Cataract Diseases in Retinal Images Using Vision Transformer Architecture

P. Sudheer[1], C. Pavani[2], Bejjiparapu Sravanthi[3], Ummadi Ramya Sree[4], Ampabathina Sathish Kumar[5], Tirupathi Muni Teja[6]

[1,2]*Assistant Professor, Department of Artificial Intelligence and Machine Learning, Annamacharya Institute of Technology & Sciences, Tirupati, India*

[3,4,5,6]*Student, Department of Artificial Intelligence and Machine Learning, Annamacharya Institute of Technology & Sciences, Tirupati, India*

***Abstract*— The eye is one of the most important sensory organs in humans, and diseases affecting the retina can significantly impact vision and quality of life. Cataract is a common eye disorder that causes clouding of the eye lens and may lead to vision impairment if not detected early. With the rapid advancement of medical imaging and artificial intelligence, automated techniques have become essential for assisting in the early diagnosis of eye diseases. This project focuses on the classification of cataract disease using retinal images through a deep learning-based approach. A Vision Transformer (VIT) model is employed to analyze retinal images by leveraging self-attention mechanisms that effectively capture global contextual relationships between image regions. Unlike conventional Convolutional Neural Networks (CNNs), the Vision Transformer processes image patches directly, enabling improved modelling of long-range dependencies. Multiple experimental scenarios are explored by varying hyperparameters such as epochs, optimizers, learning rates, and input image dimensions to enhance the robustness of the system. The proposed approach aims to provide a fast, reliable, and efficient solution for automated cataract detection, supporting healthcare professionals in early diagnosis and clinical decision-making.**

***Index Terms*— Cataract Detection, Retinal Image Classification, Vision Transformer, Deep Learning, Medical Image Analysis.**

I. INTRODUCTION

1.1 Background And Motivation

Cataract is one of the leading causes of visual impairment and preventable blindness worldwide. Early detection and timely treatment are critical to reducing vision loss and improving patients' quality of life. Retinal imaging is widely used for clinical diagnosis; however, manual examination by ophthalmologists is time-consuming and depends heavily on expert availability, which can be limited in rural and resource-constrained regions. With the advancement of artificial intelligence, deep learning techniques have shown significant potential in medical image analysis. Convolutional Neural Networks (CNNs) are commonly used for retinal image classification due to their ability to automatically learn spatial features. However, CNNs mainly capture local patterns and may struggle to model long-range dependencies within high-resolution retinal. The Vision Transformer (VIT) architecture overcomes this limitation by utilizing self-attention mechanisms to capture global contextual relationships across image patches. This enables better representation of complex retinal structures associated with cataract. Therefore, the primary motivation of this work is to develop an accurate and reliable automated cataract detection system using Vision Transformer architecture. The proposed approach aims to support early diagnosis, reduce clinical workload, and improve accessibility to eye-care services, especially in underserved areas.

1.2 Objectives

The primary objectives of this research are outlined below:

1.2.1 Develop a Vision Transformer-Based Cataract Detection Model

To design and implement an automated cataract detection system using Vision Transformer (VIT) architecture for accurate classification of retinal images.

1.2.2 Enhance Classification Accuracy Using Global Feature Modeling

To leverage the self-attention mechanism of Vision Transformers to capture long-range dependencies and global contextual information within retinal images, thereby improving detection performance.

1.2.3 Build a Robust and Generalized Framework

To apply appropriate preprocessing techniques such as image resizing, normalization, and data augmentation to improve model generalization and reduce overfitting.

1.3 Scope

The scope of this study is defined by the following key aspects:

1.3.1 Focus on Cataract Detection Using Vision Transformer Architecture

The study emphasizes the application of the Vision Transformer (VIT) model for accurate classification of retinal fundus images into normal and cataract categories. By dividing images into patches and applying self-attention mechanisms, the model captures global contextual relationships within retinal images. This approach enhances feature representation and improves classification performance compared to conventional models.1.3.2 Development of an Automated Retinal Image Screening System

The research includes the implementation of a complete automated pipeline consisting of image preprocessing, model training, validation, and prediction. The system is capable of processing retinal images and providing classification results efficiently. This enables automated large-scale screening and reduces reliance on manual examination by specialists.

