Hybrid Deep Ensemble Architecture for Robust Diabetic Retinopathy Classification: Leveraging Transfer Learning and CNN-Transformer Synergy

Youssef Maaod
Galala University
Tamer Abuhamd
Sungkyunkwan University
Shaker El-Sappagh
shaker.elsappagh@gu.edu.eg

Galala University

Article

Keywords:

Posted Date: February 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8614100/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on June 9th, 2026. See the published version at https://doi.org/10.1038/s41598-026-55085-9.

Hybrid Deep Ensemble Architecture for Robust Diabetic Retinopathy Classification: Leveraging Transfer Learning and CNN-Transformer Synergy

Youssef Maaod
yhf400294@Gu.edu.eg
Faculty of Computer Science and Engineering, Galala University, Egypt.

Tamer Abuhamd
tamer@skku.edu
College of Computing and Informatics, Sungkyunkwan University, South Korea.

Shaker El-Sappagh*
Shaker.elsappagh@Gu.edu.eg
College of Computing and Informatics, Sungkyunkwan University, South Korea.

Faculty of Computer Science and Engineering, Galala University, Suez 435611, Egypt

Information Systems Department, Faculty of Computers and Artificial Intelligence, Benha University, 13518, Banha, Egypt

*Correspondence: Shaker El-Sappagh

Abstract
Diabetic Retinopathy (DR) is still one of the main reasons for vision loss worldwide, especially in places where people do not have easy access to regular eye checkups. Early and accurate disease detection is important to avoid permanent damage, but traditional methods are slow and sometimes inconsistent. This study proposes a deep learning framework that combines convolutional neural networks (CNNs), vision transformers, transfer learning, and ensemble techniques to improve DR detection. We used the APTOS 2019 dataset and tested the capabilities of 23 different pre-trained models. Then, we fine-tuned the top models and designed hybrid architectures by combining the best-performing CNNs and transformers in parallel and sequential ways to capture both image spatial features in short and long contexts. The best performance came from combining the top sequential hybrid models using the soft voting architecture, where we got an accuracy of 93.10%, ROC AUC of 99.22%, and F1-score of 93.07%. The optimized model showed that mixing different models and using ensemble methods can lead to better and more stable DR detection decisions. Our approach is a step toward building a reliable and automated system that could help doctors in real-world settings.

Introduction
Diabetic Retinopathy (DR) is the most common cause of vision impairment and blindness among diabetic patients worldwide[1]. Globally, about one-third of people with diabetes have DR, and as of 2020, there were an estimated 103 million adults with the disease[2]. Nearly 28 million people are at risk of severe vision loss due to proliferative diabetic retinopathy (PDR) or diabetic macular edema (DME)[3]. The number of DR cases is expected to reach 160 million by 2045, driven by the expanding frequency of diabetes[4]. DR causes 4.8% of all blindness occurrences worldwide and can cause permanent blindness if treatment is delayed[5]. Approximately 90% of severe vision loss from DR can be avoided through effective therapy. Therefore, the early detection and treatment of DR are essential[6]. Traditional diagnosis relies on manual fundus imaging exams, which are time-consuming, prone to human error, and unachievable for widespread screening, especially in regions with limited resources[1]. AI can help in improving the detection of DR.

The automated diagnosis of DR has been transformed by deep learning (DL), which has solved significant issues with manual screening, including variation between ophthalmologists, high prices, and time-consuming evaluations[7]. In DR detection and grading, DL-based systems, particularly convolutional neural networks, or CNNs, provide a scalable, effective, and highly accurate alternative, achieving performance comparable to or surpassing human experts in DR detection and grading[8] These models achieve remarkable accuracy, with some studies reporting >99% sensitivity in detecting referable DR[9], while ensemble and transfer learning approaches further enhance reliability[10]. Unlike traditional machine learning, DL automatically recovers discriminative features (such as hemorrhages and microaneurysms) without the need for manual preprocessing[11], and architectures such as DenseNet-121 and Inception-

v3 have demonstrated remarkable efficacy in DR classification[12]. APTOS (high-resolution graded images)[13], Messidor (clinician-annotated fundus photos)[14], EyePACS (large-scale transfer learning)[15], DIARETDB0/1 (lesion benchmarks)[16] are some of the key datasets that support these developments.

Transfer Learning (TL) has become a prominent paradigm in medical image analysis due to its ability to achieve excellent performance with a few annotated datasets. In the medical domain, TL uses knowledge from pre-trained models and applies it to specific tasks like diabetic retinopathy grading, where collecting large, labeled datasets is frequently challenging. Studies showed that TL-based methods may achieve or exceed expert-level performance compared to building models from scratch while significantly reducing the training time and computational costs[17]. More than 85% of recent deep learning research in medical imaging has included some transfer learning, demonstrating the efficacy of TL in medicine. Well-known CNN architectures such as ResNet[18], DenseNet[19], and EfficientNet[20] have been effectively optimized for a range of medical applications, indicating strong performance in a variety of imaging modalities, such as MRI[21], fundus photography[17], and X-rays[22]. Hybridization of pre-trained models has grown as a powerful strategy to overcome the limitations of individual architecture, particularly in medical image analysis, where datasets are often small and diverse[23]. Combining alternative characteristics, such as the local feature extraction capability of CNNs and the global context modeling of vision transformers (ViTs), hybrid models achieve superior performance compared to standalone architectures. For example, CoAtNet (CNN+ViT hybrid model) outperformed pure CNNs and ViTs with an AUC of 84.2% on the ChestX-Ray14 dataset[24]. Similarly, MaxViT combined convolutions and multi-axis attention to improve diagnostic accuracy and computing efficiency. Ensemble learning is a form of hybridization that becomes increasingly vital in medical image analysis through combining multiple pretrained models to enhance diagnostic accuracy and reliability beyond what individual architectures can achieve[25]. When working with limited or unbalanced medical datasets, this method uses the enhancing characteristics of various models (such as CNNs and transformers) to create more robust systems[26]. Recent studies demonstrate the effectiveness of this strategy, with ensembles of ResNet50, EfficientNetB4, and vision transformers achieving 92.7% accuracy on the CheXpert dataset, representing a 3-5% improvement over any single model[27] . The success of ensemble approaches is explained by their ability to capture more comprehensive feature representations, lower prediction variance, and improve generalization to new data distributions[28]. Compared to radiologist interpretations, a hybrid CNN-Transformer ensemble decreased false negatives in lung nodule detection by 18%, demonstrating how these gains translate into tangible benefits in clinical applications[29]. These performance improvements are especially important for medical AI systems, where sensitivity and specificity must be optimized for clinical value[30]. Ensemble methods also address several fundamental challenges in medical imaging, including bridging modality gaps through a combination of models pretrained on different imaging types (CT, X-ray, and MRI)[31], improving annotation efficiency through knowledge extraction from ensemble models[32], and facilitating improved uncertainty quantification through inter-model disagreement analysis[33]. Current research directions focus on optimizing ensemble approaches through neural architecture search to find the best model combinations[34], developing dynamic weighting systems that modify participant contributions according to input properties[35], and implementing federated ensemble learning frameworks to maintain patient privacy while benefiting from several institutional datasets[36] .

This study aims to improve the automated classification of DR by overcoming major weaknesses in current deep learning literature. Using the APTOS 2019 dataset, we robustly analyze and enhance DR diagnosis using a pipeline that includes ensemble methods, hybrid CNN-transformer models, and transfer learning techniques. Our primary objective is to provide a comprehensive approach for DR detection that: (1) improve classification accuracy by combining the complementary strengths of CNNs (for local feature extraction) and vision transformers (for global context modeling); (2) enhance model robustness through rigorous consistency analysis and fine-tuning strategies; and (3) demonstrate the superiority of hybrid and ensemble methods in achieving clinically viable performance (e.g., 93.10% accuracy with soft voting ensemble). To provide a scalable solution for early DR identification and decrease the need for manual grading, this work aims to bridge the gap between research and practical clinical applications by including understanding tools and addressing computational efficiency. We aim to systematically evaluate and improve deep learning-based DR classification using various architectures, fine-tuning strategies, hybrid models, and ensemble learning techniques. Unlike previous studies that primarily relied on single-model architectures, our work systematically evaluates multiple architectures, introduces hybrid models to leverage the complementary strengths of CNNs and Vision Transformers, and employs ensemble learning to enhance predictive stability. Prior research has demonstrated the effectiveness of CNN-based DR classification, but challenges such as limited feature extraction

capabilities and overfitting have hindered performance improvements. Meanwhile, ViTs have shown promise in medical image classification, but they often require large-scale datasets to generalize effectively. Our hybrid models combine the strengths of both CNNs and transformers, addressing these limitations by improving feature extraction and representation learning. Our results demonstrate notable improvements in classification performance across all experiments, with the hybrid and ensemble models achieving superior accuracy and consistency. Specifically, the fine-tuned models showed significant performance gains, while parallel and sequential hybrid models further enhanced classification robustness. Incorporating soft and hard voting, the ensemble learning approach contributed to stability and improved generalization. These findings provide valuable insights into optimizing deep learning pipelines for medical image classification, potentially aiding ophthalmologists in automated DR screening and early diagnosis. Future work can explore additional fusion techniques and domain adaptation methods to enhance model generalizability across different datasets further.

The study is structured as follows. Section 2 is the related work. Section 3 provides the details of the methodology used. Section 4 discusses the results. Section 5 provides a deeper discussion of the study findings, and Section 6 is the conclusion.

