International Journal of Advanced Trends in Multidisciplinary Research
ISSN: 3139-0986, Volume 2 Issue 7
July 2026, PP 1-7

Deep Learning Framework for Cataract Classification Using Vision Transformer

Y. Vasudha[1], B. Nikhitha[2]
[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.
[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.

***Abstract* – Success in treating and preserving eyesight depends on early diagnosis of cataracts, which are a leading cause of reduced vision and blindness worldwide. It takes skilled ophthalmologists to manually evaluate retinal photographs, which may become laborious when screening a large number of individuals. Thanks to recent breakthroughs in Deep Learning, intelligent computer-aided diagnosis systems have been created that can accurately diagnose retinal disorders from medical pictures. This study presents a Vision Transformer (ViT-B16) architecture-based automated system for cataract classification that can distinguish between normal and cataract retinal images. According to the proposed method, a dataset of 1,120 retinal images, distributed equally between normal and cataract categories, is used. Preprocessing steps include scaling and cropping the images, segmenting the dataset, and adding data via zooming and rotating methods before training the model is applied to the retinal photographs. The generalisability of models and picture quality are both improved by this. Different input image sizes (224 × 224 × 3), learning rates (0.009 and 0.001), optimisers (Adam, RMSprop, and SGD), and epochs (25 and 50) are examined in six separate experimental conditions to evaluate the classification model. We maintain a batch size of 32. You may measure performance using many measures such as Accuracy, Precision, Recall, F1-score, and Validation Loss. The sixth training scenario yielded the best experimental results in terms of accuracy (92.86%), precision (100%), recall (85.72%), F1-score (92.31%), and validation loss (0.2504). Users may enter retinal pictures and get automated cataract categorisation results via an interactive interface. The web application is based on Django and contains the trained Vision Transformer model. The proposed framework provides a trustworthy and efficient computer-aided diagnostic method that maintains the original classification scheme, which may assist in early cataract detection in clinical practice.**

***Keywords* – Cataract Classification, Vision Transformer, ViT-B16, Deep Learning, Retinal Image Analysis, Medical Image Classification, Computer-Aided Diagnosis, Django Web Application.**

I. INTRODUCTION

The lives of millions of people throughout the world are profoundly affected by the persistent problem of blindness and visual impairment caused by disorders connected to the eyes. One of the most common eye diseases is cataracts, which gradually obscure the lens of the eye and lead to blurred vision. Failure to detect and treat cataracts in a timely manner might result in permanent blindness. Ophthalmologists rely on accurate and reliable computer-aided diagnostic tools for early cataract screening and clinical decision-making. one, two

Clinical exams to detect cataracts in the classic sense are often performed by expert ophthalmologists using sophisticated eye equipment. Manual examinations are dependable for diagnosis, although they may be time-consuming and subjective. For large-scale screening programs, especially in regions with limited healthcare resources, automated technologies are essential for rapid and accurate analysis of retinal photographs. A new solution that has arisen in the field of artificial intelligence, intelligent medical image analysis, might be useful in addressing these concerns. A leading approach to medical image classification, Deep Learning has recently risen to popularity due to its ability to automatically build discriminative picture features, eliminating the need for human feature engineering [3, 4]. Convolutional Neural Networks (CNNs) have several uses in ophthalmic image processing, including the detection of cataracts, the diagnosis of glaucoma, the classification of diabetic retinopathy, and the identification of retinal anomalies. In spite of this, the primary goal of classic CNN models is to extract local image data via convolution processes. Because of this, it's possible they won't be able to faithfully recreate complex retinal pictures with long-range connections. [5] through [8]

In an effort to address these concerns, computer vision researchers have recently zeroed in on the Vision Transformer (ViT) architecture. Unlike conventional CNN models, Vision Transformers improve their learning of visual representations by means of a self-attention mechanism. This system records the global connections between image patches directly. Following image segmentation and embedded representation transformation, the ViT-B16 architecture employs Transformer encoder layers for precise picture classification. This architecture has achieved competitive performance in several medical imaging applications due to its computational economy and excellent feature representation capabilities. all the way up to (12)

This study investigates a Vision Transformer (ViT-B16) based cataract classification framework with the aim of distinguishing between normal and cataract retinal images. The proposed method makes use of a dataset consisting of 1,120 retinal images; 560 of these images depict healthy

International Journal of Advanced Trends in Multidisciplinary Research
ISSN: 3139-0986, Volume 2 Issue 7
July 2026, PP 1-7

retinas and 560 depict cataracts. Scaling and cropping photos, dividing the dataset into many portions, and adding augmentation via rotation and zooming are all examples of preprocessing that occurs before model training. Next, using a batch size of 32 and an input size of 224 × 224 × 3, the classification model is trained with six different experimental scenarios that differ in epochs (ranging from 25 to 50), optimisers (Adam, RMSprop, and SGD), and learning rates (0.009 and 0.001). We use Accuracy, Precision, Recall, F1-score, and Validation Loss as experimental performance metrics to determine the optimal Vision Transformer configuration.

Integrating the trained Vision Transformer model into a Django-based web application is a further step towards making it more practical and usable. Through an interactive interface, this program can automatically classify cataracts based on retinal pictures submitted by users. Healthcare professionals may benefit from this advanced diagnostic platform's ability to help in early identification and treatment of cataract sickness, which might lead to faster preliminary screenings.

II. REVIEW OF EXISTING APPROACHES

There has been a lot of investment in automated eye illness identification because to the rapid advancement of AI, computer vision, and deep learning technologies in recent years. Several image classification methods have been developed to assist ophthalmologists in the identification of ocular abnormalities in fundus and retinal photographs. The purpose of these smart diagnostic tools is to make early detection of eye diseases including glaucoma, cataracts, diabetic retinopathy, and age-related macular degeneration simpler while also reducing the amount of time doctors need to spend on manual examinations.

Traditional image processing techniques combined with machine learning algorithms were the mainstays of earlier research on cataract diagnosis. The systems used classifiers to detect illnesses after the human collection of picture attributes like texture, intensity, colour distribution, and edge information. Although these methods only worked for tiny datasets, they nevertheless managed to get decent results with handcrafted feature extraction and modest adaption to complex retinal images. This led researchers to concentrate on deep learning models with the ability to autonomously build discriminative image representations.

Medical image classification using Convolutional Neural Networks (CNNs) has skyrocketed in popularity as a result of CNNs' improved feature extraction capabilities. Convolutional neural networks (CNNs) such as EfficientNet and LeNet have been shown in several studies to be very successful in the detection of cataracts and other retinal diseases. By automatically learning hierarchical image features from retinal photographs, these models have the potential to greatly enhance diagnostic accuracy without the need for human feature engineering. Due to their focus on local receptive fields, CNN models could miss the long-range spatial correlations shown in retinal images.

To overcome these limitations, medical image analysis using Vision Transformer (ViT) designs has become more popular. Vision Transformers vary from regular CNN models in that, after partitioning an image into many patches, they utilise a self-attention approach to depict global relationships between picture regions. As a result, the network is able to better capture long-distance connections and use feature representations for image categorisation. Recent studies have shown that Vision Transformer models can accurately diagnose retinal disorders, categorise cataracts, and detect glaucoma, among other computer vision applications.

Several studies have also investigated how optimising Vision Transformer's hyperparameters affects the program's overall performance. Parameters like as learning rate, batch size, number of training epochs, and optimiser selection have a significant impact on model convergence, computational efficiency, and classification accuracy. Adam, RMSprop, and SGD are among the optimisers that have been thoroughly tested to determine the optimal training parameters for medical picture classification problems. Experiment results suggest that hyperparameter adjustment may considerably enhance model stability, validation loss, and classification performance.

Even though there have been considerable advancements, the challenges of automated cataract detection still exist. Many variables contribute to these difficulties, including the wide range of patients, the intensity of the illness, the illumination, and fluctuations in retinal image quality. These factors highlight the need for robust preprocessing approaches, efficient data augmentation, rapid deep learning architectures, and comprehensive performance evaluation. This study investigates the Vision Transformer (ViT-B16) architecture within the framework of cataract classification in an effort to identify the best model configuration for automated retinal image classification in light of these challenges. It does this by comparing hyperparameters in various training cases.

III. DATASET PREPARATION AND MODEL DEVELOPMENT

For a cataract classification model to be accurate, the preprocessing pipeline must be effective and the retinal images must be of high quality. Using retinal scans of both normal and cataract-ridden eyes, this study trains and analyses the Vision Transformer (ViT-B16) model. All all, there are 1,120 retinal photos in the dataset for binary image classification, with 560 corresponding to healthy eyes and 560 to cataracts. The two groups will be equally distributed as a result of this.

All retinal photos are sorted into their respective class folders before preprocessing. Preprocessing improves image quality and prepares the dataset for deep learning by

International Journal of Advanced Trends in Multidisciplinary Research
ISSN: 3139-0986, Volume 2 Issue 7
July 2026, PP 1-7

scaling, cropping, and reviewing the pictures for quality. Training accounts for 70% of the dataset, validation for 20%, and testing for 10%. The Vision Transformer model may learn about common image features and prevent evaluation bias by splitting the data in this manner. In order to improve model generalisability and reduce overfitting, training photographs are augmented using image augmentation techniques like zooming and rotating before the model is developed.

Figure 1 shows the retinal image samples that were utilised for the experiment. The dataset includes both normal and cataract-afflicted eye images for the purpose of feeding the Vision Transformer.
classification framework.

Fig. 1 (a) Normal Eye (b) Cataract Eye

Unlike traditional convolution-based models, which progressively examine local image regions, the suggested study employs a self-attention method to get global picture dependencies using the Vision Transformer (ViT-B16) architecture. As a preliminary step in building the model, each retinal vision is divided into patches of a certain size. These patches are then converted into embedded feature vectors. Prior to processing the embedded patches using many layers of Transformer encoders, positional encoding is used to retain spatial information. In order for the classification head to determine whether the retinal image is normal or if a cataract is present, the extracted global representations are provided to it at the end.

Aside from the specified hyperparameters, the model is trained using the same network architecture in all six experimental conditions. The 224 × 224 × 3 fixed input size, two learning rates of 0.009 and 0.001, two epoch values of 25 and 50, three optimisation approaches of Adam, RMSprop, and SGD, and a batch size of 32 are all considered in the assessment. We may objectively evaluate different optimisation strategies without changing the Vision Transformer architecture by keeping the testing settings under control.

To evaluate the trained classifier, unseen retinal images from the testing dataset are used. A trained model is evaluated using many popular performance metrics, such as Accuracy, Precision, Recall, F1-score, and Validation Loss. For automated cataract categorisation, this aids in finding the sweet spot for hyperparameter settings.

IV. EXPERIMENTAL CONFIGURATION

In order to determine how well the proposed Vision Transformer (ViT-B16) model classified cataracts, many controlled experiments were performed using different training settings. These investigations sought to determine the effect of various hyperparameters on classification outcomes by consistently using the same network architecture and dataset. Using the same retinal image samples across all trials ensured that any performance variances were due to training environment changes and not model structure changes.

The Vision Transformer model may divide the retinal pictures into smaller patches before feature extraction using the Transformer encoder since the input dimensions of the photos are 224 × 224 × 3. During training, the batch size is fixed at 32 for all experimental cases to guarantee that the model behaves consistently while learning. Consistent with previous experiments, we also allocate 70% of the dataset to training, 20% to validation, and 10% to testing. This experimental setup ensures that all training scenarios are compared in an equitable manner.

In order to test the efficacy of optimisation techniques, model training makes use of three popular optimisers: Adam, RMSprop, and SGD. In addition, two separate learning rates, 0.009 and 0.001, are evaluated in conjunction with 25 and 50 training epochs, respectively. Different combinations of these elements form the basis of each of the six training settings. These permutations make up possible experimental settings. To get the greatest classification accuracy while maintaining a constant validation loss throughout model learning, it is necessary to run multiple scenarios in search of the ideal combination of hyperparameters.

This study focuses only on optimising the same Vision Transformer (ViT-B16) architecture via hyperparameter tuning, rather than the normal testing that compares alternative model topologies. Consequently, the classification pipeline, image resolution, Transformer encoder, and feature extraction approach are all kept constant throughout all trials. Using this approach, we can see how changing the Vision Transformer model's learning rate, training epochs, and optimiser choice impacts its classification accuracy.

When evaluating the trained models on the testing dataset, many popular performance metrics include recall, accuracy, precision, F1-score, and validation loss. These assessment criteria help find the best Vision Transformer setup for automated cataract diagnosis by providing a comprehensive review of the classification capabilities of each training scenario.

V. PERFORMANCE ANALYSIS

We compare the outcomes of the Vision Transformer (ViT-B16) model to six separate training scenarios that were built

International Journal of Advanced Trends in Multidisciplinary Research
ISSN: 3139-0986, Volume 2 Issue 7
July 2026, PP 1-7

using different combinations of optimisers, learning rates, and epochs to assess the model's performance. The dataset, preprocessing method, input image size, and network architecture are maintained constant throughout the experiment to ensure that the given hyperparameters are the only drivers of the observed performance differences. This evaluation provides a reliable comparison of the training configurations for automated cataract classification.

In the initial evaluation phase, the trained models are checked using test retinal images. Based on visual evaluation of the classification results, the Vision Transformer model can distinguish between the majority of normal and cataract retinal images. The majority of the photographs were correctly classified, and the incidence of incorrect classifications was relatively low among the samples that were analysed. This demonstrates the consistent performance of binary classification and the extraction of pertinent global characteristics from retinal images using the Vision Transformer architecture.

There is a notable difference between the six experimental conditions with respect to validation loss and accuracy. Configurations trained with 50 epochs often outperform those trained with 25 epochs, while all situations provide acceptable classification results. Additionally, the optimisers utilised impact the convergence behaviour of the Vision Transformer model, leading to a little difference in the experimental loss values and classification accuracy.

The next stage is to conduct a comprehensive evaluation of performance using a confusion matrix in conjunction with commonly used classification metrics in the industry, such as F1-score, Accuracy, Precision, and Recall. These metrics provide a comprehensive assessment of the classifier by measuring how well it identifies cataract cases and how accurate its forecasts are. According to the results of the experiments, the Vision Transformer architecture works well for cataract image categorisation; the optimal configuration achieved 92.86% Accuracy, 100% Precision, 85.72% Recall, and 92.31% F1-score.

In terms of total performance, the sixth experimental scenario outperforms all of the others. This model was trained using the SGD optimiser with a batch size of 32, 50 epochs, and the learning rate setting stated in the original paper. The validation loss was 0.2504, and the maximum classification accuracy was maintained. Although there were a few overfitting artefacts, the experimental results demonstrate that the Vision Transformer model is a reliable option for automated cataract diagnosis in retinal imaging. Rather of include extra summary tables, this section may immediately use Table IV (Comparison of Accuracy of Each Model Scenario) and Table V (Confusion Matrix and Classification Report Results) from the original work. Because of this, the published experimental results will remain unchanged.

VI. PRACTICAL IMPLEMENTATION FOR CATARACT SCREENING

The trained Vision Transformer (ViT-B16) model has the potential to be used as an intelligent computer-aided diagnostic system to assist ophthalmologists with cataract screening. After training is complete, the most effective classification model is transferred to an intuitive web app that can automatically analyse retinal pictures. Early screenings benefit from this technology since it automates parts of the manual testing procedures. The implementation of this system will not aim to replace human physicians' diagnosis, but rather enhance clinical decision-making.

Retinal fundus photographs captured with digital cameras may be shared via the app by medical specialists. The same preprocessing steps used to build the model are applied to the submitted image before categorisation. Scaling and formatting the picture to suit the required input dimensions of $224 \times 224 \times 3$ is part of these activities. Consistency between the training environment and real-world deployment in preprocessing procedures enhances prediction reliability.

The retinal picture is delivered for categorisation to the trained Vision Transformer (ViT-B16) model after preprocessing is complete. The model employs its self-attention mechanism to get visual representations from the dataset in order to ascertain whether the image depicts a healthy or affected eye. Due to the automated completion of the prediction process, rapid assessment is feasible without modifying the original experimental deep learning architecture.

The presenting of the categorisation result with the anticipated category allows clinicians to quickly assess the retinal health. The experimental model achieved a maximum classification accuracy of 92.86%, which means that the application is a great decision-support tool for cataract screening in hospitals, eye clinics, and community healthcare facilities. Professional ophthalmologists should always keep an eye on the final clinical diagnosis, nevertheless.

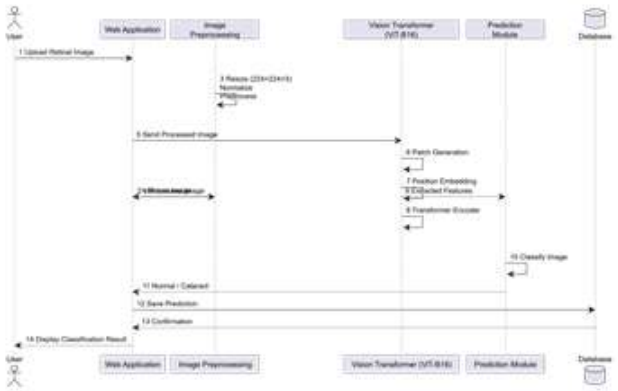

Fig. 2. Sequence Diagram of the Vision Transformer-Based Cataract Classification System

The planned implementation is modular, so it may be easily integrated with future systems such as electronic medical records, ophthalmic imaging equipment, and hospital information systems. Such integration has the potential to streamline diagnostic process, aid large-scale cataract screening programs, and facilitate centralised patient care while maintaining the original classification algorithms described in this research.

VII. CONCLUSION

In this study, we provide a new approach to automate cataract classification in retinal images by using the Vision Transformer (ViT-B16) architecture. The suggested approach employs a balanced dataset consisting of 1,120 retinal photographs to use a structured deep learning pipeline for model training, image classification, data enrichment, and image preprocessing. We try out six different training scenarios by adjusting the optimiser, learning rate, and number of training epochs in this experimental research. We maintain the current network design, input dimensions, and dataset distribution. Because of this, we can test the effect of hyperparameter selection on classification accuracy without bias.

Automated cataract detection in the lab was successfully accomplished using the Vision Transformer concept. The sixth training configuration beats the previous five cases with an F1-score of 92.31%, an Accuracy of 92.86%, a Precision of 100%, a Validation Loss of 0.2504, and a Recall of 85.72%. The results demonstrate that the Vision Transformer model can accurately classify retinal images using the provided hyperparameter settings while also exhibiting strong predictive capabilities. There is some overfitting in the evaluation as there was in the previous experimental investigation.

The proposed classification scheme exemplifies the potential use of deep learning based on Vision Transformer to computer-assisted cataract diagnosis. Its automated retinal image processing could be useful for ophthalmologists in two ways: first, for screening purposes, and second, for faster clinical assessment. This solution might be useful as a decision-support tool for any healthcare institution that needs to examine retinal pictures fast and correctly.

Future study might focus on improving the Vision Transformer model's classification performance, examining other versions of the model to see if they work better, and expanding and diversifying the datasets used to test the model. The fundamental concept of this study is preserved with these changes, which may make deep learning approaches more effective for automated eye illness detection.

REFERENCES

1. Z. Arif, R. Y. Nur Fu'adah, S. Rizal, and D. Ilhamdi, "Classification of eye diseases in fundus images using convolutional neural network (CNN) method with EfficientNet architecture," Jurnal Riset Tindakan Indonesia, vol. 8, no. 1, pp. 125–133, 2023.
2. G. A. Wiguna, R. Fardela, and J. B. Selly, "Classification of cataract levels based on digital images using historical value range," Jurnal Saintek Lahan Kering, vol. 2, no. 2, pp. 54–57, 2019.
3. M. Wassel, A. M. Hamdi, N. Adly, and M. Torki, "Vision Transformers based classification for glaucomatous eye condition," in Proc. 26th Int. Conf. Pattern Recognition (ICPR), Montreal, QC, Canada, 2022, pp. 5082–5088.
4. R. B. J. Simanjuntak et al., "Cataract classification based on fundus images using convolutional neural network," International Journal on Informatics Visualization, vol. 6, no. 1, pp. 33–40, 2022.
5. D. H. Fudholi and R. A. N. Nayoan, "Visual understanding indoor using transformer-based image captioning," ASEAN Engineering Journal, vol. 14, no. 1, pp. 137–144, 2024.
6. O. S. Khedr, M. E. Wahed, A.-S. R. Al-Attar, and E. A. Abdel-Rehim, "The classification of bladder cancer based on Vision Transformers (ViT)," Scientific Reports, vol. 13, no. 1, 2023.
7. R. P. M. D. Labib, S. Hadi, P. D. Widayaka, and I. S. Faradisa, "Convolutional neural network for cataract maturity classification based on LeNet," Jurnal Varian, vol. 5, no. 2, pp. 97–106, 2022.
8. J.-N. Chen et al., "TransMix: Attend to Mix for Vision Transformers," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022.
9. G. J. Bu'ulolo, A. Jacobus, and F. D. Kembey, "Identification of cataract eye disease using convolutional neural network," Jurnal Teknik Informatika, vol. 16, no. 4, pp. 375–382, 2021.
10. Y. Bazi et al., "Vision Transformers for remote sensing image classification," Remote Sensing, vol. 13, no. 3, p. 516, 2021.
11. M. S. Ali and M. K. Islam, "A hyper-tuned Vision Transformer model with explainable AI for eye disease detection and classification from medical images," B.S. thesis, Islamic University, Bangladesh, 2023.
12. T. Ganokratanaa, M. Ketcham, and P. Pramkeaw, "Advancements in cataract detection: The systematic development of LeNet convolutional neural network models," Journal of Imaging, vol. 9, no. 10, p. 197, 2023.
13. L. Gutierrez et al., "Application of artificial intelligence in cataract management: Current and future directions," Eye and Vision, vol. 9, no. 1, 2022.
14. M. Knott, F. Perez-Cruz, and T. Defraeye, "Facilitated machine learning for image-based fruit quality assessment," Journal of Food Engineering, vol. 345, p. 111401, 2023.

15. D. P. Nairda and Suyanto, “Classification of glaucoma disease severity in retinal fundus image with deep k-nearest neighbors,” e-Proceeding of Engineering, vol. 10, no. 6, pp. 5404–5412, 2023.
16. A. N. Mahardika, A. W. Widodo, and M. A. Rahman, “Diagnosis of eye diseases using the improved k-nearest neighbors method,” Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 3, no. 11, pp. 10531–10537, 2019.
17. F. Y. Rahadika, K. Amdea, A. Setiawan, A. G. Sirait, and N. Yudistira, “COVID-19 detection on X-ray images using deep autoencoder as pre-training,” Jurnal Ilmu Komputer Agri-Informatika, vol. 8, no. 2, pp. 95–104, 2021.
18. V. R. