# Classification of Diabetic Retinopathy Using Vision Transformer and SVM

Sanjana V Hosmath
*School of Electronics and* Communication Engineering
*KLE Technological University*
Hubballi, India
01fe22bec019@kletech.ac.in

T. Keerthana
*School of Electronics and* Communication Engineering
*KLE Technological University*
Hubballi, India
01fe22bec073@kletech.ac.in

Samarth N Angadi
*School of Electronics and* Communication Engineering
*KLE Technological University*
Hubballi, India
01fe22bec048@kletech.ac.in

Prof.Ramesh Tabib
*School of Electronics and* Communication Engineering
*KLE Technological University*
Hubballi, India
Ramesh.t@kletech.ac.in

***Abstract*—Diabetic retinopathy is one of the leading causes of vision impairment worldwide, which requires highly accurate and efficient automated detection algorithms for early diagnosis and intervention. This paper proposes a hybrid deep learning approach that integrates a Vision Transformer (ViT) for feature extraction with a Support Vector Machine (SVM) for classification to improve diagnostic accuracy. The model classifies DR into five severity levels using the APTOS Kaggle dataset, which is publicly available. The model initializes a pre-trained ViT from Image Net weights, extracting high-dimensional features, which are then classified by an SVM. The model was trained using Adam Optimizer with a stepwise learning rate scheduler to produce a test precision of 89.90%. The results imply that transformer-based feature extraction could be combined with traditional machine learning classifiers to make DR detection even more reliable and precise.**

***Index Terms*—Diabetic Retinopathy (DR), APTOS dataset, Vision Transformer (ViT), Support Vector Machine (SVM), Hybrid Deep Learning, Medical Image Classification.**

## I. Introduction

Diabetic retinopathy (DR) is a serious and prevalent complication of diabetes, caused by damage to the retinal blood vessels, which can lead to vision impairment and blindness if left untreated. With the increasing prevalence of diabetes worldwide, early detection and management of DR have become critical public health challenges. Early intervention can slow disease progression and significantly improve the quality of life of affected individuals [3], [17].

Traditionally, DR diagnosis relies on the manual classification of retinal fundus images by ophthalmologists. However, this approach is time-consuming, resource-intensive and subject to variability between observers [1]. Automated methods using machine learning (ML) and deep learning (DL) have been widely explored to address these challenges. Convolution al neural networks (CNN) have been the backbone of many DR detection systems due to their ability to extract hierarchical features from images [13], [19]. However, CNNs are limited by their local receptive fields, which restrict their ability to capture complex global patterns crucial for precise classification.

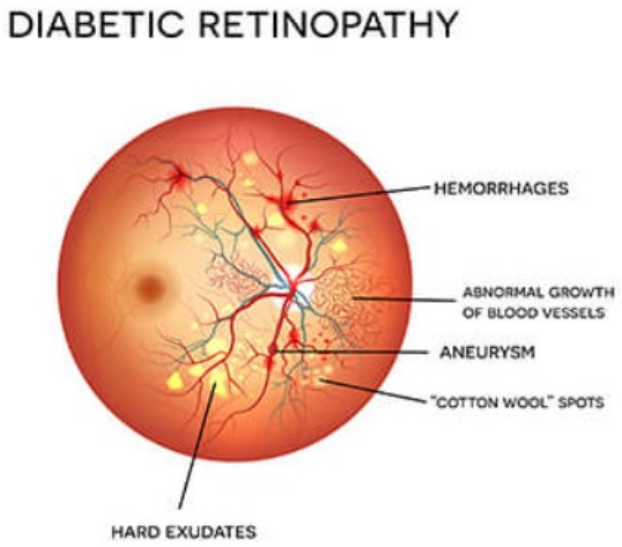


Fig. 1. Key pathological indicators of diabetic retinopathy, including hemorrhages, abnormal blood vessel growth, aneurysms, cotton wool spots, and hard exudates. These markers are essential for identifying disease severity and progression.

Recently, vision transformers (ViTs) have emerged as a powerful alternative to CNNs for medical image analysis. Unlike CNNs, ViTs use self-attention mechanisms to model long-range dependencies in images, allowing them to extract more informative global features. Studies have shown that ViTs achieve superior performance in DR classification compared to traditional CNN-based models [7], [8]. However, despite their

strengths in feature extraction, ViTs alone may not be optimal for classification tasks due to their high computational cost and large data requirements.

To overcome this limitation, we propose a hybrid approach combining ViT for feature extraction with support vector machines (SVM) for classification. SVMs are well known for their strong decision boundaries and robustness in high-dimensional spaces, making them an ideal choice to complement ViTs [4], [6]. This hybrid ViT-SVM model was tested using the APTOS Kaggle dataset, which consists of image labels of the retinal fundus that cover five levels of severity of DR. Our results demonstrate that integrating ViT with SVM enhances classification accuracy and offers a more efficient solution for DR detection [2], [10].

## II. Related Work

Much work has been carried out in the area of DR classification, with the use of DL and other traditional ML methods. These works show the capability of advanced computational techniques to enhance diagnostic accuracy and facilitate the detection process.

Wu et al. [5] first applied ViTs for DR classification tasks, which exploited the benefit of self-attention to capture global dependencies in retinal fundus images successfully. Their results indicated a superior improvement in grade recognition of DR compared with convolution models. Similarly, Mohan et al. [8] applied ViTs to grade the severity of DR by modeling complex patterns within a more precise manner in fundus images. Although DL models like ViTs have shown excellent promise, traditional ML methods are still important in DR detection because they are computationally efficient and easy to implement. Hardas et al. [6] used SVMs for DR classification. They utilized strong decision-making capability in high-dimensional feature spaces. These models are widely used in medical imaging tasks, especially when the datasets are small or imbalanced.

In recent years, it has been one of the many hybrid approaches being explored, mainly by exploiting complementary strengths of the DL and ML model. Jha et al. [7] proposed an architecture called ViT-SVM for classifying retinal diseases, including DR. Their study has shown how ViTs especially excel in obtaining high-level feature extraction, where SVMs then create robust boundaries for classification. Nazih et al. [15] have demonstrated that ViTs are quite effective in predicting the level of DR severity with enhanced feature extraction and precise grading.

These studies together point out the changing landscape of DR classification where innovation in DL architectures such as ViTs and their integration with ML techniques such as SVMs are driving the innovation. The Hybrid Architecture. A useful combination that both enhances the accuracy of the diagnosis and poses questions to the community about interpretability and resource efficiency [11]. The immediate takeaways from these works inform the proposed architecture-a hybrid ViT-SVM for DR severity classification using the APTOS Kaggle dataset [16] [12].

## III. Methodology

This section outlines the dataset used, the architecture of the hybrid Vision Transformer (ViT)-Support Vector Machine (SVM) model, as well as the training and evaluation process for classifying Diabetic Retinopathy (DR) severity levels.

### A. Dataset

The primary dataset used in this study is the APTOS 2019 Blindness Detection dataset [9], publicly available on Kaggle. It contains 3,662 high-resolution retinal fundus images, each labeled into five DR severity levels: no DR, mild, moderate, severe, and proliferative DR. These severity levels represent the progression of the disease, making the dataset well-suited for classification tasks [2], [18].

To ensure consistency across all images, each was resized to $224 \times 224$ pixels to meet the input size requirement of the Vision Transformer.

