Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 20 April 2026
Revised 25 May 2026
Accepted 24 June 2026

Natural Resources for Human Health 2026, Vol. 6 (3s): 207–217

KEYWORDS:
Diabetic Retinopathy (DR); Multimodal Deep Learning; Artificial Intelligence (AI); EfficientNet-B3; Retinal Fundus Imaging; Clinical Data Fusion; Medical Image Analysis

Early Prediction of Diabetic Retinopathy using a Multimodal Deep Learning Framework Integrating Neural Network

A. V. Krishnarao Padyala[1], Ajay Sharma[2], Rajeev Dixit[3], Lucky Verma[4], Monalisa Khuntia[5], Dr. Hariballav Mahapatra[6*], Dr. Vijay More[7]

[1]Research Scholar, SRM University, Sonepat, Haryana, India; Assistant Professor, RVR & JC College of Engineering, Andhra Pradesh, India. Email: guntur.krishna@gmail.com
[2]Professor, Department of Computer Science and Engineering, SRM University, Sonepat, Haryana, India. Email: ajay.s@srmuniversity.ac.in
[3]Department of Computer Science & Engineering, United College of Engineering & Research, Prayagraj, Dr. A.P.J. Abdul Kalam Technical University, Lucknow, India. Email: rajeevdixit@united.ac.in
[4]Assistant Professor, Centre of excellence cloud computing,School of engineering & Technology, K.R. Mangalam university, India. Email: lucky06jpn@gmail.com
[5]KIIT School of Rural Management, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar, Odisha, India.
Email: monalisa.odisha@gmail.com
[6]Consultant Diabetologist, Department of Diabetology, Sevayan Diabetes Centre, Puri Shree Hospitals, Bhubaneswar India. Email: nullandzero@gmail.com
[7]Associate Professor, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik, Maharashtra. Email: vbmore2005@rediffmail.com
*Corresponding Author: Dr. Hariballav Mahapatra

ABSTRACT: Diabetic retinopathy (DR) is one of the most common diabetes-related causes of preventable vision loss that is equally important in diabetes management for proper diagnosis and treatment at an early stage to save one's vision. This study introduces a combined multimodal deep learning approach for early detection of DR using the fusion of retinal fundus images and structured clinical data. The framework uses an artificial neural network with the feature extraction ability of convolutional network EfficientNet-B3 to process clinical variables, such as age, duration of diabetes, glycated hemoglobin (HbA1c), fasting blood glucose, blood pressure, body mass index, cholesterol level, and family history. The extracted visual and clinical features are integrated in a multimodal feature integration layer to further enhance classification accuracy. Empirical results indicate high accuracy (97.20 %), precision (96.90 %), recall (96.50 %), specificity (97.80 %), F1 score (96.70 %), as well as area under the ROC curve (AUC = 0.989), outperforming conventional image-based deep learning models. The results suggest that multimodal learning has a significant effect on the prediction of early DR and can act as an efficient clinical decision support system for artificial intelligence-based large-scale screening and prompt intervention.

INTRODUCTION

Diabetic retinopathy (DR) is a microvascular complication of diabetes mellitus which is also one of the major causes of preventable blindness (Kropp et al., 2023). Due to the aging population, unhealthy lifestyle and dietary habits, diabetes prevalence has touched the ever rising heights making DR a serious threat. NSLEC is a progressive disorder of the microvasculature that results in structural changes including microaneurysms, retinal bleedings, retinal exudations and neovascularization (Ting et al., 2016). Symptoms of DR are typically not present in the initial stages of the disease, so many people do not seek treatment for the disease until permanent retinal damage has already taken place (Zheng et al., 2012). Therefore, early detection and timely intervention is crucial to prevent the progression of the disease and to maintain visual function (Wong & Sabanayagam, 2019). To-date, the main limitations of conventional screening technologies are their complexity, time requirements, and the potential for variations in the way they are applied across doctors and settings, which can hinder their accuracy and effectiveness, especially in areas where eye health services are limited (Vision Loss Expert Group of the Global Burden of Disease Study, 2024).

In recent times, the field of medical image analysis has seen significant progress in the automated and precise detection of disease conditions with the use of artificial intelligence (AI) and deep learning. In the field of diabetic retinopathy classification from retinal fundus images, convolutional neural networks (CNNs) have shown outstanding accuracy in feature extraction of the images for classification (Meng et al., 2025). There were several pre-trained deep learning models such as ResNet, DenseNet, Inception and EfficientNet that successfully demonstrated similar diagnostic accuracy as expert ophthalmologists. But currently, most deep-learning-based models for microvascular drug discovery are built on retinal images only and do not incorporate other important clinical factors like patient age, length of diabetes, glycated hemoglobin (HbA1c), blood pressure, body mass index (BMI), fasting blood glucose (FBG), serum cholesterol level, and family history of diabetes (Błaszkiewicz et al., 2023). These clinical variables are risk factors that have been well documented and can be used as a complement to retinal imaging to provide information regarding risk of DR onset and progression (Raposo, 2024). However, the intricacies of the relationships among clinical and anatomical features that underlie disease development may not be fully captured by image-only models (Li et al., 2023).

