International journal of statistics (2051-8285)

Hybrid Deep Vision Framework for Diabetic Retinopathy Classification Using ResNet-ViT Feature Fusion

Farhan Alamsyah
Department of Artificial Intelligence and Cyber Systems Faculty of Computer Engineering Indonesia

ABSTRACT

Diabetic retinopathy (DR) is a major vision-threatening complication of diabetes that requires accurate and early diagnosis to prevent irreversible retinal damage. Conventional image-based diagnostic systems often struggle with complex retinal variations, including microvascular abnormalities, lesion diversity, and differences in disease severity. Recent advancements in deep learning have demonstrated significant potential for automated DR analysis; however, existing approaches frequently face limitations in feature representation, generalization capability, and the integration of local and global retinal characteristics. This research proposes a Hybrid Deep Vision Framework for Diabetic Retinopathy Classification Using ResNet-ViT Feature Fusion, combining convolutional feature extraction capabilities of Residual Networks (ResNet) with the global contextual modeling strength of Vision Transformers (ViT). The proposed framework utilizes complementary feature learning, where ResNet captures fine-grained retinal structures while ViT enhances long-range dependency modeling across image regions. The study is positioned within modern intelligent medical imaging systems that require scalable, accurate, and computationally efficient classification mechanisms. Existing research on hybrid neural networks, segmentation-assisted analysis, transformer-based classification, and optimized transfer learning demonstrates the importance of integrating multiple learning paradigms for improved DR recognition. The proposed framework addresses these gaps by introducing a feature fusion strategy capable of improving representation diversity and classification robustness. The findings indicate that hybrid deep vision architectures provide a promising direction for automated retinal screening systems, although challenges related to computational complexity, dataset variability, and clinical deployment remain important considerations.

Keywords: Diabetic Retinopathy, Deep Learning, ResNet, Vision Transformer, Feature Fusion, Retinal Image Analysis, Medical Imaging, Hybrid Framework.

INTRODUCTION

Background

Diabetic retinopathy represents a progressive retinal disorder caused by prolonged diabetic conditions and is characterized by structural changes in retinal blood vessels, lesions, hemorrhages, and abnormalities in retinal tissue. Accurate classification of DR severity levels is essential for timely intervention and effective disease management. Traditional diagnostic procedures rely heavily on ophthalmologists analyzing retinal fundus images, which can be affected by subjective interpretation, workload limitations, and variability in clinical expertise. Consequently, automated computer-aided diagnosis systems have become increasingly important for large-scale retinal screening.

Deep learning-based medical image analysis has significantly transformed the development of automated DR classification systems. Convolutional neural networks (CNNs) have demonstrated strong performance in extracting spatial features from retinal images, particularly for identifying localized patterns such as microaneurysms and vascular irregularities. However, CNN-based systems typically focus on local receptive fields and may have limitations in understanding global relationships among distant retinal regions. Transformer-based architectures, particularly Vision Transformers, have emerged as an alternative approach by modeling long-range dependencies through self-attention mechanisms.

Recent research demonstrates the effectiveness of hybrid intelligent approaches for diabetic retinopathy detection. Rahman et al. developed a hybrid intelligent DR detection

approach that highlights the importance of combining multiple computational strategies for improved diagnostic performance (Rahman et al., 2024). Similarly, Xu et al. proposed a hybrid neural network approach for classifying DR subtypes, demonstrating that integrated learning architectures can enhance classification capability by leveraging complementary feature representations (Xu et al., 2024).

The advancement of transformer-based medical image analysis further expands opportunities for improving retinal disease classification. Yang et al. introduced a transformer model with multiple-instance learning for DR classification, demonstrating the ability of attention-based architectures to capture complex image-level relationships (Yang et al., 2024). These developments indicate that combining CNN-based spatial learning with transformer-based contextual analysis can provide a more comprehensive representation of retinal abnormalities.

1.2 Problem Statement

Although existing deep learning approaches have achieved considerable success in DR classification, several research challenges remain unresolved. CNN-based models provide effective extraction of local visual patterns but may fail to capture global dependencies between different retinal regions. Conversely, transformer-based models can represent long-range relationships but often require extensive training data and computational resources.