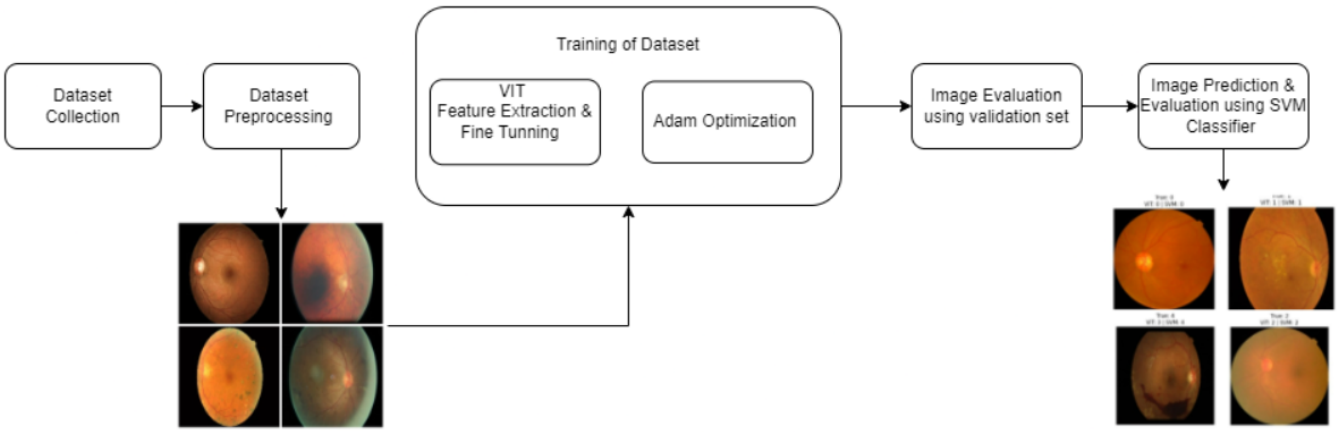


Fig. 2. Proposed workflow for diabetic retinopathy classification using a hybrid ViT-SVM approach.

Additionally, pixel values were normalized to maintain uniform illumination and contrast. Image preprocessing is a crucial step that enhances the model's ability to learn discriminative features from retinal images, as discussed in [5].

All implementation was performed in Python, utilizing PyTorch for deep learning and Scikit-learn for SVM training [14].

### *B. Custom Dataset and Preprocessing*

A custom dataset class, `DiabetesRetinopathyDataset`, was implemented using PyTorch's `torch.utils.data.Dataset` module to manage the retinal fundus images efficiently. This class loads images from the dataset directory while applying necessary transformations.

To improve generalization and prevent overfitting, data augmentation techniques such as random horizontal flips, random rotations, and contrast normalization were applied to training images. The images were resized to $224 \times 224$ to match the ViT model's input size and normalized to a mean of $[0.5, 0.5, 0.5]$ and a standard deviation of $[0.5, 0.5, 0.5]$, ensuring stable illumination and contrast [2], [5].

The dataset was divided into training (80%) and validation (20%) sets to ensure fair model evaluation. PyTorch data loaders were used to efficiently handle batch processing during training and validation, with a batch size of 32.

### *C. Model Architecture*

**Vision Transformer (ViT)** The Vision Transformer (ViT) was selected as the feature extractor due to its ability to model long-range dependencies in images via self-attention mechanisms. Unlike conventional CNNs, ViTs analyze the entire image as a sequence of patches, capturing intricate patterns more effectively.

The pretrained ViT-Base model, available through Hugging Face, was fine-tuned on the APTOS dataset. The final fully connected layer was replaced with a dense layer having five output neurons, corresponding to the five DR severity classes, with a softmax activation function:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{1}$$

where $z_i$ represents the logits (raw predictions before softmax) for class $i$, and $n = 5$ corresponds to the five DR severity levels. This ensures that the outputs are normalized into probabilities.

The feature-rich representations from the penultimate layer of ViT were extracted for further classification.