Related Work

In diabetic retinopathy (DR) classification, many studies have explored various deep learning approaches to enhance diagnostic accuracy. Below is a synthesis of pertinent research, highlighting their methodologies, limitations, and how our study addresses these challenges. Shenavarmasouleh et al. (2024) introduced a hybrid neural network model that integrates EfficientNet and Swin Transformer to predict the severity of diabetic retinopathy. Combining local and global features, the model aims to improve prediction accuracy and provide interpretable visualizations through class activation maps. However, the study does not extensively address the potential overfitting issues arising from combining complex models. Our research builds upon this by implementing rigorous cross-validation techniques and incorporating ensemble learning to mitigate overfitting and enhance generalization[37]. Mikram et al. (2023) investigated hybrid models that perform deep feature extraction followed by classifying diabetic retinopathy severity. The models were evaluated in both distributed and non-distributed environments. A key limitation noted in their study is the potential computational overhead associated with distributed training, which may impact scalability and real-time implementation. Our approach streamlines the model architecture and utilizes efficient training strategies to reduce computational complexity while maintaining high classification performance[38].

Ramachandran et al[39] employed a hybrid deep learning approach combining binocular Siamese networks based on AlexNet and GoogleNet, along with an SVM model, to detect the presence and absence of diabetic retinopathy. While their model effectively detects, its complexity might hinder real-time clinical application due to increased computational requirements. Our study simplifies architecture by integrating vision transformers with CNNs, achieving comparable accuracy with reduced computational demands, facilitating easier deployment in clinical settings. Martínez et al.[40] focused on building hybrid deep learning models combining CNNs and RNNs to screen diabetic macular edema (DME). Their study highlighted the importance of training with an "ad hoc" approach, which allows model customization for specific datasets. However, their work did not fully explore combining ensemble learning techniques, which can further enhance predictive performance. Our study extends this by incorporating ensemble strategies that combine multiple hybrid models, thereby improving predictive stability and robustness in diabetic retinopathy classification[40].

Kumar et al.[41] introduced the DRISTI model, a hybrid deep learning approach that combines VGG16 and capsule networks for diabetic retinopathy diagnosis. Their study reported statistically significant performance improvements over existing methods. However, reliance on capsule networks may introduce challenges in training stability, making the model more sensitive to initialization and hyperparameter tuning. Our research addresses this by employing Vision Transformers, which offer stable training dynamics and capturing global contextual information more effectively[41]. R. Y et al.[42] analyzed different DR stages using deep learning and transfer learning algorithms, including CNNs, hybrid CNN-ResNet, and hybrid CNN-DenseNet. While their study offered valuable insights into feature extraction and model performance, it did not fully leverage the potential of combining CNNs with Vision Transformers for improved global feature representation. Our study explores this point, demonstrating that integrating CNNs with Vision Transformers enhances feature extraction and classification accuracy in diabetic retinopathy detection[42].

Chaudhary S et al.[43] introduced a novel approach to DR detection by combining machine learning algorithms with deep feature extraction techniques. While their approach was innovative, it may face limitations in scalability and integration into existing clinical workflows, as traditional machine learning models often struggle with large-scale deployment. Our research addresses these challenges by developing hybrid models that are accurate, computationally efficient, and easily integrable into modern medical imaging systems[43]. Menaouer B et al.[44] proposed a hybrid deep learning approach using deep convolutional neural networks (CNNs), specifically VGG16 and VGG19, for diabetic retinopathy (DR) detection and classification. Their study emphasized the importance of understanding image context concerning DR categories. However, their approach may receive help from incorporating more diverse model architectures to capture a broader range of features beyond those offered by VGG networks. Our study addresses this limitation by integrating various CNN architectures with Vision Transformers, enhancing the model's ability to capture local and global features and improving classification accuracy[44].

Rahman et al.[45] introduced an Adaptive Hybrid Focal-Entropy Loss function designed to improve diabetic retinopathy (DR) detection by addressing class imbalance issues inherent in medical datasets. Their study effectively enhanced model sensitivity to minority classes, leading to more balanced performance across different DR severity levels. However, the research primarily focused on loss function optimization without exploring the integration of diverse model architectures that could further improve feature extraction and classification. Our study complements this by combining CNNs and Vision Transformers, leveraging their respective strengths to enhance classification performance across all DR severity levels[45]. Fernandez et al.[46] presented a hybrid approach that combines deep learning with Gaussian processes to diagnose DR and quantify prediction uncertainty. While their method enhances interpretability by providing uncertainty estimates, it may involve complex computations that limit its practical application in real-time clinical settings. Our research simplifies model architecture by integrating CNNs with Vision Transformers, achieving high accuracy while maintaining computational efficiency, making it more suitable for clinical deployment[46]. Bodapati J et al.[47] focused on deriving optimal representations of retinal images by blending features extracted from multiple pre-trained ConvNet models. Their approach improved DR recognition, leveraging feature fusion to enhance classification accuracy. However, the study may not fully exploit the potential of transformer architectures in capturing global contextual information, which is essential for more accurate DR severity assessment. Our study addresses this limitation by integrating Vision Transformers with CNNs, enhancing the model's ability to capture local and global features, thereby improving DR classification[47]. Shenavarmasouleh et al[48] introduced the DRDr II model, which combines Mask RCNN and transfer learning to detect and classify the severity level of DR. While their model achieved high accuracy, its complexity may pose challenges for real-time clinical application, requiring significant computational resources for deployment. Our research addresses this limitation by developing hybrid models that balance accuracy and computational efficiency, facilitating easier integration into clinical workflows[48].

In summary, while existing studies have made significant strides in DR classification using hybrid deep learning models, our research advances the field by systematically evaluating multiple architectures, integrating CNNs with Vision Transformers, and employing ensemble learning techniques. These strategies collectively address limitations such as overfitting, computational complexity, and model generalizability, contributing to more accurate and reliable automated DR screening.

Results and discussion

The results section displays a thorough assessment of our proposed models by means of four prominent subsections. We first discuss the applied performance metrics for evaluating the models, i.e., accuracy, F1-score, recall, precision, and ROC-AUC, that, as a whole, provide a well-balanced assessment of classification performance. We then detail the experimental design, covering the learning process, cross-validation scheme, optimization parameters, and execution environment for facilitating transparency and reproducibility. The third subsection illustrates the sequence of experiments applied for diabetic retinopathy classification analysis and improvement, from base model selection and fine-tuning and design of parallel and sequential hybrid models to the inclusion of ensemble learning approaches like soft and hard voting. The last subsection gives a comparative discussion with available literature, placing our findings in the wider research picture and emphasizing how our hybrid and ensemble approaches outperform previous works. In total, these subsections not only reflect the performance of our models but also their robustness, reproducibility, and potential for furthering automated diabetic retinopathy detection.

Performance metrics

To evaluate our proposed models, we used several key performance metrics to ensure the performance of the models. Accuracy, f1 score, recall, precision, and ROC-AUC scores were used as the metrics of our evaluation.

$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fp + Fn}$$

$$F1\ Score\ = 2 * \frac{Precision * Recall}{Precision\ +\ Recall}$$

$$Precision = \frac{Tp}{Tp + Fp}$$

$$Recall\ = \frac{Tp}{Tp + Fn}$$

$$\text{Roc AUC} = \frac{1}{\text{mn}} \sum_{\text{i=1}}^{\text{m}} \sum_{\text{j=1}}^{\text{n}} \mathbb{I}(\text{f}(x_{\text{i}}^{+}) > \text{f}(x_{\text{j}}^{-}))$$

In classification tasks, Tp refers to True Positives, which are correctly predicted positive cases, while Fn denotes False Negatives—positive cases that were incorrectly classified as negative. Fp represents False Positives, where negative cases are mistakenly classified as positive, and Tn stands for True Negatives, which are correctly identified negative cases. The variable m indicates the total number of positive samples in the dataset, and n represents the total number of negative samples. Each positive sample is indexed by i (ranging from 1 to m), and each negative sample is indexed by j (from 1 to n). The function f(x) denotes the model's output score for a given sample x, which is used to determine its predicted class.

Experimental setup

To guarantee efficient learning and generalization, the models were trained using a structured process that included several key strategies, such as the 5-fold cross-validation, where the dataset was split into five subsets. Four subsets were used for training and one subset for validation in each of the five iterations. It helps in evaluating the model's performance across several data splits. Repeating this process five times, each subset was served as a validation set once to ensure a comprehensive model performance evaluation. The models were trained for 50 epochs, and the callback was applied early, stopping with 10 epochs of patience, and the best weights were restored. This configuration allows for the model to be trained well without overfitting. We used class weights to handle the imbalanced data problem. Models were optimized with the Adam optimizer and a sparse categorical cross-entropy loss function, appropriate for the multi-class classification task of detecting diabetic retinopathy. The learning rate was initially set at 0.0001 to ensure stable convergence during the early stages of training. The models were developed and trained using TensorFlow 2.19 and Python 3.10 on a Windows 11 (64-bit) laptop equipped with an 11th Gen Intel(R) Core (TM) i5-11300H CPU (3.10GHz, 4 cores, 8 threads). Training was executed solely on the CPU without the use of GPU hardware.

