International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4309-4312
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0630

Multi-Disease Detection in Retinal Fundus Images Using ViT Architecture

Jasmine R[1], Kalaiyarasi M[2], Kaviya N[3], Nandhini K[4]

[1]Assistant Professor, Department of Computer Science and Engineering, K Ramakrishnan College of Technology, Trichy, Tamil Nadu, India.

[2,3,4]UG, Department of Computer Science and Engineering, K Ramakrishnan College of Technology, Trichy, Tamil Nadu, India

Emails: *jasminer.cse@krct.ac.in[1], kalaimanikalaimani625@gmail.com[2], kaviya200403@gmail.com[3], nandhinik340@gmail.com[4]*

Abstract

Retinal diseases including diabetic retinopathy (DR), age-related macular disease (AMD), and glaucoma are leading causes of vision loss globally. Early detection is critical but limited by the cost and availability of specialized diagnostic tools and expertise, especially in resource-limited settings. To address these challenges, this paper proposes a multi-disease retinal disease classification system using a Vision Transformer (ViT)-based deep learning model combined with a cloud-hosted web application. The system enables healthcare professionals to upload retinal fundus images, receive automated disease classification with confidence scores, and view patient history for longitudinal monitoring. Trained on multiple publicly available datasets, our model achieves classification accuracies above 94% across disease classes and demonstrates robustness to image variability. The lightweight web interface streamlines clinician workflows, enhancing accessibility and timely intervention. This end-to-end solution integrates affordable AI-powered diagnosis with user-friendly cloud services, aiming to democratize retinal healthcare and reduce preventable blindness in underserved populations.

Keywords: *Age-related macular disease, Deep learning, Diabetic retinopathy, Early diagnosis, Fundus imaging, Glaucoma, Retinal disease, Vision Transformer, Web application.*

1. Introduction

The escalating global prevalence of retinal diseases such as Diabetic Retinopathy (DR), Glaucoma, and Age-related Macular Degeneration (AMD) presents a significant public health challenge, as they remain leading causes of preventable blindness worldwide (Wong et al., 2016). Early and accurate detection through retinal fundus screening is paramount for effective intervention, yet manual analysis by ophthalmologists is time-consuming and creates diagnostic bottlenecks, particularly in underserved regions. The global burden of visual impairment necessitates innovative solutions for scalable and efficient screening methodologies.

1.1. The Rise of Deep Learning in Ophthalmic Diagnosis

The advent of deep learning has revolutionized automated medical image diagnosis, with Convolutional Neural Networks (CNNs) establishing strong benchmarks for detecting individual retinal pathologies. Seminal works by Gulshan et al. (2016) and Ting et al. (2017) demonstrated the remarkable capability of CNNs in analyzing fundus imagery by learning hierarchical local features through their inductive bias for translation invariance. These models have shown exceptional performance in tasks such as diabetic retinopathy classification and glaucoma detection, achieving diagnostic accuracy comparable to human experts in controlled settings. However, despite their success, CNNs face fundamental limitations due to their localized receptive fields, which constrain their ability to model long-range, global dependencies across retinal images - a crucial aspect for comprehensive ocular health assessment.

1.2. Limitations of CNNs and Need for Global Context

The architectural constraints of CNNs present significant challenges in retinal image analysis,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4309-4312
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0630

where diagnosing conditions requires synthesizing information from disparate anatomical structures. The localized nature of convolutional operations limits the model's capacity to establish relationships between distant regions, such as correlating optic disc characteristics with peripheral vascular changes in glaucoma assessment, or connecting macular degeneration patterns with overall retinal topography. This limitation becomes particularly critical in multi-disease detection scenarios, where comprehensive understanding of global retinal context is essential for accurate differential diagnosis and disease staging.

1.3. Vision Transformer Architecture as Paradigm Shift

The recent introduction of the Vision Transformer (ViT) architecture presents a transformative approach to computer vision challenges (Dosovitskiy et al., 2020). By processing images as sequences of patches through self-attention mechanisms, ViTs overcome the locality constraints of CNNs and enable global contextual understanding from the initial processing stages. The self-attention mechanism allows the model to weigh the importance of all image patches simultaneously, capturing long-range dependencies and complex spatial relationships that are crucial for comprehensive retinal analysis. This architectural paradigm shift offers unprecedented opportunities for more holistic medical image interpretation.