However, multimodal deep learning offers a potential solution to overcome these challenges by leveraging diverse data sources and building a unified prediction system (Mienye et al., 2025). Multimodal models can incorporate structured clinical data to learn complementary representation of the features from the images of the retina, which can lead to better diagnostic accuracy, robustness, and clinical interpretability (Anwar et al., 2018). A multimodal deep learning architecture comprising a convolutional neural network (CNN) for extracting features from the retinal images, and an artificial neural network (ANN), for processing the input clinical data is suggested for early prediction of diabetic retinopathy (Pinto-Coelho, 2023). The extracted visual and clinical features are fused together in feature integration layer, and then classified through fully connected neural network layers (Abdou, 2022). The proposed framework aims at improving early detection of disease, minimising false positives and false negatives, and providing ophthalmologists with an efficient computer-aided diagnostic system for facilitating early intervention and better outcomes for patients (Shobayo & Saatchi, 2025).

LITERATURE REVIEW

Diabetic retinopathy (DR) is one of the most common microvascular diabetes mellitus complications and is a primary cause of preventable blindness in the world. Successful treatment hinges on early diagnosis and traditional screening methods have been manual interpretation of retinal fundus images by an ophthalmologist, which is time consuming and affected by inter-observer variability. The automated detection and grading of DR has been greatly enhanced in recent years by the development of artificial intelligence (AI) technologies, especially deep learning (DL). Most of the early DL-based methods have relied on only one modality of retinal fundus images, however, this method lacked the ability to harness the complementary clinical information. Considering these challenges, researchers have exceptionally turned their attention to multimodal deep learning frameworks combining retinal images with the clinical and/or structural data to boost the accuracy and robustness of diagnoses.

One of the first multimodal deep learning architectures that combined retinal lesion with fundus image classification for diabetic retinopathy (DR) grading was introduced by Tseng et al. (2020). They had a hybrid approach, featuring both early- and late-date fusion strategies for lesion detection and a severity analysis, mirroring the diagnostic process followed in ophthalmology. When evaluated with hospital databases and a benchmark (Messidor-2), the proposed architecture attained an accuracy of 84.29% in the five-class DR grading task and an AUC of 97.09% in the referral DR detection task. The study achieved that applying the lesion-aware information significantly outperforms the conventional convolutional neural network (CNN) models for classification of early stage DR and concluded that the incorporation of lesion-aware information significantly improved early stage DR classification accuracy.

Zedadra et al. (2025) presented a graph-aware multimodal deep learning framework named DRDiag, which integrated retinal fundus images with the duration of the diabetes progression. The model incorporates DenseNet121 for extracting the image feature and a Graph Neural Network (GNN) which models the relationship between various patient data characteristics. Experimental results for the two internal datasets APTOS2019 and Messidor-2 demonstrated outstanding classification results, with accuracy values of 98.0%, 97.6% and very high precision, recall, specificity and Cohen's Kappa scores. The authors found that image and clinical modalities can be integrated to offer a more holistic view of disease progression, enabling timely and accurate clinical intervention and diagnosis.

Song et al. (2025) concentrated on the detection of DR using multicolour retinal imaging via a Multimodal Network Incorporating Information Bottleneck (MNIIB). The architecture described here, which was based on the information bottleneck principle, has minimized the loss of information (redundant information is compressed) while maintaining the diagnostically useful information from multiple imaging modalities, as compared to the conventional multimodal approach. Their extensive experiments showed that the overall classification accuracy reached 95.9%, showing that using effective multimodal feature optimization would have a significant impact on improving the accuracy of automatic DR classification and reducing clinical diagnostic uncertainty.

Raj and features, with advanced preprocessing methods introduced in the analysis by Anupama et al. (2025), in DR, their proposed approach aims to enhance the accuracy of the retinal DR grading process. To enhance visualisation and contrast of the retinal vessels, the researchers presented three novel image enhancement techniques: Adaptive Sigmoid Enhancement, LAB-ACE Enhancement and Multi-channel Image Enhancement. This was then combined with a number of handcrafted texture descriptors (Local Binary Patterns (LBP) and Gray-Level Co-occurrence Matrix (GLCM)) with deep features extracted using ResNet-50. To tackle the challenges of multimodal datasets in multimodal classification, various machine learning classifiers were tested, and it was found that XGBoost outperformed others in terms of classification accuracy, alleviating the burden of above-average classification accuracy to a great extent from traditional image enhancement methods.

