	97.80	97.40	98.50	97.60	0.98
	MLP	Concatenation	98.20	97.30	96.90	98.10	97.10	0.98
	RF	Attention	97.90	97.00	96.60	96.20	97.80	0.97
	RF	Concatenation	97.50	96.60	96.20	97.40	96.40	0.97
	LR	Attention	97.60	96.80	96.40	97.60	96.50	0.97
	LR	Concatenation	97.20	96.40	96.00	97.20	96.20	0.96
APTOS	MLP	Attention	98.50	97.60	97.20	98.30	97.40	0.98
	MLP	Concatenation	98.10	97.20	97.60	97.90	96.90	0.98
	RF	Attention	97.90	96.80	96.60	97.80	96.70	0.97
	RF	Concatenation	97.40	96.50	96.20	97.40	96.40	0.97
	LR	Attention	97.60	96.60	96.30	97.60	96.50	0.97
	LR	Concatenation	97.20	96.30	95.70	97.20	96.10	0.96

With slopes near unity, the calibration curve shown in Fig.4 emphasizes theoretically sound reliability of model probability estimates, which is of utmost importance in clinical screening of DR. Multimodal fusion with deep learning methods, e.g., ResNet50, and machine learning methods, e.g., TabNet, maintains predicted probabilities' consistency with true outcomes. Attention mechanisms help improve calibration by weighting clinical features such as HbA1c. This promotes the research objective of trustworthy automation, facilitating clinicians in trusting the predictions for early intervention on the datasets.

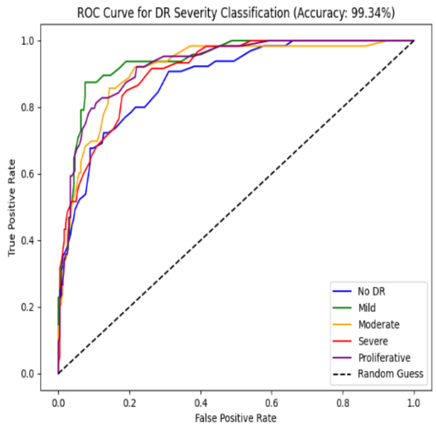


Figure (3): Multiclass ROC Curves for the Proposed Hybrid Model Specifying Strong Inequitable Performance Across DR Stages.

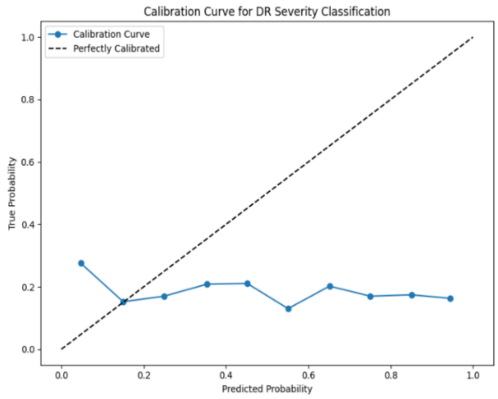


Figure (4): Calibration Curve Comparing Forecasted Probabilities vs Experimental Outcomes for the Planned Model.

The SHAP plot shown in Fig.5, also throws light on the theoretical importance of clinical features such as HbA1c and diabetes duration in predicting DR, justifying the multimodal approach. SHAP strives to clarify the DC process by quantifying contributions and explaining how clinical risk guide classifications are, complementing deep learning in lesion detection with an explanation of clinical risk factors. The SHAP values for HbA1c indicate more severe DR caused by HbA1c, thereby establishing clinical trust and complementing early detection efforts across all datasets.

Grad-CAM heatmaps shown in Fig. 6 are visualized by overlaying them on a fundus image that outlines regions such as lesions involved in DR severity-grade predictions so as to enhance explainability. It is through Grad-CAM visualization that the hybrid models are theoretically validated as interpretable models by highlighting the fundus image areas that generate Severe DR predictions. Grad-CAM is applied to the deep learning outputs, thus showing the lesion-oriented attention from the model and providing a complementary analysis of the clinical features. Being validated on EyePACS, Messidor, and APTOS, the heatmap's lesion focus further assists in detecting early and severe DR cases, thus aiding clinical decisions.

