DRDD: AI-Powered Multi-Modal System for Diabetic Retinopathy Disease Detection

Sofia A Sathya, Devi D Soundarya, Shree A Viniksha, E Yashika and K Yuvashree

March 31, 2025

DRDD: AI-Powered Multi-Modal System for Diabetic Retinopathy Disease Detection

A. Sathya Sofia,
Associate Professor
Department of Computer Science and Engineering,
PSNA College of Engineering and Technology
Dindigul, Tamil Nadu, India
sathyasofia@psnacet.edu.in

Soundarya Devi M
Department of Computer Science and Engineering,
PSNA College of Engineering and Technology
Dindigul, Tamil Nadu, India
soundaryad704@gmail.com

Viniksha Shree A
Department of Computer Science and Engineering,
PSNA College of Engineering and Technology
Dindigul, Tamil Nadu, India

Yashika E
Department of Computer Science and Engineering,
PSNA College of Engineering and Technology
Dindigul, Tamil Nadu, India

Yuvashree K
Department of Computer Science and Engineering,
PSNA College of Engineering and Technology
Dindigul, Tamil Nadu, India

Abstract

Diabetic Retinopathy is a disease due to diabetes which destroy the blood vessels supplying the retina's light sensitive tissue and resulting in vision loss and complete blindness. This disease primarily affects people of working age and increases the socioeconomic burden on individuals and the healthcare system. The automated detection for diabetic retinopathy using machine learning techniques with the help of retinal images and structured data of the patient helps to enhance the accuracy rate and efficiency of diagnosis.

This work proposes the Transformer - Based Fusion model that integrates deep learning and structured clinical data analysis to improve diabetic retinopathy detection. This model combines EfficientNet-B3, Vision Transformer, which is used in feature extraction and detect changes in retinal images, and TabNet, used to analyze structured clinical data of a patient to improve the accuracy of the result. Generative AI model is also integrated with the help of GPT-4 to give a personalized medication suggestion to patients according to their health condition and medical history. This fusion method significantly demonstrates the accuracy of 94.7% in disease detection.

Keywords: Diabetic Retinopathy, Transformer-Based Fusion model, EfficientNet-B3, Vision Transformer, TabNet, GPT-4, Retinal Fundus Image, Structured patient data, Gen AI.

I. INTRODUCTION

Diabetic Retinopathy is a significant eye disease condition that develops due to the long-term presence of high glucose levels in the blood, which will affect the retina's blood vessel. If it is left untreated, it will cause complete blindness. This progresses through four phases namely, mild, moderate, severe, and proliferative. It increases with high cholesterol, uncontrolled hypertension and poorly controlled diabetes. Common early warning signs include floaters, dark spots, blurriness, and difficulty perceiving colors. However, early detection [1] with timely treatment can significantly lower the risk of blindness.

Traditionally, ophthalmologists manually examine retinal fundus images for detection purposes, and that requires more time and specialized experts. The advancement in Machine Learning [2] enabled automated systems with good accuracy in detecting the disease. Models in deep learning [3] like ResNet, CNNs, DenseNet, and Inception depend only on retinal images, which may result in false positives. This proposed work overcomes this limitation by integrating the results of retinal images with the clinical data of the patient to improve diagnostic accuracy.