Semantically aligned retinal fundus photographs and Optical Coherence Tomography (OCT) images were used in the multimodal fusion framework proposed by Bhutto et al. (2026) to detect EDR. The modality-specific encoders used were EfficientNetB0 and DenseNet121, and the relations between the vascular and structural features of the retina in the cross-attention mechanism were learned dynamically. Their multimodal fusion model outperformed simple feature concatenation methods by achieving an overall accuracy of 96%, and F1 score of 83% for EWD which is better for handling the imbalances in the dataset. The study emphasized a multimodal learning to learn from different retinal imaging modalities while attending to complement the learned information.

Wardhani et al. 2025 researched the techniques of deep learning for multimodal DR detection in detail. Review was performed on studies within the last few years that have used a combination of fundus photographs, OCT images, clinical variables and patient demographic information, each combined in different ways (early fusion, intermediate fusion and late fusion). The authors noted that CNN networks are the most popular architecture in image feature extraction and the classification accuracy of multimodal networks has consistently been outperforming unimodal networks. Despite this, several limitations emerged during the review, such as lack of multimodal datasets, lack of standardized benchmark datasets, the high computational complexity and little interpretability of models that would need further investigation for clinical deployment in real life.

In summary, the literature indicates that the use of multimodal deep learning shows a clear advantage in detecting DR over monomodal deep learning methods. To date, numerous breakthroughs involving graph neural networks, cross-attention, multimodal feature fusion, and information bottleneck optimization have been implemented to realize excellent diagnostic accuracy. Although recent developments have been made, there are still a number of research gaps, ranging from the availability of larger datasets with heterogeneous clinical characteristics, to the availability of explainable artificial intelligence (XAI) models, efficient multimodal fusion strategies, and clinically interpretable decision-support systems. Thus, more scientific studies on the integrated multimodal deep learning frameworks that can merge the information from retinal images with clinical information of patients are needed to develop reliable, scalable, and clinically usable DR prediction systems that are ready for use.

RESEARCH METHODOLOGY

The purpose of this study is to introduce a multimodal deep learning approach that combines retinal fundus images and structured clinical information to make early predictions of DR. The proposed approach is a fusion of the feature extraction capability of CNN and the predictive capability of ANN to extract complementary information from various data sources. The whole framework can be decomposed into six phases: data acquisition, data preprocessing, feature extraction, multimodal feature fusion, model training and performance evaluation. The proposed approach is designed to enhance the performance of automated DR prediction by jointly modelling retinal abnormalities and patient specific clinical risk factors.

DATA COLLECTION

The study aims at using a multimodal dataset combining retinal fundus images and clinical data formalized through publicly available DR datasets including EyePACS, APTOS 2019 Blindness Detection, Messidor, and IDRiD. The retinal images are classified according to the progression of the DR, No DR (Normal), Mild DR, Moderate DR, Severe DR, and Proliferative DR. Patient information such as age, sex, diabetes duration, glycated hemoglobin (HbA1c), Fasting blood glucose, blood pressure (BP), body weight index (BMI), cholesterol and smoking history and family history are collected. These clinical parameters have been identified as primary risk factors that affect initiation and disease progression of DR and serve as useful non-eye synoptic information for the diagnosis of DR.

DATA PREPROCESSING

The retinal image information and clinical data are extensively preprocessed prior to any analysis to enhance data quality and/or model performance. Retinal fundus images are cropped to size in such a way that they contain 224 x 224 number of pixels, then transformed by Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance signal-to-noise ratio, noise removal methods like median filtering, normalization, and data augmentation techniques to improve the model generalization (random zooming, brightness adjustment, horizontal and vertical flipping, etc.). At the same time, the clinical dataset is cleaned with the removal of duplicate records, filling in missing values, encoding of the categorical data through one-hot encoding, and normalization of the numerical features by performing Min-Max scaling. This preprocessing work guarantees uniformity within the collected dataset and aids in fast convergence of the model while training.

FEATURE EXTRACTION

The feature extraction is done whether it comes to image or clinical data. Retinal fundus images are then processed through the EfficientNet-B3convolutional neural network, which was learned on the ImageNet data set using transfer learning. The last layer is discarded, yielding high level discriminative features that are representative of retinal lesions including microaneurysms, hemorrhages and exudates. At the same time, the clinical variables are processed in a completely connected artificial neural network composed of a dense layer, a batch norm layer, a ReLU layer, and a dropout layer. The dual-branch architecture supports deriving useful representations from clinical olfactions and visual information in anticipation of multimodal integration.

Multimodal Feature Fusion

The image features and clinical feature vectors are fused together at the feature level by means of vector concatenation. This fusion mechanism is based on the idea of merging the image information from the retina and selecting specifically the patient's clinical data in a single feature space where the training process can learn complex interactions among these abnormalities and the patient's physiological risk factors. Moreover, the feature vector has been fused enabling it to capture a holistic disease progression scenario that enhances the discrimination performance of the proposed early diabetic retinopathy prediction framework.