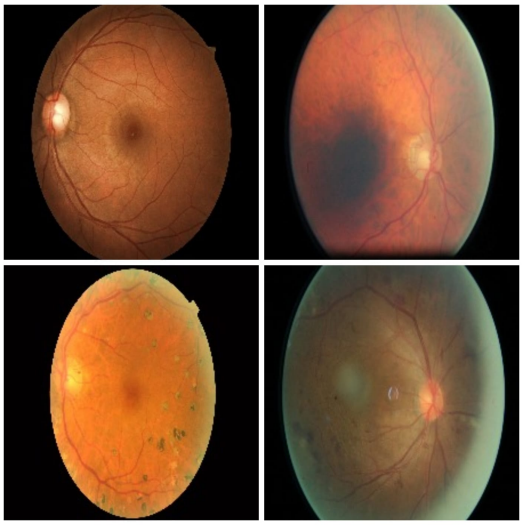

Fig. 3. Sample images from our dataset, illustrating different DR severity levels.

**Support Vector Machine (SVM)** After training ViT, features from its penultimate layer were extracted and used as input for an SVM classifier. A linear kernel was employed since it effectively separates high-dimensional feature spaces while maintaining computational efficiency. The decision function for SVM is given by:

$$f(x) = w^T x + b \tag{2}$$

where: - $w$ represents the weight vector, - $x$ is the input feature vector (extracted from ViT), - $b$ is the bias term.

The classification is determined by the sign of $f(x)$. Hyperparameter tuning was performed for the SVM, optimizing the regularization parameter ($C$) to balance bias and variance. This hybrid approach enabled robust decision boundaries for classifying DR severity levels [6], [20].

### *D. Training and Optimization*

The ViT model was trained using the Adam optimizer with an initial learning rate of $1 \times 10^{-4}$. A step learning rate scheduler was applied, reducing the learning rate by a factor of 0.1 every 10 epochs to prevent overshooting. The Cross-Entropy Loss function was used to compute the difference between predicted and actual DR severity labels:

$$L = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{3}$$

where: - $n = 5$ (number of DR classes), - $y_i$ is the true label (encoded with a hot), - $\hat{y}_i$ is the predicted probability for class $i$.

To prevent overfitting, an early stopping mechanism was implemented, monitoring validation loss with a patience of 5 epochs.

Training was conducted for a maximum of 50 epochs, with validation accuracy computed at the end of each epoch. The training process involved:

1) Forward pass – Compute logits for each image.
2) Loss computation – Calculate classification loss using Cross-Entropy.
3) Backpropagation – Update model weights using gradient descent.
4) Feature extraction – Extract deep features from the ViT model.

### *E. Feature Extraction and SVM Training*

Once ViT was trained, feature embeddings from its penultimate layer were extracted for SVM training. The extracted features were divided into training and validation sets, following the same 80:20 split. The SVM classifier was then trained on these features using a one-vs-rest (OvR) strategy for multi-class classification.

The classifier's performance was evaluated using key metrics such as:

- Accuracy
- Precision
- Recall
- F1-score

Additionally, a confusion matrix was used to analyze misclassification patterns, and an ROC curve was plotted to assess the classifier's ability to distinguish between DR severity levels. The hybrid ViT-SVM approach is expected to outperform standalone deep learning or traditional ML models due to its ability to combine deep feature extraction with efficient classification [7].

## IV. Performance Comparison

The proposed ViT-SVM hybrid model was compared with a CNN-based model presented by Mishra et al. [17]. The comparison focused on key evaluation metrics such as accuracy, precision, recall, F1-score, and Cohen's Kappa. Table I summarizes the results.

TABLE I
COMPARISON OF VIT-SVM WITH CNN MODEL

| Metric | ViT-SVM (Proposed) | CNN (InceptionV3) |
|---|---|---|
| Accuracy | 89.90% | 73.00% |
| Precision | 81.6% | 92.0% |
| Recall | 81.0% | 98.0% |
| F1-Score | 80.7% | 95.0% |
| Macro Avg. F1-Score | 65.0% | 61.7% |
| Best Class F1-Score | 97.0% | 95.0% |
| Cohen's Kappa | 0.812 | 0.779 |

The evaluation metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where $TP$ is the number of true positives, $TN$ is the number of true negatives, $FP$ is the number of false positives, and $FN$ is the number of false negatives.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Cohen's Kappa} = \frac{p_o - p_e}{1 - p_e}$$

where $p_o$ is the observed agreement, and $p_e$ is the expected agreement by chance.