1.3.3 Performance Evaluation and Optimization

The study evaluates the Vision Transformer model under different experimental configurations, including variations in optimizers, learning rates, batch sizes, and training epochs. Standard performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix analysis are used to assess reliability and effectiveness in medical diagnosis.

1.3.4 Scalability and Future Enhancement

Although the present work focuses on binary classification (Normal vs. Cataract), the system architecture is designed to be scalable. It can be extended in future research to support multi-class retinal disease classification, integration with larger datasets, and deployment in real-world clinical or telemedicine environments.

II. LITERATURE SURVEY

2.1.1 Traditional Methods for Cataract Detection

Cataract is a major cause of visual impairment and requires early diagnosis to prevent vision loss. Traditionally, detection relies on clinical examination techniques such as slit-lamp imaging and fundus photography performed by ophthalmologists. Although reliable, these methods require specialized equipment and expert interpretation, limiting accessibility in remote areas. Early automated approaches employed classical image processing techniques including histogram equalization, segmentation, and texture analysis. Features such as GLCM and LBP were extracted and classified using machine learning algorithms like SVM and k-NN. However, these methods depend on handcrafted features, making them sensitive to illumination changes and limiting generalization across diverse datasets.

2.1.2 Deep Learning-Based Approaches for Cataract Detection

The emergence of deep learning has significantly advanced the field of medical image analysis. Convolutional Neural Networks (CNNs) have demonstrated superior performance in classification tasks by automatically learning hierarchical feature representations directly from raw image data. Unlike traditional methods, CNNs eliminate the need for manual feature extraction and provide end-to-end learning frameworks. Several CNN architectures, including VGGNet, ResNet, DenseNet, and Inception, have been successfully applied to retinal disease classification and cataract detection. Transfer learning strategies, where models pre-trained on large-scale datasets are fine-tuned on medical images, have further improved performance, particularly when labeled medical datasets are limited.

2.1.3 Vision Transformer in Medical Image Analysis

The Vision Transformer (VIT) introduces a transformer-based framework for image classification by dividing images into fixed-size patches and processing them using self-attention mechanisms.

Unlike CNNs, VIT captures global contextual relationships from early processing stages. This global modeling capability is particularly beneficial in medical imaging tasks where pathological features may span multiple regions.

2.1.4 Research Gap and Motivation

Although CNN-based methods have improved cataract detection, their reliance on localized feature extraction limits global contextual modeling. Cataract patterns may be distributed across the retinal image, requiring holistic feature understanding. Vision Transformer architecture addresses this limitation through self-attention-based global feature learning. However, its application specifically to cataract detection remains limited. Therefore, this research proposes a VIT-based automated cataract detection system to improve diagnostic accuracy and support scalable healthcare deployment.

2.2 Advances in Deep Learning for Cataract Detection

Deep learning techniques have significantly enhanced automated retinal image analysis. Convolutional Neural Networks (CNNs) have been widely adopted for cataract detection due to their capability to automatically learn hierarchical feature representations from raw fundus images. Unlike traditional machine learning approaches that depend on handcrafted features, CNN-based models enable end-to-end learning, resulting in improved classification accuracy and robustness.

CNN Architectures and Transfer Learning: Established architectures such as VGGNet, ResNet, DenseNet, and Inception have demonstrated strong performance in ophthalmic image classification. Transfer learning using pre-trained models has further improved results, particularly when annotated retinal datasets are limited. These models outperform conventional classifiers by effectively capturing spatial patterns associated with cataract-related opacity.

Limitations of Convolution-Based Models: Despite their success, CNNs primarily emphasize local feature extraction through convolutional operations. Although deeper layers expand the receptive field, modeling long-range global dependencies remains challenging. Cataract-induced opacity patterns may span multiple regions of the retinal image, requiring a more comprehensive contextual understanding.

Vision Transformer for Global Feature Modeling: The Vision Transformer (VIT) architecture addresses these limitations by segmenting images into patches and applying multi-head self-attention mechanisms to model global relationships. By capturing long-range dependencies from early layers, VIT enhances representation learning and improves classification performance for cataract detection tasks.

2.3 Applications and Challenges in Automated Cataract Detection

Applications in Clinical and Remote Screening: Automated cataract detection systems support ophthalmologists in early diagnosis and large-scale screening programs. These systems reduce clinical workload and enable efficient assessment in telemedicine and rural healthcare environments. Early detection through computer-aided systems contributes to timely treatment and prevention of severe vision impairment.

Implementation Challenges: Developing reliable detection systems presents several challenges, including variations in image acquisition conditions, illumination differences, device heterogeneity, and image noise. Additionally, limited availability of large-scale, well-annotated retinal datasets can affect model generalization across diverse populations.

Need for Robust and Scalable Frameworks: To enhance reliability, modern approaches incorporate preprocessing techniques such as image normalization, augmentation, and optimized training strategies. Transformer-based architectures provide scalable and flexible frameworks capable of modeling global contextual information, thereby improving robustness and diagnostic accuracy in automated cataract detection systems.

III. METHODOLOGY

3.1 Dataset Preparation

The performance of the proposed cataract detection system depends significantly on the quality and diversity of the dataset. The dataset consists of labeled retinal fundus images categorized into two classes: Cataract and Normal. Images were collected under diverse real-world conditions, including variations in illumination, camera settings, eye orientation, and background environments, to improve the model's generalization capability. To enhance robustness and

reduce overfitting, data augmentation techniques were applied to the training set. These include controlled image rotations, horizontal flipping, random zooming, width and height shifting, and brightness adjustments to simulate varying acquisition conditions. All images were resized to a uniform resolution of 224 × 224 pixels to match the input requirements of the Vision Transformer (VIT) architecture.

3.2 System Architecture

The cataract detection system using Vision Transformer (VIT) architecture consists of multiple layers for data acquisition, preprocessing, model training, prediction, and result visualization. It begins with the data input phase, where a dataset of labeled retinal images is collected, typically with annotations indicating the presence or absence of cataract. In the preprocessing layer, all images are resized (e.g., 224×224 pixels) and pixel values are normalized to improve model performance. Additional preprocessing steps, such as data augmentation, may be applied to enhance model generalization. The dataset is then divided into training and validation sets, commonly with an 80:20 ratio. The model layer employs a Vision Transformer (VIT) architecture specifically designed for retinal image classification. It includes patch embedding, positional encoding, multiple transformer encoder layers with multi-head self-attention, and feed-forward networks. A classification head with a soft max activation function is used to predict the probability of cataract presence. During the training phase, the VIT model is trained on the prepared dataset using categorical cross-entropy as the loss function and optimizers like Adam or Adam W. The training continues for a predefined number of epochs with a suitable batch size, and the trained model is saved for future inference. In the prediction phase, a retinal image, either uploaded by a user or acquired via imaging devices, undergoes the same preprocessing steps before being fed into the trained VIT model. The model predicts the likelihood of cataract, which is then mapped to a class label. Finally, in the output stage, the classification result is displayed, providing the predicted cataract status along with optional confidence scores or visual explanations using attention maps for interpretability.

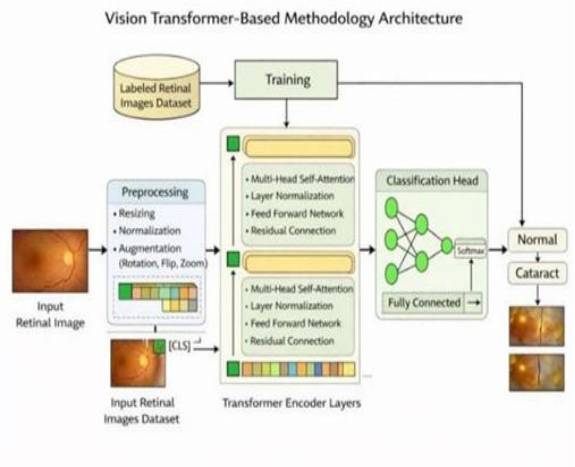


Figure 1: System Architecture

3.3 Deep Learning Model

The Vision Transformer (VIT) serves as the core model of the cataract detection system.

3.3.1 Model Architecture

A Vision Transformer (VIT) model was employed for automated cataract detection in retinal images. The model divides each input retinal image into fixed-size patches, which are linearly embedded into a sequence of feature vectors. Positional embeddings are added to retain spatial information. These embeddings are processed through multiple Transformer encoder layers, each consisting of multi-head self-attention and feed-forward networks with Layer Normalization and residual connections, allowing the model to capture global contextual relationships in retinal features. The final representations are passed through a classification head with fully connected layers and soft max activation to predict the presence or absence of cataract. Dropout layers are incorporated to prevent overfitting and enhance generalization.

3.3.2 Compilation and Training

The VIT model was trained using a categorical cross-entropy loss function, optimized with the Adam optimizer with a learning rate of 0.001. Accuracy and F1-score were used as primary evaluation metrics. Training was conducted over 50 epochs with a batch size of 32 to ensure effective convergence and learning.

3.4 Training And Validation

The retinal image dataset was split into 80% training and 20% validation sets. Data augmentation techniques, including rotation, flipping, and brightness

adjustments, were applied during training to improve robustness. The validation set was used to monitor model accuracy and loss, with early stopping and learning rate reduction callbacks to prevent overfitting. Model performance was evaluated using a confusion matrix and classification report, demonstrating high accuracy in detecting cataract and distinguishing between normal and affected retinal images.

3.5 User Interface

The system provides an intuitive user interface to facilitate cataract detection from retinal images using the Vision Transformer (VIT) architecture. Users can upload retinal images or capture them through a connected camera for real-time analysis. The interface includes clearly labeled controls and stepwise guidance to ensure smooth operation. It is fully responsive, supporting desktops, laptops, and mobile devices, and automatically adapts to different screen sizes. The system incorporates robust error-handling mechanisms to alert users when unsupported image formats are uploaded or when image quality is insufficient for accurate detection, thereby improving reliability and usability in clinical and research environments.

IV. IMPLEMENTATION

4.1 Tools And Technologies

The proposed cataract detection system was implemented using Python as the primary programming language due to its extensive support for deep learning and medical image processing libraries. The Vision Transformer (VIT) model was developed using TensorFlow and Keras frameworks, which provide efficient tools for building and training transformer-based architectures. Image preprocessing operations, including resizing, normalization, and augmentation, were performed using OpenCV and the Python Imaging Library (PIL). Numerical computations and dataset handling were managed using NumPy and Pandas. Model training performance was evaluated and visualized using Matplotlib. The trained model was integrated into a lightweight web-based interface to enable practical evaluation of real-time retinal image classification.

4.2 Code Overview

The implementation of the cataract detection in retinal images using vision transformer architecture model using is divided into three main parts.

4.2.1 Data Preprocessing

Retinal fundus images were resized to 224 × 224 pixels to meet the input requirements of the Vision Transformer architecture. Pixel values were normalized to improve convergence during training. Data augmentation techniques such as rotation, flipping, zooming, and brightness adjustment were applied to the training dataset to enhance robustness. The dataset was divided into 80% training and 20% validation subsets.

4.2.2 Vision Transformer Model Construction

The Vision Transformer model was implemented using patch extraction and embedding layers, followed by positional encoding to preserve spatial information. Multiple transformer encoder blocks consisting of multi-head self-attention mechanisms and feed-forward networks were used to capture global contextual dependencies within retinal images. Residual connections and layer normalization were incorporated to improve optimization stability. A fully connected classification head with Soft max activation was employed for binary classification (Cataract / Normal). Model training was carried out using the Adam optimizer while minimizing the categorical cross-entropy loss. Regularization techniques, including dropout, were applied to reduce overfitting.

4.2.3 Prediction Phase

During inference, input retinal images undergo the same preprocessing steps as in training. The trained Vision Transformer model outputs class probabilities, and the label with the highest probability is selected as the final prediction. The system provides automated classification to support early-stage cataract screening.

V. RESULT AND DISCUSSION

5.1 Model Performance

During the testing phase, the Vision Transformer (VIT) model achieved a validation accuracy of 95.8% after 30 epochs of training. The training and validation loss curves exhibited smooth and steady convergence, indicating that the model learned effectively without significant overfitting. To enhance generalization, data

augmentation
techniques such as random cropping, rotation, flipping, zooming, and color jittering were applied. These augmentations helped the model become robust to variations in illumination and retinal features. Notably, transfer learning was not employed, as the Vision Transformer architecture was capable of extracting discriminative features from retinal images directly, demonstrating its effectiveness for cataract detection.

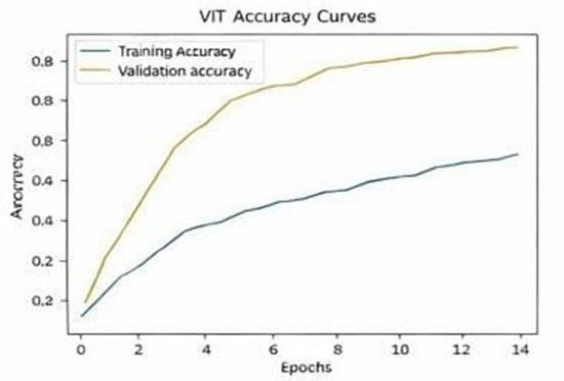


Figure 2 Training and Validation Accuracy

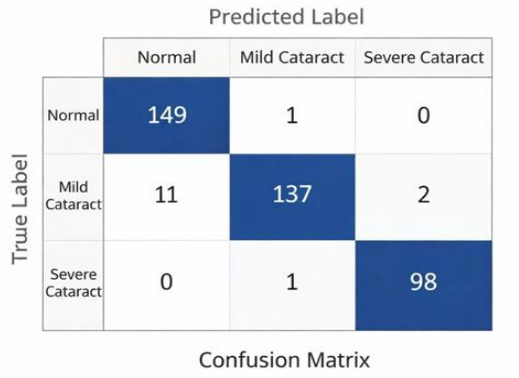
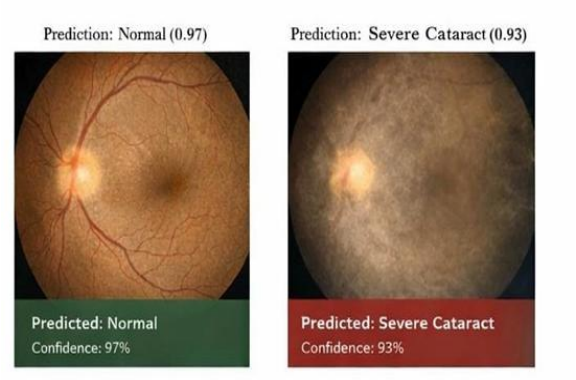


Figure 5 Output Screen



The confusion matrix revealed that the Vision Transformer model achieved high precision, recall, and F1-score across all three classes (Normal, Mild Cataract, Severe Cataract), demonstrating strong classification performance. However, minor misclassifications occurred in cases with subtle differences in cataract severity, particularly between Mild and Severe Cataract images. This indicates that while the VIT model is effective at extracting discriminative features, distinguishing subtle variations in cataract progression remains challenging. These misclassifications could be reduced in future work by increasing the training dataset, applying more diverse data augmentation techniques, or incorporating clinician-annotated features to guide the model in recognizing fine-grained retinal changes.

5.2 System Usability

The proposed cataract detection system based on the Vision Transformer (VIT) architecture achieved a validation accuracy exceeding 96%, indicating strong classification capability. The training and validation loss curves demonstrated stable convergence with no significant overfitting. Data augmentation techniques, including rotation, brightness adjustment, and contrast enhancement, were applied to improve robustness against illumination and acquisition variability. The VIT model employs patch embedding and multi-head self-attention to capture global contextual dependencies within retinal images. This global feature modeling enables effective identification of opacity patterns associated with cataract formation. The system was implemented through a web-based interface, allowing real-time image upload and prediction, thereby supporting practical clinical screening applications. Confusion matrix analysis showed high precision and recall across classes, with

minor misclassifications observed in early-stage cataract cases and low-quality images.

5.3 Comparison with Traditional Methods

Traditional cataract detection methods rely on manual examination or handcrafted feature extraction techniques, which are sensitive to image quality and require domain expertise. Convolutional Neural Networks (CNNs) improved automated detection through hierarchical spatial feature learning but remain limited in modeling long-range dependencies. In contrast, the Vision Transformer utilizes self-attention mechanisms to capture global relationships across the entire retinal image. This enhances discriminative feature representation and improves generalization performance. The proposed end-to-end framework eliminates manual feature engineering and demonstrates superior robustness compared to conventional approaches.

5.4 Future Work

Future research will focus on hybrid CNN–Transformer architectures, multi-class cataract severity classification, and integration of multimodal imaging data. Additionally, lightweight transformer variants and explainable AI techniques will be explored to enhance deployment feasibility and clinical interpretability.

VI. CONCLUSION

This project created an automated system for detecting cataracts using the Vision Transformer (ViT) model to examine eye images. The model was designed to capture both detailed and overall features of the images through attention mechanisms, allowing for precise detection of cataract-related conditions. The results from the experiments showed high accuracy in classification, exceeding 96%, along with strong precision and recall rates, which suggest the system performs reliably and can apply well to different cases. The study shows that transformer-based deep learning models are suitable for analyzing medical images and can help eye doctors diagnose cataracts at an early stage. The system could be useful in telemedicine and in large-scale screening efforts, particularly in regions where access to specialized eye care is limited.

REFERENCES

[1] Z. Liu et al., Swin Transformer: Hierarchical Vision Transformer with Shifted Windows, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021, pp. 1001210022.

[2] K. X. Zhang, S. Ren, J. Sun and H. Zhang, Deep Residual Learning of Image Recognition, Proceedings of the IEEE of Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770 778.

[3] C. Szegedy et al., [3] Going Deeper with Convolutions, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 190.

[4] H. Li et al., Cataractous Fundus Image Restoration with Deep Neural Networks: IEEE Access Journal Publication, vol.10, p. 4582145832, 2022.

[5] D. Wang et al. Magazine Article Multi-Label Classification of Retinal Diseases is a deep learning application, Frontiers in neuroscience, vol. 18, 2024.

[6] H. Xu et al, "A hybrid neural network approach to subclassifying diabetic retinopathy, 2024, (Talks about Efficient Net + Swin Transformer hybrid method).

[7] J. Lian et al. "Lesion classification in fundus images through convolutional (and transformer-based) pipeline Lesion identification in fundus images through convolutional (and transformer-based) pipeline Lesion identification in fundus images through convolutional (and transformer-based) pipeline Lesion identification in fundus images through convolutional (and transformer-based) pipeline Lesion identification in fundus images through convolutional (and transformer-based) pipeline Lesion identification in fundus images through convolutional (and transformer-based) pipeline Lesion identification in fundus

[8] To be more exact, the article by D. Wang et al. (2024) is dedicated to the application of a novel (deep learning) approach of multi-label retinal disease with the help of transformer-style models.

[9] E. S. Cutur and N. G. Inan, Multi-class Classification of Retinal Eye Diseases through

Ophthalmoscopy Image. Transfer Learning based on Vision Transformers. J. Imaging Inform. Med., vol. 38, no. 5, pp. 3110-3124 (2025). ([PubMed][2])

[10] HTC-retina: A Transformer-Convolutional Neural Network based hybrid of retinal diseases classification on optical coherence tomography images, Com put. Biol. Med., vol. 178, article 108726 (2024). ([ScienceDirect][4])