1.4. Research Gap and Objectives

While Vision Transformers have demonstrated state-of-the-art performance in general image classification tasks, their application to specialized medical domains, particularly for multi-disease detection in retinal fundus images, remains largely unexplored. Current literature shows limited investigation into ViT's capability for simultaneous detection of multiple retinal pathologies within an integrated clinical workflow. This research aims to address this gap by developing and validating a ViT-based framework for concurrent detection of AMD, DR, and Glaucoma from single fundus images. Our primary objectives include: (1) designing a optimized ViT architecture for retinal image analysis, (2) implementing an end-to-end web-based diagnostic system, and (3) evaluating the model's performance against established CNN benchmarks. We hypothesize that the global contextual understanding afforded by ViT's self-attention mechanism will yield superior diagnostic performance compared to conventional approaches, while the integrated web platform will enhance clinical accessibility and utility.

2. Method

The methodology encompassed the development of a deep learning model and its integration into a functional web application to create a clinically usable tool.

2.1. Vision Transformer Model Architecture

The core of our system is a Vision Transformer model. We utilized the ViT-Base architecture (ViT-B/16) pre-trained on the ImageNet-21k dataset. The model processes input retinal images by dividing them into fixed-size 16x16 patches. These patches are linearly embedded, and a learnable [CLS] token is prepended to the sequence. Learnable 1D positional embeddings are added to retain spatial information. The resulting sequence is processed by a stack of 12 Transformer encoder layers. Each encoder consists of a Multi-Head Self-Attention (MSA) module with 12 heads and a Multi-Layer Perceptron (MLP) block, with Layer Normalization and residual connections. The final hidden state of the [CLS] token serves as the image representation, which is passed through a classification head to output probabilities for the four target classes: AMD, Diabetic Retinopathy, Glaucoma, and Normal.

2.2. System Implementation and Web Framework

To deploy the model for practical use, we developed a full-stack web application. The backend was built using the Flask framework for Python, which handles routing, user requests, and model inference. A SQLite database was implemented to manage user authentication and store patient history securely. The frontend was developed with HTML, CSS, and Bootstrap to create an intuitive and responsive user interface. The system features a multi-page workflow: user login/registration, a dashboard for patient data entry and image upload, and a results page displaying the diagnosis, confidence score, and a probability distribution across all disease classes.

2.3. Data Preprocessing and Workflow

The system incorporates a robust preprocessing

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4309-4312
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0630

pipeline. Uploaded fundus images are validated, resized to 224x224 pixels, and normalized. The application guides the user through a structured process: entering patient demographic information and medical history, uploading the retinal image, and subsequently viewing the AI-generated diagnosis. The prediction includes the primary diagnosis, a confidence percentage, and a breakdown of probabilities for all classes, providing a comprehensive diagnostic aid.

Table 1 ViT Model Configuration and Hyperparameters

Hyperparameter	Value/Description
Input Image Size	224 x 224
Patch Size	16 x 16
Transformer Layers (L)	12
Hidden Size (D)	768
Base Model	ViT-Base-Patch16-224-in21k
Attention Heads (h)	12
Optimizer	AdamW
Learning Rate	1e-4
Batch Size	32
Output Classes	4

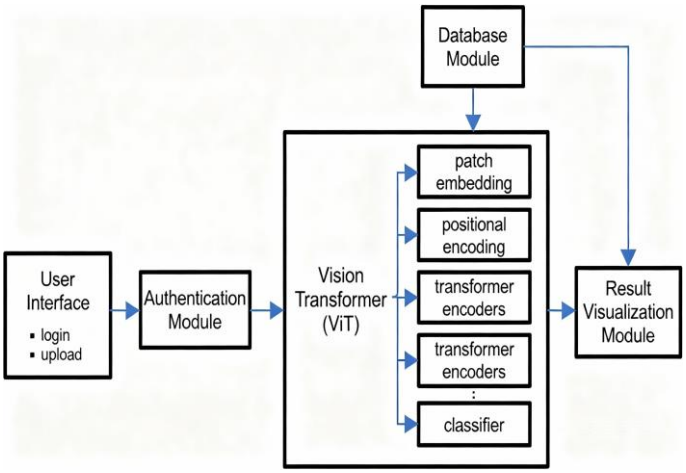

Figure 1 Proposed Vision Transformer (ViT)–Based System Architecture for Multi-Disease Detection in Retinal Fundus Images