II. LITERATURE SURVEY

Intifa Aman Taifa et al. [4] defined a system which uses DenseNet121 to achieve 95.5% accuracy in classifying multi-class and obtained 98.36% accuracy in binary classification. But the limitations in this study are computational resource limitations and restrictions in sample size. Lijuan Wang et al. [5] described a system that used 103 DR patient images and split them into 7:3 training and validation sets and applied a transfer learning technique to VGG19 and DenseNet for Diabetic Retinopathy detection that achieved an accuracy of 89% and 89.7%. Krishnan Sangeetha et al. [6] proposed a broad study of extracting features to detect and classify the disease of diabetic retinopathy. The techniques used in this, include CNN, SVM, KNN, and GoogleNet (Transfer Learning) with the accuracy of 94.43%. The challenge remains in computational cost and real-world deployments. Mohamed R. Shoaib et al. [7] introduced a novel approach that utilizes pre-trained models such as InceptionResNetV2, InceptionV3, along with a custom model named DiaCNN for diagnosing Diabetic Retinopathy. During training, its accuracy was 100% and during testing it reduced to 98.3%. Ankush Jain et al. [8] used quantum transfer learning to identify the diabetic retinopathy disease. It used pre-trained classical neural networks such as ResNet152, ResNet-101, ResNet50, ResNet-34, ResNet-18 and InceptionV3 for feature extraction and variation quantum classifier is used in classification. The practical implementation of these models is difficult and results in limitations of this system.

Inas AI-Kamachy et al. [9] presented a system that classified the disease diabetic retinopathy. It included MobileNet, VGG-16, InceptionV3, and InceptionResNetV2, and achieved AUC values as 0.70, 0.53, 0.63, and 0.69 respectively. Hossein Shakibania et al. [10] introduced a system that designed for the main purpose of identifying disease and its stage grading using only one fundus image. It achieved 98.5% accuracy in binary classification of disease and 89.6% in stage grading. The complexity of the dual-branch architecture remains as a drawback for deployment in resource-limited environments.

III. MATERIALS AND METHODS

A. Dataset

This work utilizes the **APTOS 2019** dataset created by the Asian Pacific Tele-Ophthalmology Society along with preprocessed dataset to develop an AI-based multi-modal to detect diabetic retinopathy for retinal fundus images. It contains **17,632 high-resolution images**, each labeled based on the **International Clinical Diabetic Retinopathy Scale,** which classifies DR into five severity levels:

1. No DR
2. Mild
3. Moderate
4. Severe
5. Proliferative DR

Classes	Images counts
No DR	7,805
Mild DR	2,770
Moderate DR	4,999
Severe DR	1,063
Proliferative DR	995

Table 1. Classification by severity level

The dataset is gathered from various hospitals and eye clinics, making the dataset diverse. On the other hand, numerical symptoms are acquired through electronic health records (EHRs). The dataset is preprocessed, Lately, the preprocessed data from images and numerical records have been fused, creating a multi-modal AI with enhanced diagnostic accuracy that generates medical suggestions.

B. Multi-modal AI training phases

The following is the workflow of training a multi-modal AI that generates medicine suggestions.

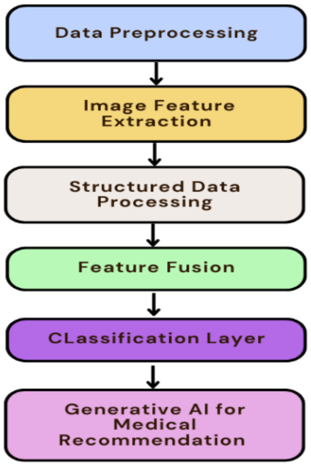


Fig. 1. Phases involved in training multi-modal AI

1. **Data Preprocessing**

a. Retinal Image Data

- **Load Dataset:** The Initial step in processing retinal images is, loading the data to utilize the 17,632 fundus images from the dataset.
- **Image Resizing: W**e would further standardize all images to the resolution of 224 x 224 pixels to fit it into the model architecture.
- **Normalization:** After resizing the retinal image, we need to scale the pixel values to the range [0,1].
- **Augmentation of data:** In enhancing the process to improve model generalization, methods including flipping, rotation, and contrast modifications are applied.

Stages	Number of Images
Training set	14,105
Validation set	1764
Testing set	1763

Table 2. Dataset classification

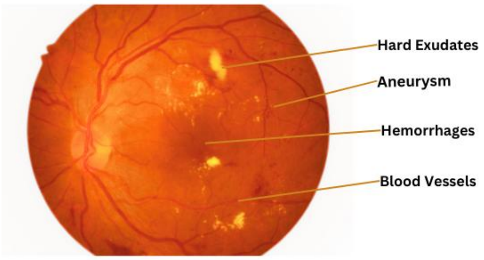
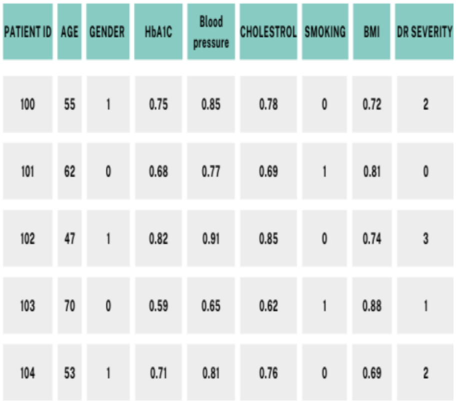


Fig. 2. Affected Retinal Image

b. Structured Clinical data

The **APTOS dataset** does not provide patient records, so we source them from electronic health records (EHRs) [11].



PATIENT ID	AGE	GENDER	HbA1C	Blood pressure	CHOLESTROL	SMOKING	BMI	DR SEVERITY
100	55	1	0.75	0.85	0.78	0	0.72	2
101	62	0	0.68	0.77	0.69	1	0.81	0
102	47	1	0.82	0.91	0.85	0	0.74	3
103	70	0	0.59	0.65	0.62	1	0.88	1
104	53	1	0.71	0.81	0.76	0	0.69	2

Table 3. Clinical data of patient

- **Load & Inspect data:** Load the structured patient data in CSV or database format.
- **Handling Missing Values:** Drop rows, impute values (fill: numerical, categorical, time-series).
- **Feature Scaling:** Neural networks perform better when data is scaled, so Min-Max scaling and Standardization (Z- score) are done.
- **Encode Categorical Variables:** Label encoding (Ex: Gender: M -> 0, F ->1), One-Hot Encoding (for multi-category values).

Outliers: Outliers are handled using z-score filtering or IQR method to remove extreme values

2. Model Architecture

I. Image Feature Extraction

Model Selection: EfficientNet-B3 [12] and Vision Transformer (ViT) [13] are chosen to extract deep visual features from retinal images.

a) <u>EfficientNet-B3:</u> Used to extract local features

- ➢ **Convolutional Layers:** The first step is to pass the images through these layers to identify lower level characteristics that include shapes and edges.
- ➢ **Depth-wise Convolutions:** This operation optimizes feature extraction, making the model capable of highlighting important regions (ex: microaneurysms, hemorrhages).
- ➢ **Global Pooling:** A global average pooling layer aggregates spatial information to create a compact vector representing the local pathological regions.

b) <u>Vision Transformer (ViT):</u> Used to extract global features

- ➢ **Patch Partition:** The image is split into non-overlapping patches.
- ➢ **Linear Embedding:** Each patch is flattened and linearly embedded into a high-dimensional feature vector.
- ➢ **Self-Attention Mechanism:** ViT applies multi-head self-attention to each patch to achieve the most relevant global features.
- ➢ **Feedforward Neural Network:** The Transformer layers process the self-attention output, producing the final set of global retinal features

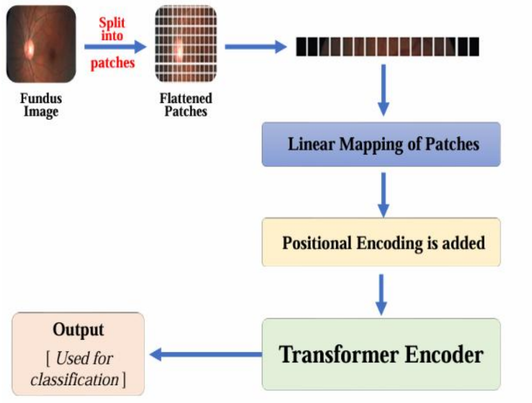

Fundus Image
Split into patches
Flattened Patches
Linear Mapping of Patches
Positional Encoding is added
Transformer Encoder
Output
[Used for classification]

Fig. 3. Vision Transformer for Fundus Image Classification

II. Clinical Feature Extraction

Model Selection: TabNet [14] is chosen for its ability to process numerical and categorical features.