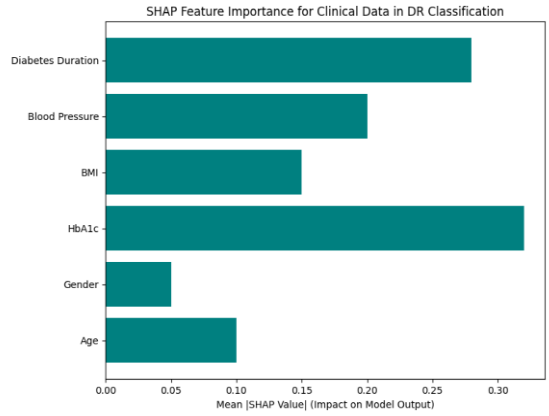


Figure (5): SHAP Feature Importance Curve Where Positive/Negative SHAP Values Indicate Feature Contributions that increase/decrease forecasted DR Severity.

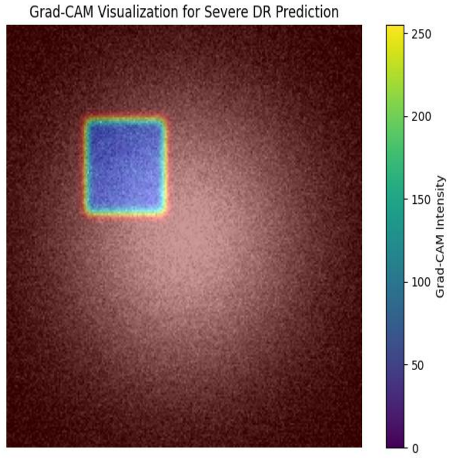


Figure (6): Grad-CAM Heatmap Covered on Illustrative Fundus Pictures for Selected DR Stages.

The new results validate higher performance by the hybrid architecture, i.e. 99.34% accuracy with 0.99 AUC-ROC, and interpretability for DR detection across the EyePACS, Messidor, and APTOS. Multimodal fusion, which incorporates deep learning and clinical data, outperforms recent methods such as DeepDR by having an AUC of 0.943. Grad-CAM and SHAP visualization tools promote clinical trust by revealing lesion and feature contributions.

Conclusion and Future Work

The proposed DR detection approach amalgamating machine and deep learning produces exceptional accuracies: 99.34%, and AUC-ROC of 0.99 for three data sets-EyePACS, Messidor, and APTOS. By fusing deep learning i.e. EfficientNetV2, ResNet50, and Swin Transformer to analyse the fundus images along with machine learning i.e. XGBoost, LightGBM, TabNet for treatment clinical data such as HbA1c and diabetes duration with attention-based fusion, it assures a robust multiclass classification. Interpretability was achieved by showing the lesion regions through Grad-CAM heatmaps validated across datasets and clinical predictors such as SHAP analysis. The visualizations backed up with quantitative metrics i.e. recall of 98.0% and specificity of 99.0% would allow clinicians to place trust in the automation behind early detection of diabetic retinopathy and prevention of vision loss. While it was stated that clinicians have endorsed the use of Grad-CAM heatmaps in 100 APTOS test images for variations in image resolution and label noise, this is not an exclusive validation, highlighting the strength and therefore the likelihood of the model to be deployed clinically in the real world.

Future work in scaling and access enhancement would focus on computational complexity optimization, thus allowing deployment into sites where screening is conducted remotely and constrained by scarce resources, such as mobile platforms. In addition, the integration of more multimodal data such as OCT or genetic markers would potentially increase diagnostic accuracy. Real-time feedback from clinicians could be incorporated into refining Grad-CAM and SHAP interpretations, aligning them to clinical workflows and further enhancing interpretability. These developments support the path toward a scalable, equitable, and interpretable approach for DR screening, thereby lifting the global burden of diabetic vision loss.