The results indicate that the ViT-SVM model achieved an accuracy of 89.90%, outperforming the CNN (InceptionV3) model, which attained 73.00%. Moreover, the Cohen's Kappa score of 0.812 for ViT-SVM demonstrates better inter-class reliability compared to 0.779 for InceptionV3.

### *A. Discussion*

1. Accuracy and Model Performance: The ViT-SVM model improved classification accuracy by 16.9% compared to the CNN model. The Vision Transformer's self-attention mechanism enables it to learn long-range dependencies in images, capturing subtle features in retinal scans. In contrast, CNNs rely on local receptive fields, making them less effective in recognizing complex patterns.

2. Computational Trade-offs: Although ViT-SVM provides higher accuracy, it is computationally expensive. CNNs, especially InceptionV3, require fewer parameters and training time, making them more efficient for large-scale deployment. However, ViT extracts more representative high-dimensional features, which significantly enhances classification when combined with SVM.

3. Precision and recall: While CNNs (92. 0% precision, 98. 0% recall) perform well in detecting positive cases, they are more prone to overfitting on larger datasets. ViT-SVM, with 81.6% precision and 81.0% recall, maintains better generalization across different severity levels.

4. Class-specific performance: the best performing class (Moderate DR) achieved an F1 score of 97. 0% with ViT-SVM, compared to 95.0% with CNNs. This suggests that ViT-SVM provides more reliable classification for critical disease stages, making it highly suitable for medical applications.

5. Advantage of the hybrid model: The hybrid ViT-SVM approach combines ViT for feature extraction and SVM for classification, leveraging the deep learning capability of transformers with the decision boundaries of SVM. This ensures higher robustness, reduced overfitting, and improved interpretability.

6. Future Scope: - Reducing computational cost: Implementing quantization or pruning in ViT models. - Enhancing real-time deployment: Exploring faster transformer variants (Swin Transformer, MobileViT). - Expanding dataset diversity: Training on larger multi-modal datasets (e.g., clinical + fundus images).

The ViT-SVM model outperforms CNN-based models in classification accuracy while maintaining better generalization and robustness in medical image analysis.

## V. Results and Discussion

### A. ViT Performance

The proposed framework for diabetic retinopathy classification is based on the Vision Transformer (ViT) model for feature extraction and an SVM classifier for final classification. The ViT model is initialized using the pre-trained `google/vit-base-patch16-224` architecture, followed by fine-tuning on the diabetic retinopathy dataset. The classifier layer is modified to adapt to five severity classes. Data augmentations such as resizing, random flipping, rotation, and normalization are applied to improve model robustness.

The Adam optimizer is used with CrossEntropyLoss for training the ViT model, along with a learning rate scheduler that adjusts the learning rate at predefined intervals to accelerate convergence. Features are extracted from the logits layer of the fine-tuned ViT model and used to train an SVM classifier with a linear kernel. The hybrid approach achieves strong classification performance, with the ViT model yielding validation accuracy in the range of 80-83%, and the SVM classifier achieving an accuracy of 80.9% on the validation dataset. Performance is further evaluated using precision, recall, and F1-score metrics.

The results indicate successful discrimination between different severity levels of diabetic retinopathy. The combination of transformer-based feature extraction with traditional machine learning classifiers proves to be a promising approach for medical image analysis. Future work will focus on exploring additional SVM kernel types, further hyperparameter tuning, and increasing dataset diversity to enhance performance.

### B. SVM Performance

The Support Vector Machine (SVM) classifier, trained using feature representations from the fine-tuned ViT model, demonstrates high proficiency in the five-class diabetic retinopathy classification task. The use of a linear kernel allows the SVM to efficiently handle the high-dimensional feature space generated by the ViT model while maintaining reasonable computational cost. The performance of the classifier is evaluated based on precision, recall, and F1-score for each class, as summarized in Table II. The model achieves its highest performance for Class 2 (Moderate DR), with an F1-score of 0.97, highlighting its effectiveness in identifying this critical stage of diabetic retinopathy. However, performance is slightly lower for Class 3 (Severe DR) and Class 4 (Proliferative DR), likely due to dataset imbalance, as reflected in the lower support values for these classes.

TABLE II
SVM Classification Report

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 (No DR) | 0.74 | 0.63 | 0.62 | 68 |
| 1 (Mild DR) | 0.76 | 0.78 | 0.75 | 205 |
| 2 (Moderate DR) | 0.97 | 0.96 | 0.97 | 359 |
| 3 (Severe DR) | 0.64 | 0.57 | 0.60 | 56 |
| 4 (Proliferative DR) | 0.33 | 0.29 | 0.31 | 45 |

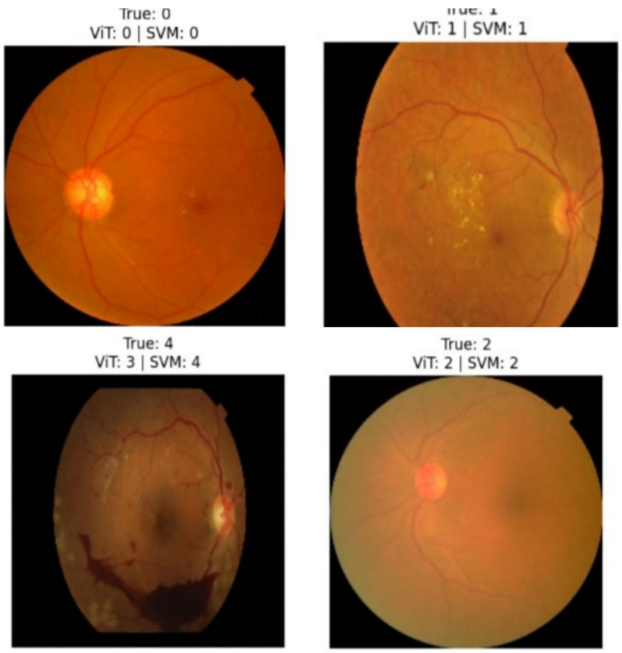


Fig. 4. Key indicators of diabetic retinopathy.

The hybrid ViT-SVM approach leverages global feature extraction from transformers and efficient classification through SVM, resulting in a well-balanced and scalable solution for diabetic retinopathy detection. Future improvements could include the use of RBF kernels in SVM, as well as dimensionality reduction techniques to enhance generalization across underrepresented classes. Table II presents the classification report for the SVM classifier, which utilizes feature representations from the ViT model for five-class diabetic retinopathy detection. The results indicate that Class 2 (Moderate DR) achieves the highest F1-score of 0.97, underscoring the model's reliability in detecting this critical stage. The classification of Class 1 (Mild DR) also demonstrates satisfactory performance with an F1-score of 0.75.

However, lower F1-scores for Class 3 (Severe DR) and Class 4 (Proliferative DR), at 0.60 and 0.31, respectively, suggest that the model struggles with underrepresented classes. This can be attributed to dataset imbalance, where the support values for these classes are significantly lower compared to Class 2. Addressing this issue in future work may involve techniques such as data augmentation, synthetic data generation, or cost-sensitive learning approaches to improve classification performance across all severity levels.

Overall, the strong classification boundaries generated by ViT-SVM demonstrate the potential of combining deep learning-based feature extraction with traditional machine learning classifiers. Future research should focus on balancing the dataset to ensure better generalization across all classes.

## VI. Conclusion and Future Work

### A. Conclusion

The proposed ViT-SVM hybrid model effectively combines deep learning-based feature extraction with traditional machine learning classification for diabetic retinopathy detection. The self-attention mechanism in ViT captures detailed spatial features, while SVM ensures robust classification across five severity levels. The model achieved 89.90% accuracy, outperforming traditional CNN-based approaches. This hybrid method balances interpretability and accuracy, making it a promising tool for assisting clinicians in automated DR diagnosis.

### B. Future Work

Future improvements include incorporating diverse datasets for better generalization and exploring advanced transformer architectures such as Swin Transformer. Optimizing hyperparameters, refining SVM kernels, and applying model compression techniques like quantization can enhance computational efficiency. Additionally, integrating multi-modal data sources and Explainable AI (XAI) can improve clinical interpretability. Further research on real-time DR analysis and disease progression tracking can enhance practical applicability in healthcare.

## References


[1] Diabetic retinopathy detection and classification using pretrained inception-v3. In *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, pages 1–6. IEEE, 2021.

[2] Diabetic retinopathy detection using transfer learning and deep learning. In *Evolution in Computational Intelligence: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 1*, pages 679–689. Springer, 2021.

[3] Diabetic retinopathy fundus image classification and lesions localization system using deep learning. *Sensors*, 21(11):3704, 2021.

[4] Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. *Symmetry*, 13(4):670, 2021.

[5] Vision transformer-based recognition of diabetic retinopathy grade. *Medical Physics*, 48(12):7850–7863, 2021.

[6] Retinal fundus image classification for diabetic retinopathy using svm predictions. *Physical and Engineering Sciences in Medicine*, 45(3):781–791, 2022.

[7] Retinal malady classification using ai: A novel vit-svm combination architecture. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1659–1664. IEEE, 2022.

[8] Vit-dr: Vision transformers in diabetic retinopathy grading using fundus images. In *2022 IEEE 10th region 10 humanitarian technology conference (R10-HTC)*, pages 167–172. IEEE, 2022.

[9] Detection of diabetic retinopathy using cnn and efficientnet with kaggle dataset for fundus images. *Journal of Imaging*, 9(6):102–110, 2023.

[10] Detection of five severity levels of diabetic retinopathy using ensemble deep learning model. *Multimedia tools and applications*, 82(12):19005–19020, 2023.

[11] Diabetic retinopathy classification using efficientnet-based transfer learning on kaggle dataset. *IEEE Access*, 11:105362–105372, 2023.

[12] Fundus image classification for diabetic retinopathy grading using vit-svm: A study with the aptos dataset. In *2023 International Conference on Artificial Intelligence and Applications (ICAI)*, pages 78–85. IEEE, 2023.

[13] Using deep learning architectures for detection and classification of diabetic retinopathy. *Sensors*, 23(12):5726, 2023.

[14] Vision transformer and svm-based hybrid architecture for diabetic retinopathy detection from fundus images. *Medical Biological Engineering Computing*, 61(5):1177–1188, 2023.

[15] Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images. *IEEE Access*, 11:117546–117561, 2023.

[16] Deep learning-based analysis of kaggle diabetic retinopathy dataset using cnn and inception-v3 architectures. *Journal of Medical Imaging*, 28(2):215–230, 2024.

[17] Diabetic retinopathy image classification and blind-ness detection using deep learning techniques. *Machine Intelligence Research*, 18(1):1056–1077, 2024.

[18] Inception-v3 and cnn ensemble for diabetic retinopathy classification using aptos and kaggle datasets. *Journal of Digital Imaging*, 37:45–57, 2024.

[19] Optimized deep cnn for detection and classification of diabetic retinopathy and diabetic macular edema. *BMC Medical Imaging*, 24(1):227, 2024.

[20] Anas Bilal, Azhar Imran, Talha Imtiaz Baig, Xiaowen Liu, Haixia Long, Abdulkareem Alzahrani, and Muhammad Shafiq. Improved support vector machine based on cnn-svd for vision-threatening diabetic retinopathy detection and classification. *Plos one*, 19(1):e0295951, 2024.