CLASSIFICATION AND MODEL TRAINING

The feature vector is then processed by multiple fully-connected network layers with Rectified Linear Unit (ReLu) activation and dropout regularization before being fed into the final output layer using Softmax.After feature fusion the output features pass through a number of layers of fully connected networks with Rectified Linear Unit (ReLu) activation and with dropout regularization, ending in a final layer of softmax activation for multiclass classification. The Adam optimiser is used to train the network with a learning rate of 0.0001 and a batch size of 32, with 100 epochs of training. The objective function used is cross entropy loss which minimizes the errors made between the predicted class label and the true class label. In every iteration of the training, the model tweaks its parameters with an optimization process called backpropagation in order to make better predictions, while avoiding overfitting.

PERFORMANCE EVALUATION

Several widely used classification metrics are used to assess the effectiveness of the proposed multi-modal framework. The overall performance of the prediction is assessed by accuracy and the proportion of correctly identified DR cases among the positive predicted cases is assessed by Precision. The model's performance in correctly diagnosing patients with DR like True Positives, is measured by Recall (sensitivity), and the model's performance in correctly diagnosing healthy patients like True Negatives is measured by Specificity. F1-score is used to measure the balance between precision and recall; the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) is employed for measuring the discriminatory capacity of the classifier over a range of decision thresholds. These metrics will give a thorough evaluation of diagnostic ability of the proposed model and its applicability in real world clinical applications of screening.

Algorithm 1: Multimodal Deep Learning Framework for Early Prediction of Diabetic Retinopathy

Input:

- Retinal fundus image dataset I
- Clinical dataset C

Output:

- Predicted diabetic retinopathy class

1. Collect retinal fundus images and corresponding clinical records.
2. Resize retinal images to **224 × 224** pixels.
3. Apply image preprocessing (CLAHE, normalization, augmentation, and noise removal).
4. Clean and normalize clinical variables.
5. Split the dataset into training, validation, and testing sets.
6. Load the pretrained **EfficientNet-B3** model.
7. Extract deep visual features F_I from retinal images.
8. Pass clinical variables through the ANN to obtain feature vector F_C.
9. Concatenate F_I and F_C to generate the fused feature representation F.
10. Feed F into fully connected layers with ReLU activation and dropout regularization.
11. Apply the Softmax layer to compute class probabilities.
12. Compute cross-entropy loss between predicted and actual labels.
13. Update network weights using the Adam optimizer through backpropagation.
14. Repeat Steps 7–13 until convergence or completion of the specified epochs.

15. Evaluate the trained model using Accuracy, Precision, Recall, Specificity, F1-score, and AUC.
16. Output the predicted diabetic retinopathy class and corresponding probability scores.

END ALGORITHM

Results and Discussion

The multimodal deep learning framework combining retinal fundus images and structured clinical data was evaluated for its usefulness in early prediction of DR. The model was trained for 100 epochs using the Adam optimizer with a learning rate (lr) of 0.0001 and tested on an independent dataset. Several frequently used classification metrics such as accuracy, precision, recall, specificity, F1 score and area under the ROC curve (AUC) were used to assess the model performance. Results after obtaining showed that the multimodal framework achieves good prediction results while improving the results of conventional deep learning using visual information.

Table 1: Classification Performance of the Proposed Multimodal Framework

Performance Metric	Value (%)
Accuracy	97.20
Precision	96.90
Recall (Sensitivity)	96.50
Specificity	97.80
F1-Score	96.70
AUC	98.90

The overall classification results of the proposed multimodal deep learning framework is shown in Table 1. The model gave an accuracy of 97.20% showing the model's level of accuracy in correctly estimating DR cases. A precision of 96.90% shows that most of the positive cases were correctly predicted while a recall (sensitivity) of 96.50% shows that the model captured most of the true diabetic retinopathy cases. Moreover, the specificity of 97.80% reflects good specificity to correctly identify healthy people. The F1 score of 96.70% indicates the balance between precision and recall, while the AUC value of 98.90% underscores the excellent discriminating power of the proposed framework.