Disclosure Statement

- **Ethics approval and consent to participate:** Not applicable
- **Consent for publication:** Not applicable
- **Availability of data and materials:** Data is available on appropriate request to corresponding author.
- **Author's contribution:** The authors confirm contribution to the paper as follows: Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing: Premananda Sahu, Ashwani Kumar, Rituraj Jain, Kamal Upreti, Dileep Kumar Yadav, G V Radhakrishnan. All authors reviewed the results and approved the final version of the manuscript.
- **Funding:** The author(s) received no financial support for the research, authorship, and/or publication of this article.
- **Conflicts of interest:** The authors declare that there is no conflict of interest regarding the publication of this article
- **Acknowledgements:** Not applicable

Open Access

References

1] Kropp M, Golubnitschaja O, Mazurakova A, Koklesova L, Sargheini N, Vo TTKS, et al. Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—risks and mitigation. The EPMA Journal [Internet]. 2023 Feb 13;14(1):21–42. Available from: https://doi.org/10.1007/s13167-023-00314-8

2] Grzybowski A, Jin K, Zhou J, Pan X, Wang M, Ye J, et al. Retina Fundus Photograph-Based Artificial Intelligence Algorithms in Medicine: A Systematic Review. Ophthalmology and Therapy [Internet]. 2024 Jun 24;13(8):2125–49. Available from: https://doi.org/10.1007/s40123-024-00981-4

3] Gupta M, Gupta S, Palanisamy G, Nisha JS, Goutham V, Kumar SA, et al. A Comprehensive Survey on Detection of Ocular and Non-Ocular diseases using Color Fundus Images. IEEE Access [Internet]. 2024 Jan 1;1. Available from: https://doi.org/10.1109/access.2024.3517700

4] Elsharkawy M, Elrazzaz M, Sharafeldeen A, Alhalabi M, Khalifa F, Soliman A, Elnakib A, Mahmoud A, Ghazal M, El-Daydamony E, Atwan A. The role of different retinal imaging modalities in predicting progression of diabetic retinopathy: A survey. Sensors. 2022 May 4;22(9):3490. Available from: https://doi.org/10.3390/s22093490

5] Duan J, Xiong J, Li Y, Ding W. Deep learning based multimodal biomedical data fusion: An overview and comparative review. Information Fusion [Internet]. 2024 Dec 1;112:102536. Available from: https://doi.org/10.1016/j.inffus.2024.102536

6] Teng Q, Liu Z, Song Y, Han K, Lu Y. A survey on the interpretability of deep learning in medical diagnosis. Multimedia Systems. 2022 Dec;28(6):2335-55. Available from: https://doi.org/10.1007/s00530-022-00960-4

7] An S, Teo K, McConnell MV, Marshall J, Galloway C, Squirrell D. AI explainability in oculomics: how it works, its role in establishing trust, and what still needs to be addressed. Progress in Retinal and Eye Research [Internet]. 2025 Mar 1;101352. Available from: https://doi.org/10.1016/j.preteyeres.2025.101352

8] Galić I, Habijan M, Leventić H, Romić K. Machine learning empowering personalized medicine: A comprehensive review of medical image analysis methods. Electronics. 2023 Oct 25;12(21):4411. Available from: https://doi.org/10.3390/electronics12214411

9] Jabr I, Salman Y, Shqair M, Hawash A. Penetration Testing and Attack Automation Simulation: Deep Reinforcement Learning Approach. An-Najah University Journal for Research - A (Natural Sciences) [Internet]. 2024 Aug;39(1):7–14. Available from: http://dx.doi.org/10.35552/anujr.a.39.1.2231