- ➢ **Sparse Attention**: It uses a sparse attention mechanism to learn features at each decision point, which is beneficial in clinical data.
- ➢ **Decision Trees:** TabNet uses these trees to split data and make predictions.
- ➢ **Feature Processing:** The attention mechanism helps focus on informative parts and produces a clinical feature vector.

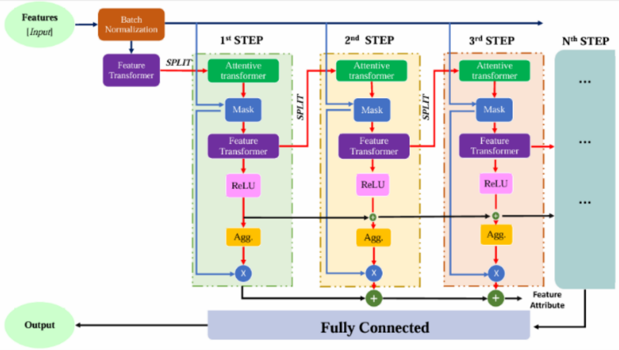


Fig. 4. TabNet Architecture Flow

3. Feature Fusion

Model: Here, we use a Multi-modal Transformer/ cross-attention Transformer to combine both image features and clinical features effectively.

Input Features:

- **Image features:** The local feature vector is taken from EfficientNet-B3, global feature vector is taken from ViT.
- **Clinical features:** These are extracted from TabNet.

Input Representation:

- **Embedding Layer:** In this layer, the inputs are made compatible, and both the image and clinical vectors are mapped.
- **Concatenation:** The embedding is tokenized as a sequence of features.

Cross-Attention Mechanism

- **Cross-Attention:** This mechanism is used to handle complex dependencies as it allows interaction between both inputs.
- **Self-Attention:** It allows focus on relevant regions of both the data.

Transformer Layers:

- The layers of the transformer process the embedded features, learning multi-modal relationships and dependencies.
- The output of the transformer is a combined feature representing the integration of both image-derived and clinical data.

4. Classification Layer

Fully Connected Layer:

- **Purpose:** After fusion, the feature vector is passed through this layer to learn higher-level abstract representations.
- **Design:** Adding more layers can increase the model's capacity to learn.
- **Activation function:** ReLU is used in hidden layers for non-linearity.

Output Layer:

- **Severity Prediction:** For severity prediction of Diabetic retinopathy, the output layer uses a softmax activation function.
- Finally, through the probability distribution of the softmax function, multiple classes of diabetic retinopathy are classified into no DR, mild, moderate, severe.

5. Medical Recommendation

- After the DR classification, GPT-4 [15] which is a pre-trained transformer model used in the medicinal field for generating personalized treatment recommendations on patient data.
- The model retrieves relevant medical knowledge using Retrieval-Augmented Generation (RAG).
- It generates treatment recommendations using ADA guidelines and real-world case studies.

IV. RESULTS AND DISCUSSION

In our proposed work, EfficientNet-B3 and Vision Transformer were integrated and trained with the APTOS dataset by using 80% of dataset images for training the model, 10% of images were utilized in validation and the rest of 10% of images were evaluated in testing the outcomes.

- **Training and Validation:** The following graph shows the accuracy comparison of training and validation. It indicates a progressive improvement in both training and validation over epochs. This represents that this model is trained successfully with the dataset to obtain more accuracy in validation and testing.