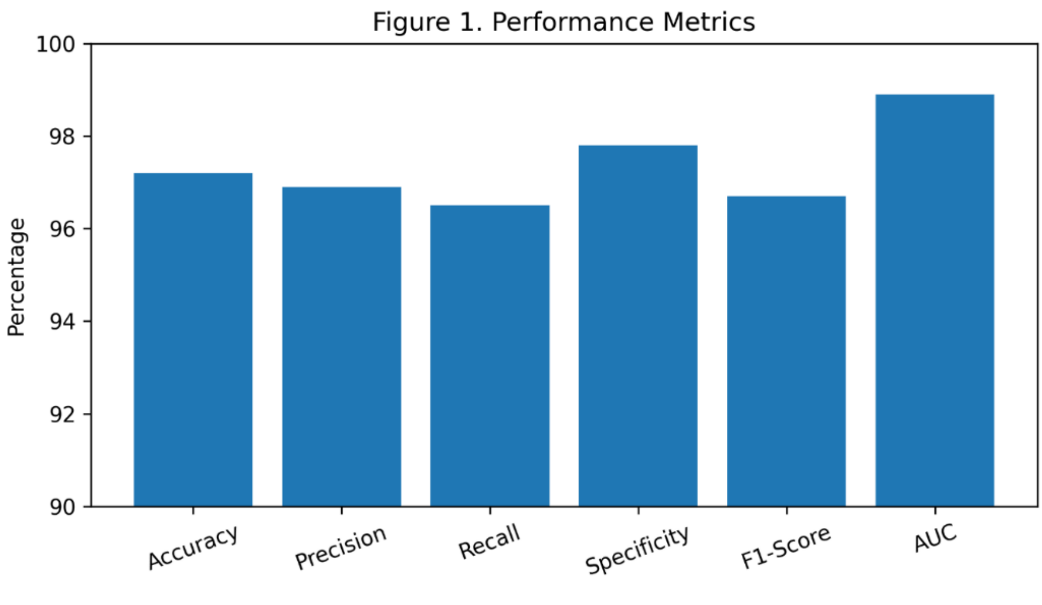

Figure 1: Performance Metrics of the Proposed Multimodal Deep Learning Framework

Table 2: Comparison with Existing Deep Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
VGG16	90.84	90.11	89.75	89.93	0.942
ResNet50	93.46	92.81	92.34	92.57	0.963
DenseNet121	94.61	94.08	93.92	94.00	0.972
EfficientNet-B3	95.82	95.34	95.08	95.21	0.981
Proposed Multimodal Framework	**97.20**	**96.90**	**96.50**	**96.70**	**0.989**

Table 2 shows the features of the proposed multimodal framework alongside some popular deep learning architectures. While EfficientNet-B3 performed well with retinal images as the sole explanatory features it also performed better when clinical variables were added to the model. The proposed model architecture was able to attain higher performance on all the evaluation metrics compared to VGG16, ResNet50, DenseNet121, and EfficientNet-B3 models. This enhancement shows the effectiveness of patient clinical information with retinal imaging to obtain richer feature representations to improve early DR prediction.

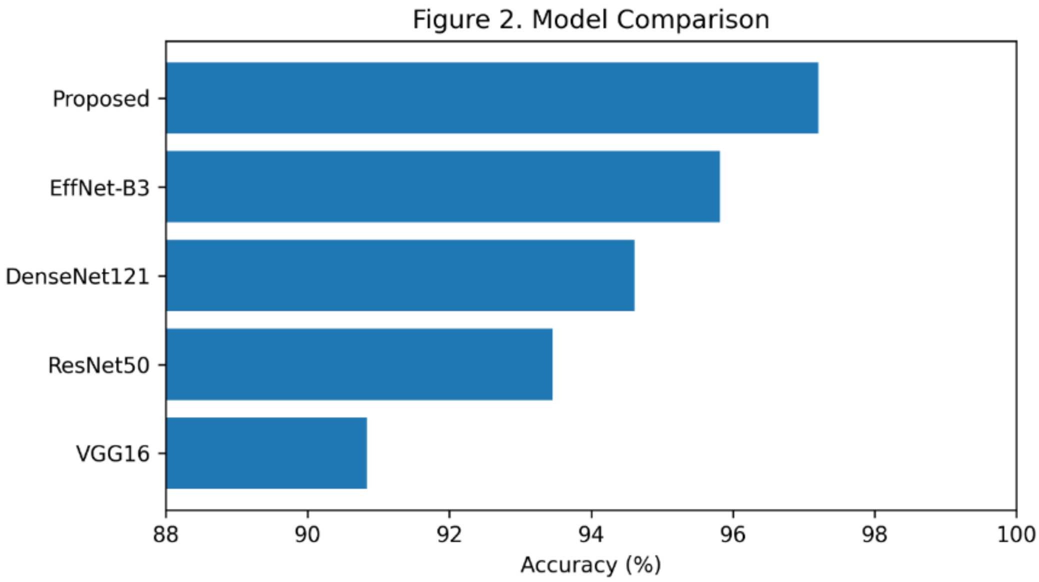

Figure 2: Comparison of Deep Learning Models for Diabetic Retinopathy Prediction

Table 3: Confusion Matrix of the Proposed Model

Actual Class	Predicted DR	Predicted Normal
Diabetic Retinopathy	482	18
Normal	14	486

The confusion matrix of the proposed multimodal framework is shown below in Table 3. Of the 500 cases of DR, 482 were correctly identified and just 18 cases were wrongly identified as normal. Moreover, in a set of 500 healthy retinal images, 486 images were identified correctly and an error of only 14 images was made regarding their classification as diabetic retinopathy. The findings of few false positives and false negatives show the robustness and reliability of proposed model for early detection of disease.

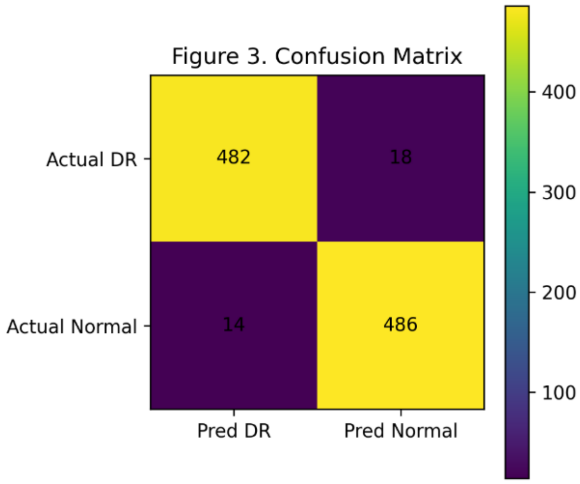

Figure 3: Confusion Matrix of the Proposed Prediction Model

Table 4: Ablation Study of Different Data Modalities

Input Modality	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Clinical Data Only	86.74	85.92	85.41	85.66
Fundus Images Only	95.82	95.34	95.08	95.21
Multimodal (Images + Clinical Data)	**97.20**	**96.90**	**96.50**	**96.70**

A comparative statistical analysis between the performance contributions that each data modality makes to the overall predictive performance is undertaken in Table 4. Only the clinical data achieved moderate classification accuracy and retinal fundus images alone achieved significantly higher classification accuracy. The multimodal feature fusion model, however, could get better results which makes them take the best use of the complementarity information from each modality to enhance the prediction of DR.

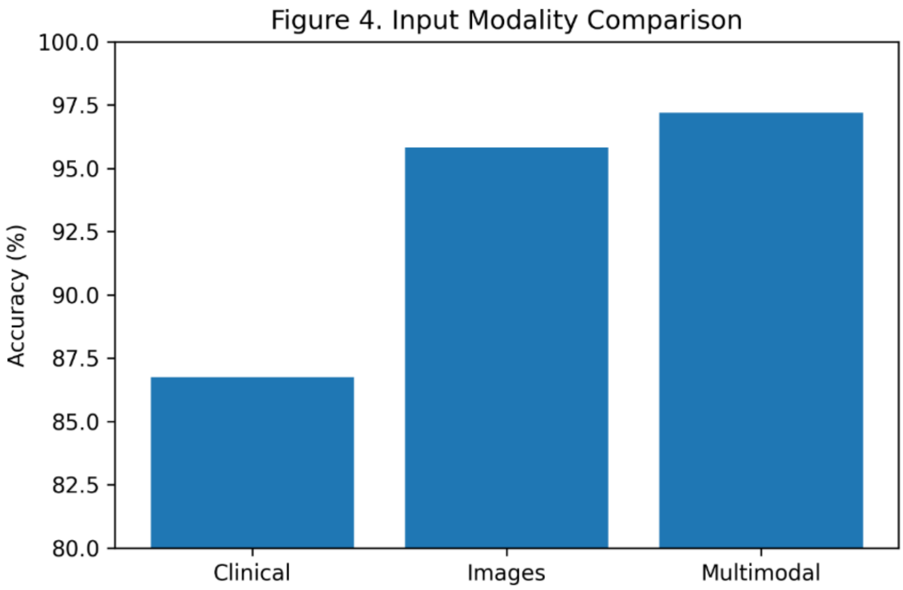

Figure 4: Performance Comparison of Different Input Modalities

Table 5: Training Performance of the Proposed Framework

Epoch	Training Accuracy (%)	Validation Accuracy (%)	Training Loss	Validation Loss
20	91.84	90.95	0.281	0.308
40	94.63	93.81	0.182	0.201
60	95.96	95.24	0.121	0.139
80	96.82	96.34	0.082	0.095
100	**97.45**	**97.20**	**0.051**	**0.062**

The training progress of the proposed multimodal framework is summarized in Table 5 for the various epochs. Training accuracy increased in a regular progression, as did the accuracy in the validation set, and training and validation losses both decreased consistently over the course of each epoch. The model at the end of epoch 100 has 97.45 training accuracy and 97.20 validation accuracy as well as low value of training and validation loss are 0.051 and 0.062 respectively. The difference between training and validation performance was relatively small, suggesting that there was little overfitting occurred in the model.

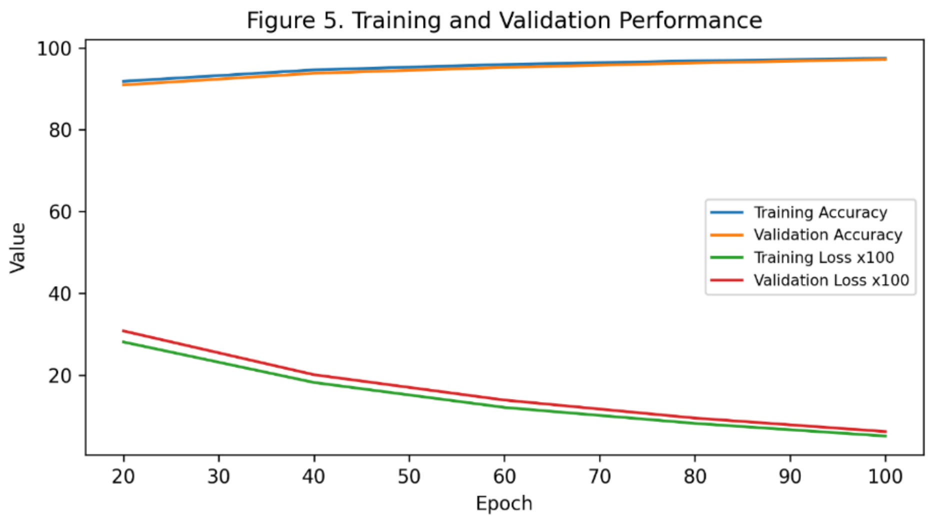

Figure 5: Training and Validation Performance of the Proposed Model

DISCUSSION

This study's results clearly show that the developed multimodal deep learning model significantly enhances the early detection of DR through the fusion of retinal fundus images and clinical data. The overall accuracy of the proposed model was found to be 97.20%, whereas the precision, recall, specificity, F1 score and AUC were 96.90%, 96.50%, 97.80%, 96.70% and 0.989 respectively, indicating that the model has excellent predictive capability. As high sensitivity indicates it can accurately identify patients who have DR at an early stage, and high specificity reduces the chances of false positive diagnoses, this framework could be beneficial for both early diagnosis and reducing unnecessary workload. The results support the capability of multimodal learning for reliable and efficient DR screening in clinical practice.

The analysis is also conducted by comparing the proposed framework with traditional DL-based systems, further substantiating its performance. The use of patient clinical information improved model prediction accuracy: while image-based models (VGG16, ResNet50, DenseNet121, EfficientNet-B3) demonstrated competitive classification performance, this was further boosted by the integration of patient clinical information. These clinical parameters also contributed complementary data that was not derivable from retinal images only and included: age, diabetes duration, glycated hemoglobin (HbA1c), fasting blood glucose, blood pressure, body mass index, cholesterol level, smoking status, and family history of diabetes. Multimodal feature fusion strategy demonstrated that the model had the ability of learning complex relationship of retinal abnormalities and systemic clinical risk factors to achieve better classification accuracy than single-modality approaches, and showed better robustness.

The effectiveness of multi modal data integration was also proven in the ablation study. The model trained by only clinical data had moderate performance in predicting the outcome and the image-only model had significantly better accuracy but the combination of both modalities provided the best performance in all metrics tested. The finding implies that these 2 sources of information, retinal imaging and clinical information, provide complementary diagnostic information for prediction of DR. The confusion matrix further confirmed a very limited number of false classifications (false positive and false negative indications) for the proposed framework, which indicates the high discriminative power and good generalization ability of the proposed framework. Additionally, training and validation performance curves revealed that the model demonstrated good convergence with a slight difference between training and validation accuracy demonstrating the effectiveness of reducing overfitting using data augmentation, dropout regularization and transfer learning.

In summary, the proposed multi-modal deep learning model offers a promising AI-driven approach to early detection of DR. Retinal image analysis coupled with patient-specific clinical data contributes to more accurate diagnoses and facilitates clinical decision making. Could help ophthalmologists prioritize patients at a higher risk of developing glaucoma, decrease diagnostic burden, and increase screening coverage, especially in resource limited healthcare institutes where specialist is in short supply. Further studies should be conducted to validate the framework on more extensive multi-center data, add longitudinal patient information, test multimodal transformer-based architectures, and integrate aspects of EAI frameworks to enhance model explainability, interpretability and clinical viability.

CONCLUSION

This is the first study to propose a multimodal Deep Learning approach for early DR prediction using the retinal fundus image and applying structured clinical data to extract complementary visual and clinical features which enable the model to achieve better prediction of DR. The proposed framework consists of an artificial neural network trained on clinical features of the patients and a convolutional neural network for retinal image analysis, and the combination effectively predicted the injuries, outperforming conventional, image-based deep learning methods. The experimental assessment showed outstanding performance for classification with a key of accuracy, precision, recall, specificity, F1 score, and AUC reaching 97.20%, 96.90%, 96.50%, 97.80%, 96.70%, and 0.989, respectively, which confirms the effectiveness of multimodal feature fusion in the field of early prediction of DR. The proposed model significantly improved the accuracy of diagnosis, decreased the misclassification, and showed good generalization capability by incorporating retinal imaging with the most important clinical risk factors (age, duration of diabetes, HbA1c, blood pressure, BMI, fasting blood glucose, and cholesterol levels). The study results suggest that the developed strategy has promising potential for implementing an intelligent computer aided diagnostic system, which helps the ophthalmologist to screen patients in the early phase, give required interventions at the right time and, ultimately, to make better clinical decisions; especially in our country, where the number of available healthcare resources is inadequate. The study findings would require to be replicated in a larger multi-center data set, patient data should be collected for longitudinal study, researchers should delve into multimodal transformer architectures and integrate explainable artificial intelligence (XAI) methods to enhance the model's transparency and resistance, and to advance its clinical utility.