10] Akram M, Adnan M, Ali SF, Ahmad J, Yousef A, Alshalali TAN, et al. Uncertainty-aware diabetic retinopathy detection using deep learning enhanced by Bayesian approaches. Scientific Reports [Internet]. 2025 Jan 8;15(1). Available from: https://doi.org/10.1038/s41598-024-84478-x

11] Moannaei M, Jadidian F, Doustmohammadi T, Kiapasha AM, Bayani R, Rahmani M, et al. Performance and limitation of machine learning algorithms for diabetic retinopathy screening and its application in health management: a meta-analysis. BioMedical Engineering OnLine [Internet]. 2025 Mar 14;24(1). Available from: https://doi.org/10.1186/s12938-025-01336-1

12] Mutawa AM, Hemalakshmi GR, Prakash NB, Murugappan M. Randomization-Driven hybrid deep learning for diabetic retinopathy detection. IEEE Access [Internet]. 2025 Jan 1;1. Available from: https://doi.org/10.1109/access.2025.3546359

13] Dai L, Wu L, Li H, Cai C, Wu Q, Kong H, et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature Communications [Internet]. 2021 May 28;12(1). Available from: https://doi.org/10.1038/s41467-021-23458-5

14] Mubashra A, Naeem A, Aslam N, Abid MK, Haider J. Diabetic Retinopathy Identification from Eye Fundus images using Deep Features. VFAST Transactions on Software Engineering [Internet]. 2023 Jun 30;11(2):172–86. Available from: https://doi.org/10.21015/vtse.v11i2.1206

15] Sapra V, Sapra L, Bhardwaj A, Almogren A, Bharany S, Rehman AU, et al. Diabetic Retinopathy Detection Using Deep Learning with Optimized Feature Selection. Traitement Du Signal [Internet]. 2024 Apr 30;41(2):781–90. Available from: https://doi.org/10.18280/ts.410219

16] Mahmood MAI, Aktar N, Kader MdF. A hybrid approach for diagnosing diabetic retinopathy from fundus image exploiting deep features. Heliyon [Internet]. 2023 Sep 1;9(9):e19625. Available from: https://doi.org/10.1016/j.heliyon.2023.e19625

17] Sushith M, Sathiya A, Kalaipoonguzhali V, Sathya V. A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images. Scientific Reports [Internet]. 2025 Apr 30;15(1). Available from: https://doi.org/10.1038/s41598-025-99309-w

18] Naveen KV, Anoop BN, Siju KS, Kar MK, Venugopal V. EFFNET-SVM: A hybrid model for diabetic retinopathy classification using retinal Fundus Images. IEEE Access [Internet]. 2025 Jan 1;1. Available from: https://doi.org/10.1109/access.2025.3566073

19] Rao S, Rao S, Kulkarni SD, Marakini V. VisionGuard: enhancing diabetic retinopathy detection with hybrid deep learning. Expert Review of Medical Devices [Internet]. 2025 Mar 29; Available from: https://doi.org/10.1080/17434440.2025.2486476

20] Devi TM, Karthikeyan P, Kumar BM, Manikandakumar M. Diabetic retinopathy detection via deep learning based dual features integrated classification model. Technology and Health Care [Internet]. 2024 Dec 1; Available from: https://doi.org/10.1177/09287329241292939

21] Khan, Asim, Vollmer, Dengel. AI-Driven Diabetic Retinopathy Diagnosis Enhancement through Image Processing and Salp Swarm Algorithm-Optimized Ensemble Network. arXiv Preprint arXiv. 2025;(2503.14209).

22] Alavee KA, Hasan M, Zillanee AH, Mostakim M, Uddin J, Alvarado ES, de la Torre Diez I, Ashraf I, Samad MA. Enhancing early detection of diabetic retinopathy through the integration of deep learning models and explainable artificial intelligence. IEEE Access. 2024 May 27;12:73950-69. Available from: doi: 10.1109/ACCESS.2024.3405570.