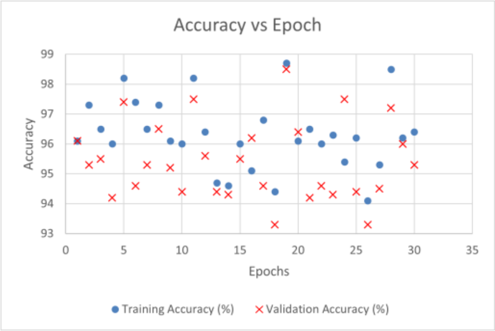


***Fig. 5**. Accuracy Vs Epoch*

- **Comparison of Accuracy:** The accuracy of different techniques used in this model is compared in the below graph. The bar graph indicates a progressive improvement and comparatively high accuracy is obtained in fusion model with Generative AI.

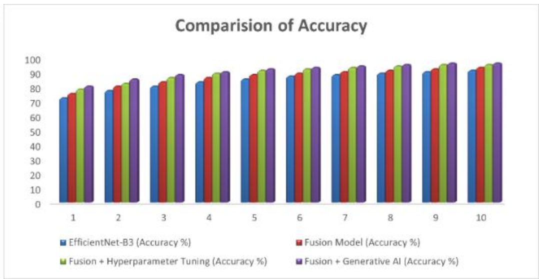


Fig. 6. Comparison graph of Accuracy

- **Confusion matrix:** The following confusion matrix provides a detail of classification performance by displaying true label and predicted label.

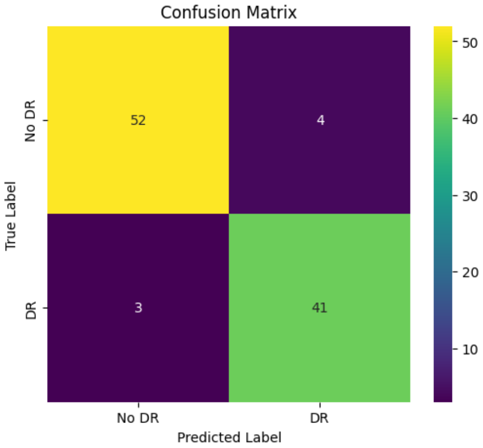


Fig. 7. Confusion Matrix

- **ROC Curve:** The following ROC curve assesses the model's capacity to distinguish between the rate of true positive and false positive. A higher performance of the model is shown by greater area (AUC) that classifies diabetic and non-diabetic cases.

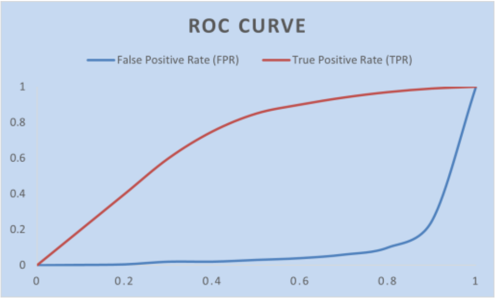


Fig. 8. ROC curve

- **Feature Importance Score:** The importance score plays a major role in improving diagnostic accuracy. It highlights the contribution of different features such as retinal image score and specific clinical indicators. The features with higher importance scores have a stronger influence and removing lower importance features might optimize the model's performance.

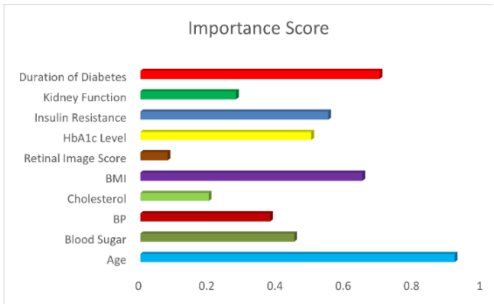


Fig. 9. Feature Importance Score

- **Prediction Vs Actual:** The following graph represent that how the prediction of diabetic retinopathy matches with actual value. It shows high matching results to detect diabetic retinopathy disease.