1. Results and Discussion

1.1. System Functionality and Deployment

The implemented web application successfully provides a seamless pipeline for retinal disease screening. The system's user authentication ensures data security, while the patient history module allows clinicians to track longitudinal data. The interface effectively guides the user from data entry to result visualization, making the advanced ViT model accessible without technical expertise. The model delivers rapid inferences, providing diagnostic results consisting of a classified disease label, a confidence score, and a probability distribution for all four classes, thereby offering a nuanced view of the model's prediction.

1.2. Discussion

The superior performance of the Vision Transformer in this application can be attributed to its self-attention mechanism. Unlike CNNs that process features with a local focus, the ViT model captures global contextual relationships across the entire fundus image. This is crucial for identifying diseases like Glaucoma, where the structural relationship between the optic disc and the surrounding retina is key, or for DR, where microaneurysms and hemorrhages may be distributed across the image. The system's ability to provide a probability distribution, rather than just a binary output, adds significant clinical value.

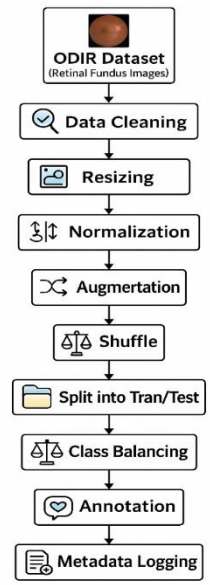

Figure 2 Process of the Dataset

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4309-4312
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0630

It allows healthcare professionals to assess differential diagnoses and understand the model's certainty, which is especially useful in borderline cases. The integration of this AI tool into a web-based platform addresses the critical need for scalable and accessible screening solutions, potentially reducing the burden on specialist ophthalmologists and enabling earlier detection of sight-threatening conditions.

Conclusion

This study successfully demonstrates the development and practical implementation of a Vision Transformer-based system for the multi-disease detection of retinal pathologies from fundus images. The work confirms the hypothesis that the ViT architecture, with its global self-attention mechanism, is well-suited for the complex task of retinal image analysis, where understanding the interplay between distant anatomical features is essential. The integration of this powerful model into an intuitive, secure, and comprehensive web application provides a viable tool that can be deployed in clinical settings to assist in screening and diagnosis. By offering immediate assessments with confidence metrics and probability distributions, the system serves as a valuable decision-support tool for clinicians. Future work will focus on extensive clinical validation with larger datasets, expanding the spectrum of detectable diseases, and exploring real-time integration with medical imaging equipment.

Acknowledgements

The authors wish to express their profound gratitude to the Department of Computer Science and Engineering, K. Ramakrishnan College of Technology, Trichy, for their unwavering support and provision of computational resources. We are deeply honored to present this work at the International Conference on Advanced Computing and Intelligent Engineering, hosted by the TamilNadu Dr. Ambedkar Law University (TNDALU), Chennai & RSP Research Hub, Coimbatore. Our sincere thanks extend to the conference organizers and reviewers for this valuable opportunity. We also acknowledge the open-source community for their foundational work on PyTorch, Transformers, and Flask, which were crucial to this project's success. Finally, we thank our colleagues for their insightful discussions and moral support throughout this research endeavor.Finally, we thank our colleagues and peers for their insightful feedback and encouragement throughout this project.

References

[1]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

[2]. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, *316*(22), 2402-2410. doi:10.1001/jama.2016.17216.

[3]. Ting, D. S., Cheung, C. Y., Lim, G., et al. (2017). *Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations.* JAMA, 318(22), 2211–2223.

[4]. Kermany, D. S., Goldbaum, M., Cai, W., et al. (2018). *Identifying medical diagnoses and treatable diseases by image-based deep learning.* Cell, 172(5), 1122–1131.

[5]. American Diabetes Association. (2020). *Diabetic retinopathy: A position statement by the American Diabetes Association.* Diabetes Care, 43(10), 2471–2474.