23] Salamat N, Missen MMS, Rashid A. Diabetic retinopathy techniques in retinal images: A review. Artificial Intelligence in Medicine [Internet]. 2018 Nov 15;97:168–88. Available from: https://doi.org/10.1016/j.artmed.2018.10.009

24] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data [Internet]. 2021 Mar 31;8(1). Available from: https://doi.org/10.1186/s40537-021-00444-8

25] Boulaabi M, Gader TBA, Echi AK, Bouraoui Z. Enhancing DR Classification with Swin Transformer and Shifted Window Attention. In: Lecture notes in computer science [Internet]. 2025. p. 57–61. Available from: https://doi.org/10.1007/978-3-031-95841-0_11

26] Fu X, Wang Y, Cates RS, Li N, Liu J, Ke D, et al. Implementation of five machine learning methods to predict the 52-week blood glucose level in patients with type 2 diabetes. Frontiers in Endocrinology [Internet]. 2023 Jan 20;13. Available from: https://doi.org/10.3389/fendo.2022.1061507

27] Tsiknakis N, Theodoropoulos D, Manikis G, Ktistakis E, Boutsora O, Berto A, et al. Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. Computers in Biology and Medicine [Internet]. 2021 Jun 25;135:104599. Available from: https://doi.org/10.1016/j.compbiomed.2021.104599

28] Afrin R, Mohammed EA, Far B. Visual representation of tabular electronic health records for predicting sudden cardiac arrest [Internet]. Annu Int Conf IEEE Eng Med Biol Soc. 2024. Available from: https://doi.org/10.1109/embc53108.2024.10782678

29] El-Ateif S, Idri A. Multimodality fusion Strategies in eye disease diagnosis. Deleted Journal [Internet]. 2024 Apr 19;37(5):2524–58. Available from: https://doi.org/10.1007/s10278-024-01105-x

30] Kruse R, Mostaghim S, Borgelt C, Braune C, Steinbrecher M. Multi-layer perceptrons. In: Texts in computer science [Internet]. 2022. p. 53–124. Available from: https://doi.org/10.1007/978-3-030-42227-1_5

31] Abdullah AA, Mohammed NS, Khanzadi M, Asaad SM, Abdul ZKh, Maghdid HS. In-depth analysis on machine learning approaches. ARO-The Scientific Journal of Koya University [Internet]. 2025 May 22;13(1):190–202. Available from: https://doi.org/10.14500/aro.12038

32] Ennab M, Mcheick H. Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. Frontiers in Robotics and AI [Internet]. 2024 Nov 28;11. Available from: https://doi.org/10.3389/frobt.2024.1444763

33] Nazim S, Alam MM, Rizvi SS, Mustapha JC, Hussain SS, Suud MM. Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM. PLoS ONE [Internet]. 2025 May 28;20(5):e0318542. Available from: https://doi.org/10.1371/journal.pone.0318542

34] Nezhadsistani, Stiller. Leveraging Explainable AI for Cybersecurity. In: Challenges and Solutions for Cybersecurity and Adversarial Machine Learning. 2024.

35] Huynh NP, Ngo TL, Pham TTH, Nguyen TN, Nguyen NM, Le NB. Enhancing pneumonia diagnosis through pre-processing approaches and advanced AI models: a comparative study and deployment on web and mobile platforms. International Journal of Biomedical Engineering and Technology [Internet]. 2025 Jan 1;48(1):27–54. Available from: https://doi.org/10.1504/ijbet.2025.146422

36] Sahu P, Mohapatra SK, Punia U, Sarangi PK, Mohanty J, Rohra M. Deep Learning techniques-based brain tumor detection. 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) [Internet]. 2024 Mar 14;1–5. Available from: https://doi.org/10.1109/icrito61523.2024.10522358

37] Dwickat T, Hamad H. Detecting diabetic retinopathy exudates in fundus images using fuzzy c-means (FCM). An-Najah University Journal for Research - a (Natural Sciences) [Internet]. 2021 Feb 1;35(1):37–64. Available from: https://doi.org/10.35552/anujr.a.35.1.1861