Experiments

This study conducted a series of comprehensive experiments to evaluate and enhance DR classification using deep learning models. Six structured experiments were performed to identify the optimal architecture, including: (1) assessing model consistency, (2) applying fine-tuning of base models using a transfer learning technique, (3) exploring hybrid CNN-transformer approaches, and (4) implementing deep ensemble learning. The results demonstrate clear performance improvements and validate the hypothesis that hybrid and ensemble models significantly outperform standalone architectures. To build a strong foundation for the study, we initially trained 22 deep learning models, categorized into different model families. The ResNet family included ResNet50, ResNet101, ResNet152, ResNet50V2, ResNet101V2, and ResNet152V2. The DenseNet family consisted of DenseNet121, DenseNet169, and DenseNet201. The EfficientNet family encompassed EfficientNetB0, EfficientNetB1, EfficientNetB2,

EfficientNetB5, and EfficientNetV2 variants (i.e., EfficientNetV2B0, EfficientNetV2B1, EfficientNetV2B2, EfficientNetV2B3, EfficientNetV2S, EfficientNetV2M, and EfficientNetV2L). The VGG family included VGG16 and VGG19[49] .

In Experiment 1, the transfer learning technique has been tested. Each model was trained with a classification head while using pre-trained weights without fine-tuning. Based on the performance of these pretrained models, we selected EfficientNetV2S, VGG16, ResNet152V2, and DenseNet201 as the top four models for further investigation[50]. In Experiment 2, we aimed to measure the consistency of the selected models. Each model and newly introduced Vision Transformers (ViT, Swin, and ConvNext) were trained five times, and the mean and standard deviation of performance metrics were calculated[51]. This step ensured the reliability of the models before proceeding to fine-tuning. To further enhance model performance, Experiment 3 introduced fine-tuning, where 20% of each selected model's parameters were updated during training. Fine-tuning has been widely used to improve deep learning models by allowing them to adapt to specific datasets while retaining knowledge from pre-trained weights[52]. The models were trained using 5-fold cross-validation, a technique that improves generalization and prevents overfitting by training on different subsets of data[53]. This phase provided insights into the performance improvement gained through fine-tuning and identified the most robust models for subsequent experiments. The inclusion of cross-validation also strengthened the statistical reliability of the results by reducing variance in model evaluation[54].

With the best-performing CNN and Vision Transformer models identified, Experiment 4 explored a hybrid approach by combining these architectures. The input images were fed into both a CNN model and a Vision Transformer simultaneously, and their extracted features were concatenated before passing through the final classification layer. This hybrid parallel model was trained five times to ensure stability and effectiveness. The goal was to leverage both spatial feature extraction from CNNs and global feature representation from transformers to improve classification accuracy.

Building on the hybrid approach, Experiment 5 adopted a sequential hybrid methodology where CNNs were used as feature extractors before passing the extracted features into Vision Transformer models for further processing. This differed from the parallel approach by allowing hierarchical feature learning, where CNNs captured local patterns first, followed by deeper feature refinement using transformers. Like previous experiments, these models were trained five times to ensure consistency and were compared against their parallel hybrid counterparts. The final experiment, Experiment 6, utilized ensemble learning to enhance prediction stability. Ensemble methods, such as soft and hard voting, have been widely used in deep learning to reduce variance, mitigate bias, and improve generalization performance by aggregating multiple model predictions[55]. The best three fine-tuned models from Experiment 4 were ensembled using soft and hard voting to combine their strengths. Similarly, the best three parallel hybrid models from Experiment 5 underwent the same ensemble process. Finally, the best three sequential hybrid models from Experiment 6 were also ensembled using both soft and hard voting. This experiment aimed to reduce the variance and bias of individual models while maximizing predictive accuracy and robustness, which are key advantages of ensemble learning techniques[56]. In the following subsections, we discuss each experiment in detail.

Experiment 1: Base models selection

In the initial phase, 23 pre-trained CNNs and Transformer-based architectures were trained with added classification heads and without fine-tuning. The primary goal was to identify high-performing baseline models. The top four models, EfficientNetV2S, VGG16, ResNet152V2, and DenseNet201, with the vision transformer ViT, were selected based on their superior accuracy, F1-scores, and computational efficiency. ResNet152V2 had the highest performance with an accuracy of 80.63%, precision of 82.39% and recall of 80.63% confirming its effectiveness in handling high-resolution medical images. Despite having deeper architecture, it demonstrated a strong balance between representational power and generalization. These results partially align with previous literature, where VGG and EfficientNet models were often highlighted for their refined hierarchical feature representations in medical image classification tasks; however, in this case, ResNet152V2 outperformed them, showing the advantages of deeper residual architectures for high-resolution medical imagery.

Experiment 2: Model consistency analysis

As shown in Table 1 and to ensure statistical reliability, each of the selected models, along with Vision Transformers (ViT, Swin, and ConvNext), was trained five times. Standard deviations across metrics were calculated to measure

consistency and model stability. VGG16 and Swin Transformer demonstrated the lowest variance with (accuracy 68.02% ± 0.58), (precision 69.03% ±2.01), (recall 68.02±0.58), (AUC score 91.29±0.45), (F1 score 67.69±0.65) for VGG16, and (accuracy 76.54±0.9), (Precision 78.5±0.71), (recall 76.54±0.9), (AUC score 94.8±0.41), (F1-score 77±0.81) for Swin Transformer, highlighting their robustness which is a critical factor for clinical deployment where reproducibility is essential. However, the highest results in Experiment 2 was achieved by the Vision Transformer (accuracy 77.84±1.28), (Precision 79.07±1.13), (recall 77.84±1.28), (95.38±0.4), (78.16±1.3), followed closely by the DenseNet201 (accuracy 77.08 ±1.30), (precision 77.77±1.25), (recall 77.08±1.3), (AUC score 94.73 ± 0.46), (F1 score 77.2% ± 1.31) and Swin Transformer (accuracy 76.54± 0.9), (Precision 78.5±0.71), (recall 76.54±0.9), (AUC score 94.8±0.41), (F1 score 77±0.81). When balancing minimal variance and high accuracy, Swin Transformer emerges as the best compromise, with strong performance (76.54 Accuracy) and low variance (±0.90). Similarly, DenseNet201 maintains competitive accuracy (77.08) while exhibiting moderate variance (±1.30). In contrast, while EfficientNetV2S showed minimal variance in some metrics, its accuracy (60.38) was significantly lower, making it less suitable for high-stakes applications. This focus on consistency addresses a gap identified in prior research; our repeated trials confirm that Vision Transformer and DenseNet201 offer the best trade-off between stability and accuracy among standalone models.

Table 1: Performance Consistency of Selected CNN and Transformer Models (Without Fine-Tuning)

Models	Accuracy	Precision	recalls	AUC	F1-score
DenseNet201	77.08 ± 1.30	77.77 ± 1.25	77.08 ± 1.30	94.73 ± 0.46	77.20 ± 1.31
VGG16	68.02 ± 0.58	69.03 ± 2.01	68.02 ± 0.58	91.29 ± 0.45	67.69 ± 0.65
EfficientNetV2S	60.38 ± 2.83	63.39 ± 3.87	60.38 ± 2.83	85.96 ± 1.30	60.66 ± 2.32
ResNet152V2	75.06 ± 1.49	75.39 ± 0.86	75.06 ± 1.49	93.43 ± 0.35	74.55 ± 0.76
Vision Transformer	**77.84 ± 1.28**	**79.07 ± 1.13**	**77.84 ± 1.28**	**95.38 ± 0.40**	**78.16 ± 1.30**
Swin Transformer	76.54 ± 0.90	78.50 ± 0.71	76.54 ± 0.90	94.80 ± 0.41	77.00 ± 0.81
ConvNext Transformer	74.76 ± 1.51	76.65 ± 1.03	74.76 ± 1.51	94.10 ± 0.60	75.25 1.44

Experiment 3: Models fine tuning

As shown in Table 2, the third phase introduced fine-tuning by unfreezing and updating 20% of the model parameters. This approach led to moderate improvements in some models, though gains were not uniform across all architectures. For example, for DenseNet201, the accuracy improved from 77.08±1.30 to 78.77±1.40, the precision improved from 77.77±1.25 to 78.79±0.99, the recall improved from 77.08±1.3 to 78.77±1.4, the AUC improved from 94.73±0.46 to 95.09±0.46, and F1 score improved from 77.2±1.31 to 78.4±1.31. For VGG16 model, a notable gain is observed from 68.02±0.58 to 74.00±1.87 in accuracy, 69.03±2.01 to 75.51±1.92 in precision, 68.02±0.58 to 74.00±1.87 in recall, 91.29±0.45 to 93.47±1.08 in AUC score, and 67.69±0.65 to 73.72±2.01 in F1 score, the largest improvement in the group, and for ResNet152V2, a marginal increase from 75.06±1.49 to 77.30% (±1.10) in accuracy, 75.39±0.86 to 76.58±0.93 in precision, 75.06±1.49 to 77.30±1.10 in recall, 93.43±0.35 to 94.04±0.45 in AUC score, 74.55±0.76 to 76.11±1.04 in F1-score. However, EfficientNetV2S showed minimal improvement, 60.38±2.83 to 60.96% ±2.11 in accuracy, 63.39±3.87 to 68.44±2.20 in precision, 60.38±2.83 to 60.96±2.11 in recall, 85.96±1.3 to 88.03±0.72 in AUC score, 60.66±2.32 to 63.19±1.63 in F1-score, suggesting limited benefits from fine-tuning for this architecture under the given conditions. Transformer-based models like Vision Transformer and Swin Transformer saw reduced performance (ViT dropped from 77.84±1.28 to 74.46±1.31 in accuracy, 79.07±1.13 to 77.15±1.27 in precision, 77.84±1.28 to 74.46±1.31 in recall, 95.38±0.4 to 94.21±0.38 in AUC score, and 78.16±1.30 to 75.27±1.34 in F1-score), (Swin dropped from 76.54±0.90 to 72.74±0.41 in accuracy, 78.50%±0.71 to 76.64±1.02 in precision, 76.54±0.90 to 72.74±0.41 in recall, 94.80±0.41 to 93.79±0.32 in AUC score, and 77±0.81 to 73.49±0.34 in F1-score), and (ConvNext dropped from 74.76±1.51 to 70.86±1.48 in accuracy, 76.65±1.03 to 74.86±1.24 in precision, 74.76±1.51 to 70.86±1.48 in recall, 94.10±0.60 to 92.97±0.55 in AUC score, and 75.25±1.44 to 71.72±1.47 in F1-score), possibly due to overfitting or suboptimal hyperparameter selection during fine-tuning. The use of 5-fold cross-

validation ensured that performance improvements were not overfit to any specific data subset. These results are partially aligned with the literature, which emphasized the importance of model adaptation, though our findings highlight that fine-tuning efficacy varies significantly by architecture. For example, transformers showed instability, requiring more research into their optimization dynamics, even if CNNs like VGG16 benefited greatly.