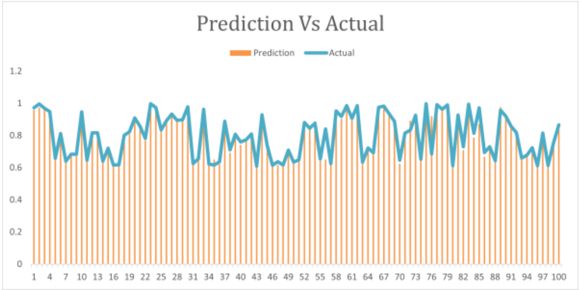


Fig. 10. Prediction Vs Actual

Comparison Table:

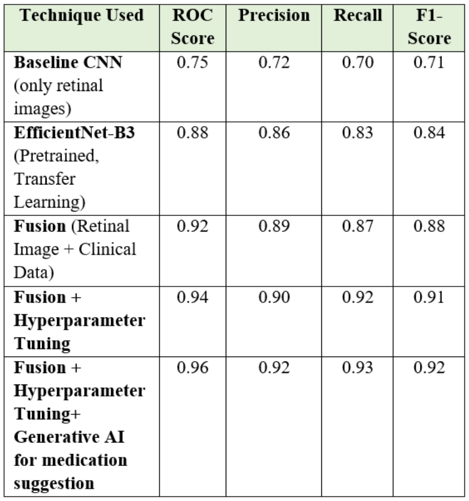
The following table compares the accuracy of different techniques used in this model.

Technique Used	Training Accuracy	Validation Accuracy	Testing Accuracy
Baseline CNN (only retinal images)	80%	75%	72%
EfficientNet-B3 (Pretrained, Transfer Learning)	91%	88%	87%
Fusion (Retinal Image + Clinical Data)	93%	91%	90%
Fusion + Hyperparameter Tuning	95%	93%	93%
Fusion + Hyperparameter Tuning+ Generative AI for medication suggestion	96%	94%	94%

Table 4. Accuracy Table

The results obtained from the model demonstrate the prediction reliability with more accurate classification. After detecting the severity of diabetic retinopathy, the system suggests personalized suggestion with treatment plans based on patient's history and clinical parameters. Despite the fact that the model performs well, the challenge may occur if the image quality is not good and imbalance in data. The improved quality of retinal images and accurate clinical data of patient should be given to the model to avoid this problem and perform well.

Key measures such as recall, precision, ROC score, and F1-score were used to assess the model's performance using various approaches. The results are summarized in the table below.



Technique Used	ROC Score	Precision	Recall	F1-Score
Baseline CNN (only retinal images)	0.75	0.72	0.70	0.71
EfficientNet-B3 (Pretrained, Transfer Learning)	0.88	0.86	0.83	0.84
Fusion (Retinal Image + Clinical Data)	0.92	0.89	0.87	0.88
Fusion + Hyperparameter Tuning	0.94	0.90	0.92	0.91
Fusion + Hyperparameter Tuning+ Generative AI for medication suggestion	0.96	0.92	0.93	0.92

Table 5. Comparison table of techniques

V. CONCLUSION

This proposed work presents an AI-powered Diabetic Retinopathy Detection system that leverage the deep learning with Generative AI to improve the diagnostic accuracy and suggest treatment planning. It classifies diabetic retinopathy cases with high precision by using retinal images along with patient clinical data. The innovative contribution of this work is AI powered medication recommendation system that provides personalized treatment plan with lifestyle modification to lower the risk of disease based on AI-generated insights. The future work will focus on expanding the dataset with diverse retinal images to improve the generalization and deploy the system in real-time healthcare applications for automated diabetic retinopathy screening and treatment guidance. This integration of automated diagnostic and Generative AI for medical recommendations represents a significant step toward intelligent and data-driven healthcare solutions.

REFERENCES

[1] Vijay, K., P. Krithiga, S. Kavirakesh, S. Swetha, and B. Vishal. The Early Detection in the Disease Diabetic Retinopathy with the help of Deep Convolutional Neural Network. In International Conference on Advances in Artificial Intelligence and Machine Learning in Big Data Processing, *pp. 315-327. Cham: Springer Nature Switzerland*, (2023).

[2] Sofia, A. S., Sowmiya, K., Soundarya, K., & Theepiga, M. APD-ML: Air Pollution Detection Using Machine Learning Algorithms. In *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)* (pp. 1-5). IEEE. (2023).

[3] Rajendran, Sowmia Kanakam, Dennise Mathew, Babu Rajendiran, and Vijay Kandasamy. Diabetic retinopathy detection using deep learning techniques. In AIP Conference Proceedings, vol. 2790, no. 1. AIP Publishing, (2023).

[4] Taifa, I. A., Setu, D. M., Islam, T., Dey, S. K., & Rahman, T. A hybrid approach with customized machine learning classifiers and multiple feature extractors for enhancing diabetic retinopathy detection. *Healthcare Analytics*, *5*, 100346. (2024).

[5] Wang, L., Li, B., Pan, J., Zhang, C., & Wang, T. Optical coherence tomography image recognition of diabetic retinopathy based on deep transfer learning. *Journal of Radiation Research and Applied Sciences*, *17*(3), 101026. (2024).

[6] Sangeetha, K., Valarmathi, K., Kalaichelvi, T., & Subburaj, S. A broad study of machine learning and deep learning techniques for diabetic retinopathy based on feature extraction, detection and classification. *Measurement: Sensors*, *30*, 100951. (2023).

[7] Shoaib, M. R., Emara, H. M., Zhao, J., El-Shafai, W., Soliman, N. F., Mubarak, A. S., ... & Esmaiel, H. Deep learning innovations in diagnosing diabetic retinopathy disease: The potential of transfer learning and the DiaCNN model. *Computers in Biology and Medicine*, *169*, 107834. (2024).

[8] Jain, A., Gupta, R., & Singhal, J. Diabetic Retinopathy Detection Using Quantum Transfer Learning. *arXiv preprint arXiv:2405.01734*. (2024).

[9] Al-Kamachy, I., Hassanpour, R., & Choupani, R. Classification of diabetic retinopathy disease using the pre-trained deep learning models. *arXiv preprint arXiv:2403.19905*. (2024).

[10] Shakibania, H., Raoufi, S., Pourafkham, B., Khotanlou, H., & Mansoorizadeh, M. Dual branch deep learning network for detection and stage grading of diabetic retinopathy. *Biomedical Signal Processing and Control*, *93*, 106168. (2024).

[11] Breeyear, J. H., Mitchell, S. L., Nealon, C. L., Hellwege, J. N., Charest, B., Khakharia, A., ... & Giri, A. Development of electronic health record based algorithms to identify individuals with diabetic retinopathy. *Journal of the American Medical Informatics Association*, *31*(11), 2560-2570. (2024).

[12] Kansal, I., Khullar, V., Sharma, P., Singh, S., Hamid, J. A., & Santhosh, A. J. Multiple model visual feature embedding andselection method for an efficient ocular disease classification. *Scientific Reports*, *15*(1), 5157. (2025).

[13] Cutur, E. S., & Inan, N. G. Multi-class Classification of Retinal Eye Diseases from Ophthalmoscopy Images Using Transfer Learning-Based Vision Transformers. *Journal of Imaging Informatics in Medicine*, 1-15. (2025).

[14] Khan, Q. W., Iqbal, K., Ahmad, R., Rizwan, A., Khan, A. N., & Kim, D. An intelligent diabetes classification and perception framework based on ensemble and deep learning method. *PeerJ Computer Science*, *10*, e1914. (2024).

[15] Gopalakrishnan, N., Joshi, A., Chhablani, J., Yadav, N. K., Reddy, N. G., Rani, P. K., ... & Venkatesh, R. Recommendations for initial diabetic retinopathy screening of diabetic patients using large language model-based artificial intelligence in real-life case scenarios. *International journal of retina and vitreous*, *10*(1), 11. (2024).