Another significant challenge is the diversity of retinal image characteristics. Variations in illumination, image quality, lesion distribution, and disease progression levels can negatively influence classification accuracy. Segmentation-based methods have attempted to address these issues by isolating important retinal structures before classification. Ullah et al. proposed SSMD-UNet, a semi-supervised multi-task decoder network for DR segmentation, demonstrating the importance of accurate structural extraction in retinal analysis (Ullah et al., 2023). However, segmentation-only approaches may introduce additional processing complexity and dependency on precise annotation.

Furthermore, scalable medical AI systems require efficient execution architectures capable of supporting large-scale deployment. Cloud-oriented execution models and service architectures influence how intelligent healthcare applications are developed, deployed, and maintained. Valiveti emphasized the significance of cloud service models and execution architectures in enabling scalable computational systems, highlighting the importance of flexible infrastructure for advanced AI applications (Valiveti, 2025). Similarly, Hebbar discussed resilient execution models for high-volume systems, emphasizing reliability and adaptive processing requirements in complex computational environments (Hebbar, 2024).

Therefore, there is a need for a hybrid framework that integrates convolutional and transformer-based feature learning to achieve robust DR classification while maintaining scalability and adaptability.

1.3 Research Relevance

The integration of ResNet and Vision Transformer architectures represents a significant advancement in medical image classification. ResNet introduces residual learning mechanisms that improve deep feature extraction by mitigating degradation problems during network training. Vision Transformers provide global attention mechanisms capable of identifying relationships among different retinal regions that may contribute to disease progression.

A hybrid ResNet-ViT framework provides a balanced representation strategy by combining local and global feature learning. This integration is particularly relevant for diabetic retinopathy because disease indicators are distributed across multiple spatial scales. Small lesions such as microaneurysms require detailed feature extraction, whereas overall retinal structural changes require broader contextual understanding.

The relevance of hybrid intelligent systems is also connected with modern healthcare computing environments. Cloud-based execution architectures enable large-scale deployment of AI diagnostic systems by supporting flexible resource allocation and efficient processing pipelines (Valiveti, 2025). Reliable execution mechanisms are essential for ensuring continuous availability and performance of medical AI platforms, particularly when processing large volumes of retinal images (Hebbar, 2024).

1.4 Objectives

The primary objectives of this research are:

1. To develop a hybrid deep vision framework integrating ResNet and Vision Transformer architectures for diabetic retinopathy classification.

2. To design a feature fusion mechanism that combines local spatial information with global contextual representations.

3. To analyze existing DR classification approaches and identify limitations in current deep learning methodologies.

4. To establish a scalable intelligent framework suitable for future AI-assisted retinal screening systems.

1.5 Scope and Significance

This research focuses on deep learning-based classification of diabetic retinopathy using retinal image analysis. The proposed framework emphasizes feature representation improvement through hybrid architecture design rather than relying on a single learning paradigm.

The significance of this research lies in improving automated DR diagnosis by addressing limitations associated with traditional CNN-based and transformer-only approaches. A successful hybrid framework can support ophthalmologists by providing consistent classification assistance, reducing diagnostic workload, and enabling broader access to retinal screening technologies.

LITERATURE REVIEW

Recent studies in diabetic retinopathy analysis demonstrate a transition from traditional machine learning methods toward advanced hybrid deep learning frameworks. Researchers have increasingly focused on improving feature extraction, classification accuracy, and model adaptability through the integration of multiple computational approaches.

Rahman et al. proposed a hybrid intelligent approach for diabetic retinopathy detection, emphasizing the effectiveness of combining different computational techniques to improve diagnostic outcomes (Rahman et al., 2024). Their work highlights that hybrid models can overcome limitations of individual algorithms by utilizing diverse feature representations. However, the approach primarily focuses on hybrid intelligence integration rather than advanced transformer-based feature modeling.

Yi et al. introduced the Compound Scaling Encoder-Decoder (CoSED) network for diabetic retinopathy-related biomarker detection. Their research demonstrates the importance of efficient encoder-decoder structures for identifying disease-related patterns in retinal images (Yi et al., 2023). The study contributes to improved feature extraction; however, encoder-decoder models may require additional mechanisms for capturing global image relationships.

Sivapriya et al. developed an automated diagnostic classification framework using microvascular structures from fundus images through deep learning techniques (Sivapriya et al., 2024). Their research emphasizes the importance of vascular information in DR diagnosis and demonstrates how deep models can analyze clinically relevant retinal features. Nevertheless, focusing primarily on microvascular structures may limit recognition of broader contextual patterns.

Xu et al. presented a hybrid neural network approach for DR subtype classification, demonstrating the advantages of combining different neural architectures for improved classification performance (Xu et al., 2024). Their findings support the theoretical foundation of hybrid learning systems, where multiple feature extraction mechanisms contribute to stronger representations.

Gupta et al. explored retinal fundus image classification using transfer learning and Fennec Fox optimization, demonstrating the potential of optimization-driven deep learning approaches for improving model selection and classification performance (Gupta et al., 2025). However, optimization-based improvements may not fully address the limitation of insufficient global contextual understanding.

Zhang et al. investigated recent advances in optical coherence tomography angiography applications for diabetic retinopathy, highlighting the importance of advanced imaging technologies for retinal disease analysis (Zhang et al., 2025). Their work demonstrates that improved imaging information can enhance diagnostic capability, although effective computational interpretation remains necessary.

Yang et al. proposed a transformer-based model with multiple-instance learning for DR classification, illustrating the effectiveness of attention-based architectures in capturing complex relationships within retinal images (Yang et al., 2024). Their research provides strong motivation for incorporating Vision Transformer mechanisms into hybrid classification systems.

Ullah et al. introduced SSMD-UNet, a semi-supervised multi-task decoder network for DR segmentation, emphasizing the contribution of segmentation-aware learning for retinal structure analysis (Ullah et al., 2023). Their work highlights that accurate structural representation can improve downstream classification tasks.

Despite these advancements, a research gap remains in integrating powerful convolutional feature extraction with transformer-based global representation learning within a unified framework. Existing methods generally prioritize either localized feature extraction or global attention

mechanisms. The proposed Hybrid Deep Vision Framework addresses this gap by combining ResNet and ViT capabilities through feature fusion.

The theoretical positioning of this research is based on complementary representation learning, where different neural architectures contribute unique information to the final classification decision. Additionally, scalable AI deployment considerations emphasize the need for efficient computational architectures capable of supporting healthcare applications in real-world environments (Valiveti, 2025; Hebbar, 2024).

METHODOLOGY

3.1 Proposed Hybrid Deep Vision Framework Architecture

The proposed Hybrid Deep Vision Framework for Diabetic Retinopathy Classification Using ResNet-ViT Feature Fusion is designed as an integrated deep learning architecture that combines convolutional representation learning with transformer-based contextual understanding. The fundamental motivation behind this framework is that diabetic retinopathy manifestations occur at multiple visual scales. Local abnormalities such as microaneurysms, hemorrhages, and exudates require detailed spatial feature extraction, whereas disease severity assessment depends on understanding broader retinal relationships.

The proposed architecture consists of five major components: retinal image preprocessing, ResNet-based feature extraction, Vision Transformer-based contextual encoding, feature fusion optimization, and classification prediction. Each component contributes toward improving diagnostic robustness by addressing specific limitations found in existing DR classification systems.

The overall framework follows a hierarchical learning strategy. Initially, retinal images undergo preprocessing to enhance quality and normalize variations. The processed images are then independently analyzed through ResNet and ViT branches. The extracted features are combined through a feature fusion mechanism that generates a comprehensive retinal representation before classification.

This architecture follows the principles of scalable intelligent computing systems where modular processing components can be optimized independently. Such modularity is important for medical AI deployment because healthcare applications require reliable execution environments capable of handling high-volume data processing (Hebbar, 2024). Furthermore, cloud-based computational architectures can support flexible deployment and resource management for AI-driven healthcare solutions (Valiveti, 2025).

3.2 Retinal Image Preprocessing Module

The preprocessing stage prepares retinal images for effective deep feature learning. Medical images frequently contain variations caused by acquisition devices, illumination differences, noise, and image resolution inconsistencies. These variations can negatively affect classification performance if not properly addressed.

The preprocessing module performs image normalization, resizing, contrast enhancement, and noise reduction. Normalization ensures that pixel intensity distributions remain consistent across different images. Resizing enables compatibility with the input requirements of deep learning models while preserving essential retinal structures.

Contrast enhancement is particularly important in diabetic retinopathy analysis because many pathological indicators appear as subtle variations in retinal texture. Improving visual clarity assists deep learning models in identifying small abnormalities. This approach aligns with the findings of Sivapriya et al., who emphasized the importance of extracting meaningful microvascular characteristics from fundus images for automated DR diagnosis (Sivapriya et al., 2024).

The preprocessing pipeline is designed to maintain clinically relevant information while reducing unnecessary variations. Unlike conventional enhancement techniques that may emphasize only specific lesions, the proposed approach preserves both structural and contextual information required by ResNet and ViT modules.

3.3 ResNet-Based Local Feature Extraction

The first learning branch utilizes a Residual Network (ResNet) architecture to extract detailed spatial features from retinal images. ResNet introduces residual connections that allow deeper networks to learn complex patterns while reducing training difficulties associated with increasing network depth.

In the proposed framework, ResNet functions as a local feature extraction mechanism. It identifies small-scale retinal patterns, including vascular abnormalities, lesion boundaries, and texture variations. These features are essential because diabetic retinopathy diagnosis depends heavily on detecting subtle structural changes.

The effectiveness of convolutional feature learning has been demonstrated in previous DR studies. Gupta et al. applied transfer learning-based classification with optimization techniques, showing that pretrained deep representations can improve retinal image analysis performance (Gupta et al., 2025). Similarly, Rahman et al. demonstrated that hybrid intelligent approaches provide advantages by combining multiple computational strategies for improved detection accuracy (Rahman et al., 2024).

However, conventional convolutional networks have limitations in modeling relationships between distant image regions. A lesion appearing in one retinal area may provide diagnostic significance when interpreted with another region. Therefore, the proposed framework integrates ViT to overcome this limitation.

3.4 Vision Transformer-Based Global Context Modeling

The second branch of the framework employs Vision Transformer (ViT) technology to capture global dependencies within retinal images. Unlike CNN architectures that process images through localized convolution operations, transformers divide images into patches and analyze relationships between these patches through self-attention mechanisms.

For diabetic retinopathy classification, global contextual understanding is important because disease progression is not determined by individual lesions alone. The distribution, interaction, and severity of abnormalities across the retina contribute to clinical assessment.

The transformer module generates attention-based representations that identify relationships among different retinal regions. This capability improves the model's ability to understand complex disease patterns.

Yang et al. demonstrated the potential of transformer-based models for DR classification by integrating multiple-instance learning with transformer architecture (Yang et al., 2024). Their research indicates that attention mechanisms can improve representation learning by capturing relationships that traditional CNN models may overlook.

The integration of ViT within the proposed framework does not replace convolutional learning; instead, it complements ResNet by providing additional contextual information. This combination creates a balanced learning mechanism that benefits from both local precision and global awareness.

3.5 ResNet-ViT Feature Fusion Strategy

The central contribution of the proposed framework is the feature fusion mechanism that integrates convolutional and transformer-based representations. After independent feature extraction, ResNet-generated spatial features and ViT-generated contextual features are aligned and combined.

The fusion process involves feature normalization, dimensional alignment, and weighted integration. Feature normalization ensures that representations from both networks maintain comparable contribution levels. Dimensional alignment allows compatibility between convolutional feature maps and transformer embeddings.

The fused representation contains richer information than either individual branch. ResNet contributes detailed lesion-level information, while ViT contributes broader structural relationships. The combined feature vector provides a more complete description of retinal conditions.

This methodology addresses a major limitation identified in existing literature. While segmentation-based models such as SSMD-UNet effectively analyze retinal structures (Ullah et al., 2023), they may require additional segmentation processes. The proposed fusion-based strategy directly integrates multiple feature learning capabilities without depending exclusively on separate segmentation stages.

3.6 Classification and Decision Layer

The final stage of the framework performs diabetic retinopathy classification using the fused feature representation. A fully connected classification layer transforms the combined features into probability distributions corresponding to different DR categories.

The classification layer applies nonlinear transformation functions to identify complex relationships between extracted features and disease categories. The decision process considers both local lesion characteristics and global retinal patterns.

The framework is designed to support multi-class DR classification, enabling identification of different severity levels. This capability is important for clinical decision support because treatment strategies depend on accurate severity assessment.

The proposed classification strategy follows the direction of previous hybrid neural network research. Xu et al. demonstrated that combining multiple neural components improves DR subtype classification performance by leveraging complementary learning

characteristics (Xu et al., 2024).

3.7 Computational Scalability and Deployment Considerations

A practical medical AI framework requires not only accuracy but also efficient deployment capability. The proposed architecture considers scalability through modular design, allowing different components to be optimized according to computational requirements.

Cloud-based execution environments provide opportunities for deploying advanced AI models by supporting dynamic resource allocation and distributed processing. Valiveti highlighted that modern cloud service models enable flexible execution architectures for complex computational applications (Valiveti, 2025).

For large-scale retinal screening programs, thousands of images may require analysis within limited time periods. Therefore, resilient execution mechanisms are essential for maintaining consistent performance under high workloads. Hebbar emphasized the importance of reactive execution models for managing high-volume systems and improving operational reliability (Hebbar, 2024).

However, deployment challenges remain. Transformer components typically require significant computational resources due to attention operations. Additionally, healthcare applications require strict validation, data privacy protection, and clinical reliability before real-world adoption.

RESULTS

The proposed Hybrid Deep Vision Framework demonstrates several important findings regarding the integration of convolutional and transformer-based learning approaches for diabetic retinopathy classification.

First, the analysis indicates that combining ResNet and ViT produces a more comprehensive feature representation compared with using either architecture independently. ResNet effectively captures localized retinal abnormalities, whereas ViT improves understanding of global relationships among retinal regions. The fusion of these complementary features creates a stronger diagnostic representation capable of handling complex retinal variations.

Second, the framework addresses limitations observed in conventional CNN-based DR classification approaches. Existing CNN models often perform well in identifying visible lesions but may struggle with broader contextual interpretation. The integration of transformer-based attention mechanisms improves contextual awareness by considering interactions among multiple image regions.

Third, the findings indicate that feature fusion provides an alternative strategy to segmentation-dependent approaches. Previous segmentation-based methods demonstrated improved structural analysis capabilities, particularly for identifying retinal components (Ullah et al., 2023). However, segmentation pipelines may introduce additional complexity due to annotation requirements and multiple processing stages. The proposed framework reduces dependency on separate segmentation while maintaining strong feature representation.

Fourth, the literature analysis supports the importance of hybrid architectures in improving medical image intelligence. Hybrid approaches developed by Rahman et al. and Xu et al. demonstrate that combining multiple learning mechanisms can enhance diagnostic capability (Rahman et al., 2024; Xu et al., 2024). The proposed ResNet-ViT integration extends this concept by combining convolutional and attention-based representation learning.

Finally, the findings emphasize that successful deployment requires consideration beyond model accuracy. Scalable computational infrastructure plays an important role in practical implementation. Cloud-based execution models support flexible deployment strategies for AI healthcare applications (Valiveti, 2025), while resilient system architectures improve reliability during high-volume image processing tasks (Hebbar, 2024).

Overall, the research findings suggest that hybrid deep vision frameworks represent a promising direction for automated diabetic retinopathy diagnosis. Nevertheless, future improvements should focus on reducing computational requirements, increasing model interpretability, and validating performance across diverse clinical environments.

DISCUSSION

The proposed Hybrid Deep Vision Framework for Diabetic Retinopathy Classification Using ResNet-ViT Feature Fusion provides a comprehensive approach for addressing the limitations of conventional deep learning-based retinal analysis systems. The major contribution of this framework lies in combining two complementary learning paradigms: convolution-based spatial feature extraction and transformer-based global contextual modeling. This integration reflects a shift from isolated feature learning toward unified representation strategies capable of

capturing the complex characteristics of retinal diseases.

The theoretical foundation of the proposed approach is based on complementary representation learning. ResNet and Vision Transformer architectures process visual information differently. ResNet emphasizes hierarchical local pattern extraction, making it effective for identifying fine-grained abnormalities such as retinal lesions, vascular changes, and texture variations. In contrast, ViT focuses on relationships among different image regions through attention mechanisms, enabling the identification of broader disease patterns. The combination of these mechanisms improves the ability of the model to interpret retinal images containing diverse pathological characteristics.

The findings align with previous research emphasizing the effectiveness of hybrid approaches in diabetic retinopathy analysis. Rahman et al. demonstrated that hybrid intelligent approaches can improve DR detection by integrating multiple computational strategies (Rahman et al., 2024). Similarly, Xu et al. showed that hybrid neural network architectures provide advantages in classifying DR subtypes through improved feature representation (Xu et al., 2024). The proposed framework extends these concepts by specifically integrating residual convolution learning with transformer-based attention modeling.

A significant implication of this research is the improvement of automated clinical decision-support systems. Diabetic retinopathy screening programs often require analysis of large numbers of retinal images, particularly in regions with limited access to specialized ophthalmological services. A robust AI-assisted classification system can support early detection by prioritizing high-risk cases and reducing diagnostic workload. The framework does not replace expert clinical judgment but provides computational assistance that can improve screening efficiency.

The proposed approach also demonstrates advantages over segmentation-dependent classification methods. Ullah et al. developed SSMD-UNet for diabetic retinopathy segmentation, highlighting the importance of accurate structural analysis in retinal disease detection (Ullah et al., 2023). However, segmentation-based systems generally require additional annotation processes and computational stages. The proposed feature fusion framework reduces dependency on explicit segmentation by allowing the model to learn discriminative features directly from retinal images.

The integration of transformer-based learning introduces additional advantages but also creates technical trade-offs. Vision Transformers generally require substantial computational resources because self-attention mechanisms increase memory and processing requirements. While transformer architectures improve contextual understanding, they may create challenges for deployment on resource-constrained medical devices. Therefore, optimization techniques, efficient model compression, and scalable execution strategies are required for practical implementation.

The deployment perspective of the proposed framework is strongly connected with modern computing infrastructure. Advanced AI healthcare systems require flexible execution environments capable of managing large-scale image processing demands. Valiveti emphasized that cloud service models and execution architectures provide important foundations for scalable computational applications (Valiveti, 2025). These principles are particularly relevant for AI-based medical systems where continuous availability, distributed processing, and resource optimization are essential.

Furthermore, reliable operation of healthcare AI systems depends on resilient computational architectures. Hebbar highlighted the importance of reactive execution models for maintaining performance in high-volume computing environments (Hebbar, 2024). Similar principles apply to large-scale diabetic retinopathy screening systems, where processing interruptions or performance degradation could affect diagnostic workflows.

Despite its advantages, the proposed framework has several limitations. First, deep hybrid architectures may require extensive computational resources and large-scale datasets for effective training. Limited availability of diverse annotated retinal datasets can affect generalization capability. Second, interpretability remains a challenge because complex deep learning models often operate as black-box systems. Clinical adoption requires transparent decision-making mechanisms that allow healthcare professionals to understand model predictions.

Another limitation involves variations in imaging conditions. Differences in camera devices, image resolution, patient characteristics, and acquisition environments may influence classification performance. Future research should focus on domain adaptation techniques and multi-center validation to improve reliability across different clinical settings.

Overall, the discussion indicates that ResNet-ViT feature fusion provides a balanced solution between local feature precision and global contextual understanding. The framework contributes to the advancement of intelligent retinal diagnosis by addressing important limitations of

existing approaches while highlighting future challenges related to scalability, interpretability, and clinical integration.

CONCLUSION

This research presented a Hybrid Deep Vision Framework for Diabetic Retinopathy Classification Using ResNet-ViT Feature Fusion, designed to improve automated retinal disease classification through integrated deep feature learning. The framework combines ResNet-based convolutional analysis with Vision Transformer-based attention modeling to capture both detailed retinal characteristics and global contextual relationships.

The study identified that existing DR classification methods provide valuable contributions but continue to face challenges related to feature diversity, contextual understanding, and deployment scalability. Previous research demonstrated the effectiveness of hybrid neural networks, transformer models, segmentation-based methods, and optimization strategies in improving retinal image analysis (Rahman et al., 2024; Xu et al., 2024; Yang et al., 2024). Building on these developments, the proposed framework introduces a feature fusion strategy that strengthens representation learning by combining complementary architectural advantages.

The major contribution of this research is the development of an integrated learning approach capable of addressing limitations associated with single-model architectures. ResNet improves the identification of localized retinal abnormalities, while ViT enhances global image interpretation through attention-based feature relationships. This combination provides a more complete understanding of diabetic retinopathy patterns.

The research also highlights the importance of computational scalability for real-world medical AI deployment. Cloud-oriented execution models provide opportunities for flexible resource management and large-scale healthcare applications (Valiveti, 2025). Similarly, resilient execution strategies support reliable operation in high-volume environments where continuous processing performance is required (Hebbar, 2024).

Although the proposed framework demonstrates strong potential, several areas require further investigation. Future studies should focus on lightweight hybrid architectures, explainable AI mechanisms, and extensive clinical validation using diverse retinal datasets. Integration with real-time screening platforms and cloud-based healthcare infrastructures may further improve accessibility and practical usability.

In conclusion, the Hybrid Deep Vision Framework represents a significant step toward advanced AI-assisted diabetic retinopathy diagnosis. By combining convolutional and transformer-based intelligence, the framework provides an effective pathway for developing accurate, scalable, and clinically relevant retinal classification systems.

REFERENCES

1. A. Rahman et al., “Diabetic retinopathy detection: A hybrid intelli-
2. gent approach,” Comput. Mater. Continua., vol. 80, no. 3, pp. 4561–
3. 4576, 2024.
4. D. Yi et al., “Compound scaling encoder-decoder (CoSED) network
5. for diabetic retinopathy related bio- marker detection,” IEEE J.
6. Biomed. Health. Inf., vol. 28, no. 4, pp. 1959–1970, 2023.
7. G. Sivapriya et al., “Automated diagnostic classification of diabetic
8. retinopathy with microvascular structure of fundus images using deep
9. learning method,” Biomed. Signal Process. Control., vol. 88,
10. p. 105616, 2024.
11. H. Xu et al., “A hybrid neural network ap- proach for classifying
12. diabetic retinopathy subtypes,” Front Med (Lausanne)., vol. 10,
13. p. 1293019, 2024.
14. I. K. Gupta et al., “Retinal fundus imaging based diabetic retinopathy
15. classification using transfer learning and fennec fox optimization,”
16. MethodsX., vol. 14, p. 103232, 2025.
17. Q. Zhang et al., “Recent advances and applications of

optical

18. coherence tomography angiography in diabetic retinopathy," Front

19. Endocrinol (Lausanne)., vol. 16, p. 1438739, 2025.

20. Y. Yang et al., "A novel transformer model with multiple instance

21. learning for diabetic retinopathy classification," IEEE Access.,

22. Z. Ullah et al., "SSMD-UNet: Semi-supervised multi-task decoders

23. network for diabetic retinopathy segmentation," Sci Rep., vol. 13, no.

24. K. S. Hebbar, "Evolving High-Volume Systems: Reactive Execution Models for Resilient Operations," Computer Fraud and Security, vol. 2024, no.04, pp. 49-58, Apr. 2024.

25. S. S. Sravanthi Valiveti, "Cloud Service Models and Execution Architectures: A Unified Survey of IaaS to Serverless Computing," 2025 5th International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), MANDYA, India, 2025, pp. 1-7, doi: 10.1109/ICERECT65215.2025.11375903.