REFERENCE

[1] Zedadra, A., Zedadra, O., Salah-Salah, M. Y., & Guerrieri, A. (2025). Graph-aware multimodal deep learning for classification of diabetic retinopathy images. *IEEE Access*, *13*, 74799-74810.

[2] Tseng, V. S., Chen, C. L., Liang, C. M., Tai, M. C., Liu, J. T., Wu, P. Y., ... & Chen, Y. H. (2020). Leveraging multimodal deep learning architecture with retina lesion information to detect diabetic retinopathy. *Translational Vision Science & Technology*, *9*(2), 41-41.

[3] Bhutto, J. A., Rahman, Z., He, Z., Dahri, F. H., & Laghari, A. A. (2026). Cross-attention-based multimodal fusion for early diabetic retinopathy detection. *Scientific Reports*.

[4] Song, J., Tian, C., Qi, Y., Gao, X., Cui, M., Yang, J., ... & Xi, X. (2025). Detection of diabetic retinopathy using multicolor image by multimodal network incorporating information bottleneck (MNIIB). *Scientific Reports*, *16*(1), 1835.

[5] Anupama, B. C., Rao, S. N., Malini, M. B., & Athreya, V. V. (2025). Comparative analysis of novel preprocessing techniques and deep learning based multi modal feature fusion for diabetic retinopathy grading. *Scientific Reports*, *16*(1), 1743.

[6] Wardhani, K. D. K., Kasim, S., Hassan, R., Hidayat, R., & Sujon, K. M. (2025). Deep Learning for Diabetic Retinopathy Detection: A Review of Multimodal Data Fusion Approaches.

[7] Kropp, M., Golubnitschaja, O., Mazurakova, A., Koklesova, L., Sargheini, N., Vo, T. T. K. S., ... & Thumann, G. (2023). Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—risks and mitigation. *Epma Journal*, *14*(1), 21-42.

[8] Ting, D. S. W., Cheung, G. C. M., & Wong, T. Y. (2016). Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. *Clinical & experimental ophthalmology*, *44*(4), 260-277.

[9] Zheng, Y., He, M., & Congdon, N. (2012). The worldwide epidemic of diabetic retinopathy. *Indian journal of ophthalmology*, *60*(5), 428.

[10] Wong, T. Y., & Sabanayagam, C. (2019). The war on diabetic retinopathy: where are we now?. *Asia-Pacific Journal of Ophthalmology*, *8*(6), 448-456.

[11] Vision Loss Expert Group of the Global Burden of Disease Study. (2024). Global estimates on the number of people blind or visually impaired by diabetic retinopathy: a meta-analysis from 2000 to 2020. *Eye*, *38*(11), 2047.

[12] Meng, Y., Liu, Y., Duan, R., Liu, B., Lin, Z., Ma, Y., ... & Li, T. (2025). Global, Regional, and National Epidemiology of Vision Impairment due to Diabetic Retinopathy Among Working-Age Population, 1990–2021. *Journal of Diabetes*, *17*(7), e70121.

[13] Błaszkiewicz, M., Walulik, A., Florek, K., Górecki, I., Sławatyniec, O., & Gomułka, K. (2023). Advances and perspectives in relation to the molecular basis of diabetic retinopathy—a review. *Biomedicines*, *11*(11), 2951.

[14] Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in public health*, *11*, 1273253.

[15] Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., & Ilono, P. (2025). Deep convolutional neural networks in medical image analysis: A review. *Information*, *16*(3), 195.

[16] Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., & Khan, M. K. (2018). Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, *42*(11), 226.

[17] Pinto-Coelho, L. (2023). How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. *Bioengineering*, *10*(12), 1435.

[18] Abdou, M. A. (2022). Literature review: Efficient deep neural networks techniques for medical image analysis. *Neural Computing and Applications*, *34*(8), 5791-5812.

[19] Shobayo, O., & Saatchi, R. (2025). Developments in deep learning artificial neural network techniques for medical image analysis and interpretation. *Diagnostics*, *15*(9), 1072.

[20] Raposo, H. (2024). Intelligent imaging: A systematic review of artificial intelligence techniques in disease detection, segmentation, and classification. *Segmentation, and Classification (May 13, 2024).*