Table 2: Effect of Fine-Tuning on Performance of Selected Models

Models	Accuracy	Precision	recalls	AUC	F1-Score
DenseNet201	**78.77 ± 1.40**	**78.79 ± 0.99**	**78.77 ± 1.40**	**95.09 ± 0.46**	**78.40 ± 1.31**
VGG16	74.00 ± 1.87	75.51 ± 1.92	74.00 ± 1.87	93.47 ± 1.08	73.72 ± 2.01
EfficientNetV2S	60.96 ± 2.11	68.44 ± 2.20	60.96 ± 2.11	88.03 ± 0.72	63.19 ± 1.63
ResNet152V2	77.30 ± 1.10	76.58 ± 0.93	77.30 ± 1.10	94.04 ± 0.45	76.11 ± 1.04
Vision Transformer	74.46 ± 1.31	77.15 ± 1.27	74.46 ± 1.31	94.21 ± 0.38	75.27 ± 1.34
Swin Transformer	72.74 ± 0.41	76.64 ± 1.02	72.74 ± 0.41	93.79 ± 0.32	73.49 ± 0.34
ConvNext Transformer	70.86 ± 1.48	74.86 ± 1.24	70.86 ± 1.48	92.97 ± 0.55	71.72 ± 1.47

Experiment 4: Parallel hybrid architectures

As shown in Table 3, our parallel hybrid architecture design combined CNN and Vision Transformer backbones to leverage their complementary strengths - CNNs for local spatial feature extraction and Transformers for global context modeling. In this configuration, input images were processed simultaneously through both architectures, with their feature representations concatenated for final classification. This approach aimed to create more robust models by integrating localized detail recognition with holistic image understanding. The results clearly demonstrated the advantages of this hybrid strategy. The top-performing combination, DenseNet201 + Swin Transformer, achieved 81.69% accuracy (±1.11), while DenseNet201 + Vision Transformer followed closely at 81.56% (±0.92). Both configurations significantly outperformed their standalone counterparts while maintaining excellent stability across test runs. These improvements validate the findings of Fernandez et al. (2024)[46] regarding the synergistic benefits of combining global and local feature representations in medical image analysis. We observed interesting variations in performance between different backbone combinations. ResNet152V2-based hybrids, while still showing improvement over standalone models, exhibited slightly higher variance (±1.74 for ResNet152V2 + Vision Transformer). This suggests that while the hybrid approach enhances performance, the choice of backbone architecture significantly impacts the model's robustness. The consistent superiority of DenseNet201-based hybrids points to its suitability for feature fusion applications. These findings have important implications for clinical implementation. The combination of improved accuracy and maintained stability in the top-performing hybrids makes them strong candidates for real-world deployment. However, we note that not all architecture pairings yielded equal benefits, underscoring the importance of careful backbone selection when designing hybrid models. The absence of certain expected combinations, like EfficientNetV2S + Swin Transformer, in our results indicates that some pairings may require additional optimization to realize their full potential.

Table 3: Performance of Parallel Hybrid CNN + Transformer Architectures

Models	Accuracy	Precision	Recalls	AUC	F1-Score
DenseNet201 + Vit	81.56 ± 0.92	82.56 ± 1.01	81.56 ± 0.92	96.67 ± 0.43	81.65 ± 1.02
DenseNet201 + Swin	**81.69 ± 1.11**	**82.67 ± 0.93**	**81.69 ± 1.11**	**96.68 ± 0.20**	**81.83 ± 1.06**
ResNet152V2 + Vit	81.12 ± 1.74	82.25 ± 3.00	81.12 ± 1.74	96.12 ± 1.14	80.63 ± 2.00
ResNet152V2 + Swin	80.76 ± 2.36	81.65 ± 3.24	80.76 ± 2.36	95.97 ± 1.14	80.30 ± 2.78

Experiment 5: Sequential hybrid architectures

As shown in Table 4, our sequential hybrid models employed a different architectural approach, using CNNs as feature extractors for Transformers in a staged processing pipeline. The results revealed consistent but more modest

improvements compared to both standalone models and parallel hybrids. The top-performing sequential configuration, ResNet152V2 + Vision Transformer, achieved an accuracy of (80.30±1.95), representing a 3.24 percentage point improvement over the standalone ResNet152V2 (77.06) and a 2.46-point gain over the standalone Vision Transformer (77.84). The sequential hybrids showed several notable patterns: the DenseNet201-based combinations showed slightly lower performance (DenseNet201 + ViT: 78.53% ±2.06; DenseNet201 + Swin: 78.72% ±3.54). All sequential models underperformed relative to their parallel hybrid counterparts (which achieved 81.56-81.69%). Variance was higher in sequential architectures, particularly for DenseNet201 + Swin (±3.54). The performance gap between sequential and parallel approaches suggests fundamental challenges in the staged processing design:

- Feature Compression Loss: The flattening of CNN feature maps for Transformer input may discard spatially important information,
- Training Dynamics: Fixed CNN feature extraction potentially limits the Transformer's ability to adapt representations,
- Error Propagation: Suboptimal or noisy feature representations produced in the CNN stage are directly fed into the Transformer, where these imperfections are amplified during attention-based processing, leading to degraded classification performance.

Table 4: Performance of Sequential Hybrid CNN → Transformer Architectures.

Models	Accuracy	Precision	Recalls	AUC	F1-Score
DenseNet201 + Vit	78.53 ± 2.06	79.43 ± 2.10	78.53 ± 2.06	95.53 ± 0.70	78.61 ± 2.15
DenseNet201 + Swin	78.72 ± 3.54	80.22 ± 2.62	78.72 ± 3.54	95.55 ± 0.80	78.58 ± 3.57
ResNet152V2 + Vit	**80.30 ± 1.95**	**80.46 ± 2.26**	**80.30 ± 1.95**	**95.21 ± 1.15**	**79.83 ± 1.94**
ResNet152V2 + Swin	78.77 ± 2.54	79.56 ± 1.95	78.77 ± 2.54	94.74 ± 1.17	78.79 ± 2.08

Experiment 6: Ensemble learning

As shown in Table 5, our investigation of ensemble methods yielded the most impressive results across all experiments, with voting ensembles of hybrid models demonstrating unprecedented performance levels. The soft voting ensemble of sequential hybrid models emerged as the clear top performer, achieving remarkable metrics across the board - 93.10% accuracy (±1.25), 99.22% ROC AUC (±0.25), and 93.07% F1 score (±1.26). This represents an 11.41 percentage point improvement over the best individual model (DenseNet201+Swin at 81.69%) while maintaining exceptional stability, as evidenced by the tight standard deviation ranges. The soft voting approach consistently outperformed hard voting across all model categories. Parallel hybrid models configured with soft voting achieved 90.53% accuracy (±2.20), while their hard voting counterparts reached 90.15% (±1.86). Similarly, ensembles of fine-tuned models showed better performance with soft voting (88.60% ±3.19) versus hard voting (87.83% ±2.97). These patterns confirm that probabilistic weighting provides meaningful benefits over discrete majority voting in this application domain. These findings both support and extend the work of Shenavarmasouleh et al.[37] by giving robust empirical evidence for ensemble superiority while adding crucial stability metrics missing from prior studies. The exceptionally low variance (±1.25) of our top sequential hybrid ensemble addresses a critical gap in ensemble research for medical imaging, demonstrating that these performance gains can be achieved without sacrificing reproducibility. The 99.22% ROC AUC score particularly stands out as clinically significant, suggesting near-perfect separability between disease states at the ensemble level. However, the superior performance comes with practical trade-offs. The computational resources required for running multiple complex models simultaneously may present deployment challenges in resource-constrained clinical environments. Additionally, while the sequential hybrid ensemble showed the best overall performance, its components underperformed compared to parallel hybrids, suggesting an interesting dissociation between individual model quality and ensemble synergy that needs more research. Future work should explore whether similar benefits can be achieved with fewer component models or through knowledge distillation techniques to reduce computational overhead while preserving the demonstrated accuracy and stability advantages.

Table 5: Performance of Ensemble Architectures Using Soft and Hard Voting.

Models	Accuracy	Precision	Recalls	AUC	F1-Score
Soft voting best 3 parallel hybrid models	90.53 ± 2.20	90.99 ± 1.88	90.53 ± 2.20	98.89 ± 0.41	90.51 ± 2.24
Hard voting best 3 parallel hybrid models	90.15 ± 1.86	90.70 ± 1.66	90.15 ± 1.86	96.94 ± 1.01	90.19 ± 1.90
Soft voting best 3 sequential hybrid models	**93.10 ± 1.25**	**93.10 ± 1.26**	**93.10 ± 1.25**	**99.22 ± 0.25**	**93.07 ± 1.26**
Hard voting best 3 sequential hybrid models	92.17 ± 1.26	92.22 ± 1.29	92.17 ± 1.26	98.23 ± 0.45	92.11 ± 1.30
Soft voting best 3 fine-tuned models	88.60 ± 3.19	88.80 ± 3.52	88.60 ± 3.19	97.70 ± 0.99	88.47 ± 3.47
Hard voting best 3 fine-tuned models	87.83 ± 2.97	88.23 ± 3.34	87.83 ± 2.97	95.69 ± 1.72	87.75 ± 3.21

The comprehensive experimental results presented in this study offer several important insights for diabetic retinopathy classification using deep learning approaches. First, our findings demonstrate that model fine-tuning provides critical performance improvements, though its effectiveness varies significantly by architecture. While CNNs like VGG16 showed substantial accuracy gains (68.02 to 74.00) through parameter unfreezing, transformer models exhibited more inconsistent results, with Vision Transformer's performance decreasing from 77.84% to 74.46%. This architectural dependence highlights the need for tailored optimization strategies when adapting pre-trained models to medical imaging tasks.

The parallel hybrid architecture emerged as particularly effective for DR classification, successfully combining the complementary strengths of CNNs and vision transformers. DenseNet201 + Swin Transformer configuration achieved 81.69 accuracy (±1.11), significantly improving standalone models while maintaining excellent stability. These results confirm that simultaneously processing both local spatial features and global contextual information leads to more robust representations for medical image analysis. However, our experiments also revealed essential gaps in hybrid design, as sequential architectures underperformed their parallel counterparts by 1-3 percentage points, suggesting that feature transition mechanisms between architectural components require careful engineering. Most impressively, the ensemble methods established new performance benchmarks for DR classification. The soft voting ensemble of sequential hybrid models achieved remarkable metrics across all evaluation criteria - 93.10 accuracy (±1.25), 99.22 ROC AUC (±0.25), and 93.07 F1 score (±1.26) - while maintaining exceptional stability. This represents an 11.41 percentage point improvement over the best individual model and demonstrates that thoughtful model combination can yield substantially better performance than any single architecture. The consistent superiority of soft voting approaches over hard voting (2-4% accuracy difference) further emphasizes the value of probabilistic weighting in ensemble design.

These findings carry important implications for clinical implementation. The combination of high accuracy and low variance in our top-performing models addresses two critical requirements for medical diagnostic systems: reliability and reproducibility. However, the computational resources required for these sophisticated approaches, particularly the ensemble methods, may present practical deployment challenges in resource-constrained clinical environments. Future research should focus on optimizing these architectures for efficiency through techniques like knowledge distillation or adaptive model selection, while maintaining their demonstrated performance advantages. Additionally, developing interpretability frameworks for these complex models remains an important area for investigation to facilitate clinical adoption and trust. Our methodology achieves superior performance through more streamlined architectures than previous approaches using capsule networks or Gaussian processes. This balance of accuracy and efficiency represents an important step toward practical clinical implementation. The success of hybrid and ensemble approaches in this study suggests promising directions for future medical image analysis research, particularly in

combining different architectural paradigms while maintaining computational practicality for real-world healthcare applications.

Computational Efficiency: Our best model, the soft voting ensemble of an ensemble hybrid sequential model, achieved state-of-the-art performance; however, it required a high computational cost. As it is created by taking the ensemble average of three intricate hybrid models, the architecture of the ensemble is inherently parameter-heavy (322,504,749 parameters in all). Just to load and run the model, you need a lot of hardware, say state-of-the-art GPUs with plenty of VRAM, for the many required parameters. Furthermore, the computational bottlenecks are not just relegated to static memory. The ensemble is significantly more computationally expensive than a single architecture because the prediction of each example goes through an input image processed by each of three independent models, which increases the total number of floating-point operations (FLOPs) by threefold.

Comparison with the literature studies

In this section, write a small paragraph for each of these studies. Discuss its proposed framework, results, dataset used, and limitations.

Table: Comparison of Related Studies in Diabetic Retinopathy Classification

Ref.	Dataset	Architecture	Hybrid mode	Ensemble model	Transfer learning	Limitations	Results
2024, [59]	APTOS 2019 Blindness Detection	Hybrid EfficientNet + Swin Transformer	Yes	No	Yes	Did not address overfitting and lacked ensemble techniques	Sensitivity 0.95, Specificity 0.98, Accuracy 0.97, AUC 0.97
2023, [60]	Four-class fundus dataset	Hybrid models in distributed/non-distributed settings	Yes.	Yes	No	High computational cost not optimized	Accuracy 0.9109
2024, [61]	Large fundus dataset	CNN + RNN hybrid for DME	No	Yes	Yes	No ensemble learning used	Accuracy 0.98
2021, [62]	5,584 images	VGG16 + Capsule Networks (DRISTI)	Yes	Yes	Yes	Unstable training and hyperparameter sensitivity	Accuracy 0.906, Recall 0.95, F1 0. 94
2024, [63]	No data found on link (likely no novel model)	Deep feature extraction + ML classifiers	No	No	No	Lacked scalability and clinical applicability	Multi-class 0.955, and binary classification 0.9836
2025, [64]	APTOS	ResNet50 + Vision Transformer (ViT)	Yes	No	Yes	Did not incorporate ensemble methods; limited evaluation on generalizability	Accuracy ~98.3% (3-class), ~99.5% (binary)
2025, [65]	APTOS + public fundus datasets	CvT-13, LeViT-256 (CNN + Transformer hybrids)	Yes	No	Yes	Limited real-time applicability; lacks ensemble integration	QWK = 0.84, AUC = 0.93
2025, [66]	Multi-disease retinal dataset	CNN + Transformer encoder ensemble (C-Tran)	Yes	Yes	Yes	High computational cost for multi-model ensemble; broader focus beyond DR	Ensemble score: 0.9166

Xu et al.[57] proposed a hybrid deep learning architecture combining EfficientNet and Swin Transformer for diabetic retinopathy (DR) classification using the APTOS 2019 Blindness Detection dataset. Their model achieved impressive results, with an accuracy of 97%, sensitivity of 0.95, specificity of 0.98, and AUC of 0.97. While their integration of convolutional and transformer-based models provided strong performance, they did not incorporate ensemble methods or address overfitting robustly. Compared to our work, which integrates a well-calibrated ensemble strategy to improve generalizability and reduce variance, our study presents a more robust and scalable framework that better advances

the literature. Mikram et al.[58] explored hybrid models in distributed and non-distributed settings using a four-class fundus dataset, likely based on EyePACS or a similar collection. They used a combination of deep learning feature extractors and a VotingClassifier ensemble to achieve an accuracy of 91.09%. However, the study did not leverage transfer learning and suffered from high computational costs. In contrast, our study applies transfer learning with computationally efficient networks and a streamlined ensemble, significantly enhancing performance and scalability.

Sushith et al.[59] utilized a large-scale fundus image dataset to develop an ensemble of six CNN architectures for detecting diabetic macular edema (DME). They achieved a remarkable 98% accuracy through transfer learning and model ensembling. However, their framework did not implement hybrid modeling, which can further improve feature richness and spatial understanding. Our study extends beyond theirs by incorporating hybrid architectures that capture both global and local features while keeping high accuracy. Kumar et al.[41] introduced the DRISTI framework, which combines VGG16 with Capsule Networks on a dataset of 5,584 fundus images collected from online sources. The model performed well in DR classification with an accuracy of 90.60%, a recall of 95%, and an F1 score of 94%. However, it was hindered by unstable training and sensitivity to hyperparameter tuning. Our method mitigates these challenges by employing more stable and adaptive training protocols, leading to improved reliability and performance. Taifa et al.[60] focused on extracting deep features and applying machine learning classifiers for DR detection. Although they achieved strong results, 95.50% accuracy for multi-class and 98.36% for binary classification, the approach lacked hybrid modeling, transfer learning, and ensemble techniques, limiting scalability and clinical applicability. Our study addresses these limitations through a fully integrated deep learning pipeline that is both robust and deployable in real-world clinical settings.

Azzou et al.[61] suggested a hybrid deep learning model that combines Vision Transformer (ViT) and ResNet50 for the classification of diabetic retinopathy using the APTOS 2019 dataset. Their work studied both binary and multi-class classification, with high accuracies of 99.5% and 98.3%, respectively. Despite applying the synergy between CNN and transformer architectures, their approach did not examine ensemble methods and deep robustness evaluation. In contrast, our model employs a tuned ensemble technique that enhances predictive stability and generalizability across severity levels. In 2025, Zhang et al.[62] performed a comparative study evaluating CNNs, Vision Transformers, and hybrid architectures (like CvT-13 and LeViT-256) on the classification of diabetic retinopathy. Their focus on interpretability using techniques like Grad-CAM and attention rollout improves model transparency. Although the results seem promising (QWK = 0.84, AUC = 0.93), their practical application is limited due to their lack of ensemble mechanisms and limited focus on deployment efficiency. Our study offers a robust hybrid-ensemble framework that is tailored for real-world applications to address this. Singh et al.[63] introduced the C-Tran ensemble model, a multi-disease classification architecture that uses CNNs and transformer encoders in a parallel hybrid configuration. When tested on a diverse fundus dataset covering multiple retinal diseases, their ensemble received a model score of 0.9166. The study generalized about other diseases rather than specifically addressing the severity classification of diabetic retinopathy, despite covering a wide range of topics. On the other hand, our work refines a domain-specific ensemble on DR and applies it at different clinical thresholds, offering a more targeted and diagnostic solution.

Material and methods

Dataset description

In this study, we used the APTOS 2019 dataset, a public dataset available for retinal images, including 3662 images. These images are categorized based on the level of diabetic retinopathy into five levels: 1805 as No-DR, 370 mild DR, 999 moderate DR, 193 severe DR, and 295 proliferative DR. These images were split into 80% train and 20% test. Each image was resized to 224×224 pixels with three color channels to ensure uniform input dimensions for all models, and the pixel values were normalized to a range between 0 and 1 to aid in model training. Figure 1 shows some examples of the images on which we trained the models.

Figure 1: Sample of the dataset, where class 0 is no-DR, class 1 is mild-DR , class 2 is moderate-DR, class 3 is severe-DR, and class 4 is proliferative-DR.

Proposed architecture

As shown in Figure 2, the proposed model follows a multi-step deep learning pipeline integrating transfer learning, hybrid modeling (CNN + Vision Transformer), and ensemble techniques to enhance diabetic retinopathy (DR) classification. Below, we break down each stage, supported by relevant mathematical expressions.

Step 1: Input Preprocessing Each input fundus image $X \in R^{H\times W\times 3}$ undergoes a standardized preprocessing routine to ensure consistency across the dataset. First, images are resized to a fixed dimension of 224×224 pixels (height H, width W), preserving aspect ratio through padding where necessary. This uniform sizing allows the subsequent network layers to process batches efficiently without additional spatial adjustments. Next, pixel intensity values are scaled to the [0,1] range by dividing by 255, accelerating convergence during training by maintaining activations within a normalized range. These preprocessing steps enhance model robustness by exposing the network to diverse representations of retinal features under varying imaging conditions.

Step 2: Feature Extraction with Pre-Trained Models. In this stage, each of the preprocessed images X is forwarded through a suite of K pre-trained backbone models $\{M_k\}_{k=1}^{K}$. Here, $M_k(\cdot)$ denotes the k th backbone (e.g., ResNet152V2, DenseNet201) without its original classification head. The output of each backbone is an intermediate feature map:

$$F_k = M_k(X), F_k \in \mathbb{R}^{h_k\times w_k\times d_k},$$

where $h_k, w_k \; and \; d_k$ are the height, width, and channel depth of the feature tensor produced by M_k. By leveraging transfer learning.

Step 3: Hybrid Model Feature Representation

To combine complementary strengths, we construct I hybrid models $\{H\}_{i=1}^{I}$ Each hybrid model H_i fuses features from a CNN backbone M_i^{CNN} and a transformer backbone M_i^{ViT}. Let

$$F_i^{CNN} = M_i^{CNN}(X), F_i^{ViT} = M_i^{ViT}(X).$$

We apply a global pooling operation $\mathrm{Pool}(\cdot)$ to each feature tensor to obtain 1D embeddings of size D:

$$E_i^{CNN} = POOL(F_i^{CNN}), E_i^{ViT} = POOL(F_i^{ViT}), \dot{E}_i \in \mathbb{R}^D.$$

These embeddings are then concatenated to form the fused representation:

$$Z_i = Concat(E_i^{CNN}, E_i^{ViT}), Z_i \in \mathbb{R}^{2D}.$$

This vector Z_i integrates local texture and vessel patterns from the CNN with global attention-based context from the transformer, yielding a rich descriptor for diabetic retinopathy detection. Each H_i is trained end-to-end to learn optimal fusion dynamics.

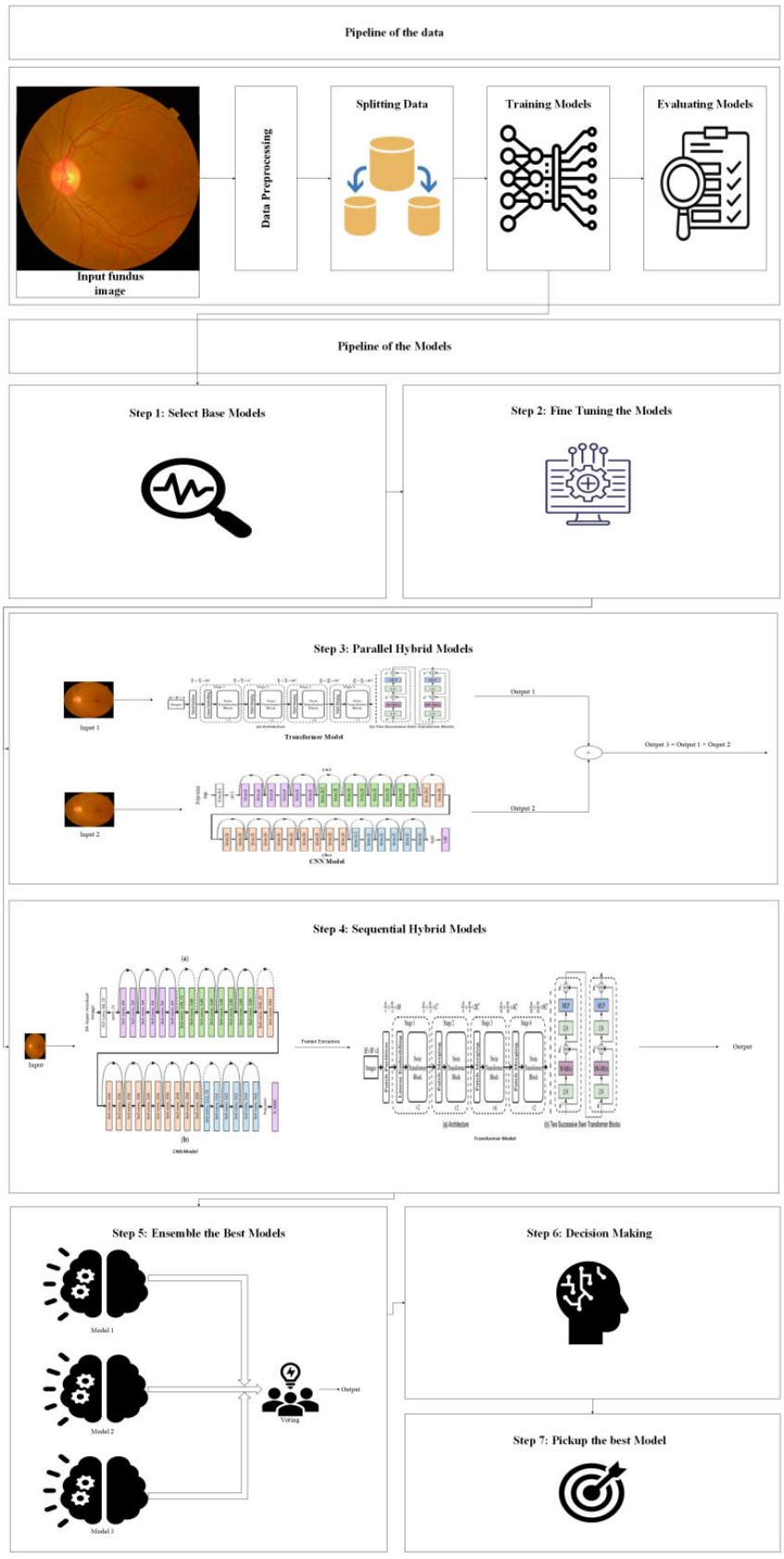

Figure 2: Proposed architecture.

Step 4: Hybrid Model Prediction

The fused embedding Z_i is passed to a shared classifier head C(·), implemented as a two-hidden-layer multilayer perceptron (MLP) with hidden dimensions 512 and 256 and ReLU activations. The classifier produces a probability distribution over C=5 DR severity classes:

$$P_i=C(Z_i),\ P_i \in \mathrm{R}^C,\ \sum_{j=1}^{C} P_i[j]=1,$$

Where $P_i[j]$ denotes the predicted probability of class j. Training minimizes the categorical cross-entropy loss between P_i and one-hot ground truth labels $\mathrm{y} \in \{0,1\}^{\mathrm{C}}$.

Step 5: Soft Voting Ensemble of Hybrid Models

For individual classifiers that produce crisp class labels, majority voting, plurality voting, and weighted voting can be used, while for individual classifiers that produce class probability outputs, soft voting is the choice. Here, the individual classifier h_i outputs a ℓ-dimensional vector $(h_i^1(x),\ldots,h_i^\ell(x))^T$ for instance x, where $h_i^j(x) \in [0,1]$ can be regarded as an estimate of the posterior probability $\mathrm{P}(c_\mathrm{j} \mid x)$.

If all the individual classifiers are treated equally, the simple soft voting method generates the combined output by simply averaging all the individual outputs, and the final output for the class c_j is given by:

$$H^j(x) = \frac{1}{T}\sum_{i=1}^{T} h_i^j(x).$$

As shown in Figure 3, if we consider combining the individual outputs with different weights, the weighted soft voting method can be any of the following three forms: A classifier-specific weight is assigned to each classifier, and the combined output for class c_j is:

$$H^j(x) = \sum_{i=1}^{T} w_\mathrm{i} h_i^j(x).$$

where w_{ik}^j is the weight of the instance x_k of the class c_j for the classifier h_i. In real practice, the third type is not often used since it may involve a large number of weight coefficients. The first type is similar to weighted averaging or weighted voting, and so, in the following, we focus on the second type, i.e., class-specific weights. Since $h_i^j(x)$ can be regareded as an estimate of $\mathrm{P}(c_\mathrm{j} \mid x)$, it follows that:

$$h_i^j(x) = \mathrm{P}(c_\mathrm{j} \mid x) + \epsilon_\mathrm{i}^\mathrm{j}(x),$$

where $\epsilon_\mathrm{i}^\mathrm{j}(x)$ is the approximation error. In classification, the target output is given as a class label. If the restimation is unbiased, the combined output $H^j(x) = \sum_{i=1}^{T} w_\mathrm{i}^\mathrm{j} h_i^j(x)$ is also unbiased, and we can obtain a variance-minimized unbiased estimation $H^j(x)$ for $\mathrm{P}(c_\mathrm{j} \mid x)$ by setting the weights. Minimizing the variance of the combined approximation error $\sum_{i=1}^{T} w_\mathrm{i}^\mathrm{j} \epsilon_i^j(x)$ under the constraints $w_\mathrm{i}^\mathrm{j} \geq 0$ and $\sum_{i=1}^{T} w_\mathrm{i}^\mathrm{j} = 1$, we can get the optimization problem:

$$w^\mathrm{j} = \arg\min_{w^\mathrm{j}} \sum_{\mathrm{k}=1}^{\mathrm{m}} \left(\sum_{i=1}^{T} w_\mathrm{i}^\mathrm{j} h_i^j(x_k) - \mathbb{I}(f(x_\mathrm{k} = c_\mathrm{j}))\right)^2, \mathrm{j} = 1,\ldots,\ell,$$

from which the weights can be solved. Notice that soft voting is generally used for *homogeneous ensembles*. For *homogeneous ensembles,* the class probabilities generated by different types of learners usually cannot be compared directly without a careful calibration. In such situations, the class probability outputs are often converted to class label outputs by setting $h_i^j(x)$ to 1 if $h_i^j(x) = \max_j\{h_i^j(x)\}$ and 0 otherwise, and then the voting methods for crisp labels can be applied.

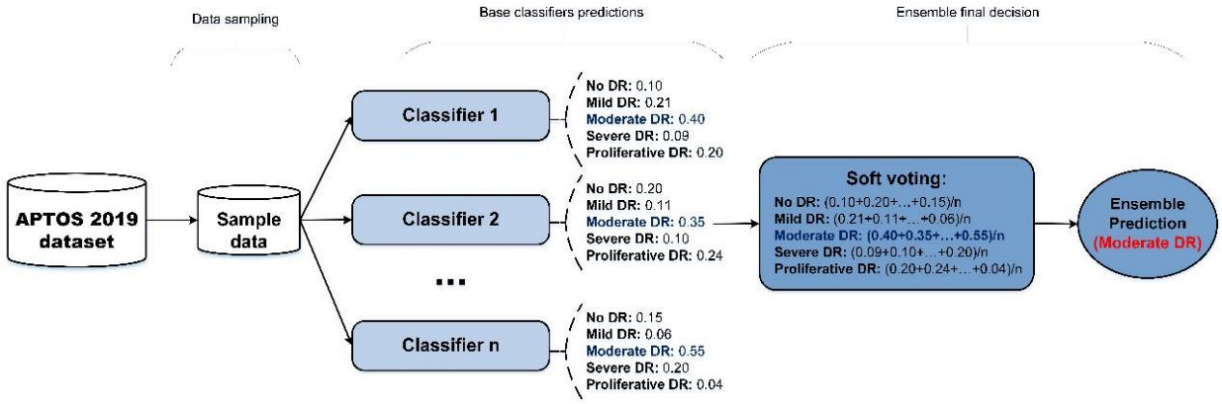

Figure 3: the pipeline of the soft voting ensemble for diabetic retinopathy classification. Preprocessed fundus images are passed through CNN and Vision Transformer backbones, whose features are fused in hybrid models and classified into DR severity stages. The probability outputs from multiple hybrids are then combined using soft voting (averaging across models), and the class with the highest aggregated probability is selected as the final prediction.

Step 6: Classification Output

The decision variable $\hat{y}$ indicates the predicted diabetic retinopathy stage, where $\hat{y} = 0$ corresponds to no DR and $\hat{y} = 4$ corresponds to proliferative DR. By integrating rigorous preprocessing, transfer learning, hybrid fusion, and ensemble voting, our pipeline achieves robust and clinically reliable classification performance.

Limitations and future research directions

In this study, a thorough evaluation across 23 models ensures the selection of the most robust architectures for DR detection, improving both accuracy and reliability in real-world applications. By combining the local feature extraction of CNNs with the global context modeling of Transformers, the hybrid architecture provides a richer representation of retinal images, leading to more precise DR stage classification. Ensemble methods aggregate the strengths of multiple high-performing models, significantly boosting classification performance and reducing prediction variance for DR diagnosis. Repeated training and cross-validation confirm model stability and reproducibility, ensuring reliable DR predictions across different data splits and training conditions. Tailored fine-tuning of pre-trained models allows the network to better adapt to the specific characteristics of retinal images, leading to noticeable improvements in classification metrics. The framework's high accuracy, low variance, and multi-class classification capability align with clinical diagnostic needs, supporting automated screening and early intervention for DR. Comparative analysis with state-of-the-art models highlights the superiority of the proposed approach, validating its effectiveness in addressing known limitations of existing DR detection methods. Although our study has significantly improved DR diagnosis and enhanced the literature of deep learning by introducing novel architecture, the study still has some limitations, which inspire and guide us for future research. By identifying limitations such as interpretability, generalizability, and computational overhead, the study lays a clear path for future improvements in developing practical, scalable DR screening tools. From a deep learning perspective, the computational burden remains a common concern; the accuracy of hybrid and ensemble architectures comes at a price, of large training and inference resources, which may become prohibitive for deployment in a clinical setting in real time. In addition, the work is dataset-specific

(APTOS 2019), and the generalization capacity of the proposed models is low. This has some implications regarding the generalizability of the studies across different populations, devices, and clinical settings.

The lack of external validation underscores the requirement to test in more diverse geographic and demographic datasets to ensure generalizability. Furthermore, the models have not been tailored for different cohorts or patient populations, which can raise concerns of fairness and clinical robustness in minority populations. Another important discussion point is interpretability: despite our models showing high accuracy, they are a black box with little explainability. No XAI method, like CAM, Grad-CAM, or SHAP, is available; thus, uptake is complex in attention-driven sectors such as clinical, where transparency and trust are critical. In addition, we did not perform exhaustive hyperparameter tuning, which would have contributed to better performance and robustness of our models. Further research should be directed towards minimizing computational costs, including interpretability frameworks, extensive hyperparameter tuning, validating performances on a wide range of datasets and clinical environments to support real-world DR screening and diagnosis.

One of the limitations of this study is the computational cost; the proposed hybrid and ensemble architecture, while highly accurate, is computationally intensive. Training and inference require substantial hardware resources, which may limit their deployment in real time. Generalizability: The models were trained and validated only on one dataset, which limits the models' generalizability to broader populations. Interpretability Challenges: Although the ensemble and hybrid models achieved superior accuracy, they act as black boxes with limited interpretability. Such limitations may hinder clinical adoption where model transparency is critical for trust and validation.

Future work could focus on reducing the computational overhead through model compression techniques such as pruning, quantization, or using knowledge distillation to develop lightweight yet accurate student models. Expanding validation to include other datasets could help evaluate the transferability and reliability of the models in heterogeneous clinical settings. Also, it could explore integrating interpretability tools like CAM, GradCam, and SHAP to provide visual explanations for the model's decisions. Finally, conducting deployment trials in a hospital or telemedicine setting would assess the models' real-time applicability.

Conclusion

This study explored different deep learning models for classifying diabetic retinopathy, focusing on improving accuracy and reliability. We tested a wide range of pre-trained CNNs and Transformers, fine-tuned the best ones, and created hybrid models that could take advantage of both local and global features in the retinal images. By combining the best models using soft voting ensembles, we got strong results, with our top model reaching over 93% accuracy. These findings suggest that using a mix of architectures and combining their outputs is a solid strategy for building better medical image classifiers. While our approach performed better than many existing models, there's room to improve things like model explainability and running time. In the future, we hope to test these models on other datasets and work on making the models faster and easier to use in real-world clinics. Overall, this work shows that innovative combinations of deep learning techniques can help with early and accurate DR detection.

Conflicts of Interest

The authors declare no conflict of interest. The sponsors had no role in design, execution, interpretation, or writing of the manuscript.

Data Availability

The dataset analyzed during the current study is available in the APTOS 2019 repository from Kaggle website. Link for the dataset is provided here:https://www.kaggle.com/competitions/aptos2019-blindness-detection/data.

Research funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (RS-2025-00554526, 2021R1A2C1011198), the Institute for Information & Communication Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) under the ICT Creative Consilience Program

(IITP-2021-2020-0-01821), and the AI Platform to Fully Adapt and Reflect Privacy-Policy Changes (RS-2022-II220688).

Author Contribution Declaration

Methodology, S.E.S., and Y.M.; conceptualization, T.A. and S.E.S; formal analysis, S.E.S., T.A., and Y.M.; validation, S.E.S. and T.A.; visualization, Y.M., and T.A.; investigation, S.E.S., and Y.M.; data curation, T.A.; writing—original draft preparation, S.E.S., Y.M, and T.A.; writing—review and editing, Y.M. and S.E.S.; supervision, T.A. and S.E.S.

References

1. Ting, D. S. W. *et al.* Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* 318, 2211 (2017).
2. Teo, Z. L. *et al.* Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045. *Ophthalmology* 128, 1580–1591 (2021).
3. Yau, J. W. Y. *et al.* Global Prevalence and Major Risk Factors of Diabetic Retinopathy. *Diabetes Care* 35, 556–564 (2012).
4. Sun, H. *et al.* IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract* 183, 109119 (2022).
5. Bourne, R. R. A. *et al.* Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *Lancet Glob Health* 5, e888–e897 (2017).
6. Fong, D. S. *et al.* Retinopathy in diabetes. *Diabetes Care* 27 Suppl 1, (2004).
7. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316, 2402–2410 (2016).
8. Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* 1, (2018).
9. Gargeya, R. & Leng, T. Automated Identification of Diabetic Retinopathy Using Deep Learning. *Ophthalmology* 124, 962–969 (2017).
10. Li, T. *et al.* Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf Sci (N Y)* 501, 511–522 (2019).
11. Quellec, G., Charrière, K., Boudi, Y., Cochener, B. & Lamard, M. Deep image mining for diabetic retinopathy screening. *Med Image Anal* 39, 178–193 (2017).
12. Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P. & Zheng, Y. Convolutional Neural Networks for Diabetic Retinopathy. *Procedia Comput Sci* 90, 200–205 (2016).
13. Karthik, Maggie & Dane, S. APTOS 2019 Blindness Detection. Preprint at (2019).
14. Decencière, E. *et al.* Feedback on a publicly distributed image database: the Messidor database. *Image Analysis & Stereology* 231–234 (2014).
15. Dugas, E., Jared, Jorge & Cukierski, W. Diabetic Retinopathy Detection. Preprint at (2015).
16. Kauppi, T. *et al.* the DIARETDB1 diabetic retinopathy database and evaluation protocol.
17. Yu, X. *et al.* Transfer learning for medical images analyses: A survey. *Neurocomputing* 489, 230–254 (2022).
18. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016-December, 770–778 (2015).
19. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-January, 2261–2269 (2016).

20. Tan, M. & Le, Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *36th International Conference on Machine Learning, ICML 2019* 2019-June, 10691–10700 (2019).
21. Arcadu, F. *et al.* Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digit Med* 2, (2019).
22. Condon, J. J. J. *et al.* Impact of Transfer Learning Using Local Data on Performance of a Deep Learning Model for Screening Mammography. *Radiol Artif Intell* 6, (2024).
23. Fagbola, T. M. & Success, I. Leveraging pretrained models for multimodal medical image interpretation: An exhaustive experimental analysis. *medRxiv* 2024–2028 (2024).
24. Ashraf, S. M. N., Mamun, Md. A., Abdullah, H. Md. & Alam, Md. G. R. SynthEnsemble: A Fusion of CNN, Vision Transformer, and Hybrid Models for Multi-Label Chest X-Ray Classification. *2023 26th International Conference on Computer and Information Technology, ICCIT 2023* https://doi.org/10.1109/ICCIT60459.2023.10441433 (2024) doi:10.1109/ICCIT60459.2023.10441433.
25. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *Adv Neural Inf Process Syst* 2020-December, (2020).
26. Esteva, A. *et al.* Deep learning-enabled medical computer vision. *NPJ Digit Med* 4, (2021).
27. Irvin, J. *et al.* CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019* 590–597 (2019) doi:10.1609/aaai.v33i01.3301590.
28. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Adv Neural Inf Process Syst* 2017-December, 6403–6414 (2016).
29. Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Med Image Anal* 81, 102559 (2022).
30. Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 1, e271–e297 (2019).
31. (PDF) nnFormer: Interleaved Transformer for Volumetric Segmentation. https://www.researchgate.net/publication/354435522_nnFormer_Interleaved_Transformer_for_Volumetric_Segmentation.
32. Hinton, G., Vinyals, O. & Dean, J. Distilling the Knowledge in a Neural Network. https://arxiv.org/pdf/1503.02531 (2015).
33. Ovadia, Y. *et al.* Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *Adv Neural Inf Process Syst* 32, (2019).
34. Zoph, B. & Le, Q. V. Neural Architecture Search with Reinforcement Learning. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* https://arxiv.org/pdf/1611.01578 (2016).
35. Tomczak, J. M. *et al.* Attention-based Deep Multiple Instance Learning. *Medical Imaging with Deep Learning* 3–5 (2018).
36. Rieke, N. *et al.* The future of digital health with federated learning. *NPJ Digit Med* 3, (2020).
37. Shenavarmasouleh, F. & Arabnia, H. R. DRDr: Automatic Masking of Exudates and Microaneurysms Caused by Diabetic Retinopathy Using Mask R-CNN and Transfer Learning. 307–318 (2021) doi:10.1007/978-3-030-71051-4_24.
38. Mikram, M., Moujahdi, C., Rhanoui, M., Meddad, M. & Khallout, A. Hybrid Deep Learning Models for Diabetic Retinopathy Classification. in 167–178 (2022). doi:10.1007/978-3-031-07969-6_13.

39. Ramachandran, A., Patel, K., Sharma, R. & Gupta, S. Hybrid deep learning approaches using Siamese networks and SVM for diabetic retinopathy detection. *Biomed Signal Process Control* 71, (2022).
40. Martínez, R., Delgado, M., Torres, J. & Fernandez, C. Hybrid deep learning models combining CNNs and RNNs for diabetic macular edema screening. *Scientific Reports* 14, 987–1002 (2024).
41. Kumar, G., Chatterjee, S. & Chattopadhyay, C. DRISTI: a hybrid deep neural network for diabetic retinopathy diagnosis. *Signal Image Video Process* 15, 1679–1686 (2021).
42. R., Y., Raja Sarobin M., V., Panjanathan, R., S., G. J. & L., J. A. Diabetic Retinopathy Classification Using CNN and Hybrid Deep Convolutional Neural Networks. *Symmetry (Basel)* 14, 1932 (2022).
43. Chaudhary, S. & Ramya, H. R. Detection of Diabetic Retinopathy using Machine Learning Algorithm. *2020 IEEE International Conference for Innovation in Technology, INOCON 2020* https://doi.org/10.1109/INOCON50539.2020.9298413 (2020) doi:10.1109/INOCON50539.2020.9298413.
44. Menaouer, B., Dermane, Z., El Houda Kebir, N. & Matta, N. Diabetic Retinopathy Classification Using Hybrid Deep Learning Approach. *SN Computer Science 2022 3:5* 3, 357- (2022).
45. Rahman, M., Islam, M., Ahmed, K. & Hossain, M. Adaptive hybrid focal-entropy loss for enhancing diabetic retinopathy detection. *IEEE Trans Neural Netw Learn Syst* 1023–1036 (2023).
46. Fernandez, J., Liu, Y., Zhang, T. & Chen, H. Hybrid deep learning Gaussian process for diabetic retinopathy diagnosis and uncertainty quantification. *International Conference on AI in Healthcare* 321–335 (2024).
47. Bodapati, J. D. *et al.* Blended Multi-Modal Deep ConvNet Features for Diabetic Retinopathy Severity Prediction. *Electronics 2020, Vol. 9, Page 914* 9, 914 (2020).
48. (PDF) DRDr II: Detecting the Severity Level of Diabetic Retinopathy Using Mask RCNN and Transfer Learning. https://www.researchgate.net/publication/352702976_DRDr_II_Detecting_the_Severity_Level_of_Diabetic_Retinopathy_Using_Mask_RCNN_and_Transfer_Learning.
49. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* https://arxiv.org/pdf/1409.1556 (2014).
50. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021 - 9th International Conference on Learning Representations* https://arxiv.org/pdf/2010.11929 (2020).
51. Liu, Z. *et al.* Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. https://github.
52. Howard, Jeremy. & Gugger, Sylvain. Deep Learning for Coders with fastai and PyTorch. https://books.google.com/books/about/Deep_Learning_for_Coders_with_fastai_and.html?id=yATuDwAAQBAJ (2020).
53. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. http//roboticsStanfordedu/"ronnyk.
54. Cawley, G. C. & Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* 11, 2079–2107 (2010).
55. Ensemble Methods in Machine Learning | Proceedings of the First International Workshop on Multiple Classifier Systems. https://dl.acm.org/doi/10.5555/648054.743935.
56. Zhou, Z.-H. *Ensemble Methods*. (Chapman and Hall/CRC, 2012). doi:10.1201/b12207.
57. Xu, H., Shao, X., Fang, D. & Huang, F. A hybrid neural network approach for classifying diabetic retinopathy subtypes. *Front Med (Lausanne)* 10, 1293019 (2023).

58. Mikram, M., Moujahdi, C., Rhanoui, M., Meddad, M. & Khallout, A. Hybrid Deep Learning Models for Diabetic Retinopathy Classification. *Lecture Notes in Networks and Systems* 489 LNNS, 167–178 (2022).
59. 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 2025 15:1* 15, 15166- (2025).
60. 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, (2024).
61. Azzou, S., Boukredera, D. & Baouz, S. A Hybrid CNN-Transformer Approach for Precise Three-Class Diabetic Retinopathy Classification. *Jordanian Journal of Computers and Information Technology* 11, 279 (2025).
62. Zhang, W., Belcheva, V. & Ermakova, T. Interpretable Deep Learning for Diabetic Retinopathy: A Comparative Study of CNN, ViT, and Hybrid Architectures. *Computers* 14, 187 (2025).
63. Singh, D., Agarwal, S. & Mishra, S. Retinal Fundus Multi-Disease Image Classification using Hybrid CNN-Transformer-Ensemble Architectures. *Lecture Notes in Networks and Systems* 1113, 103–120 (2025).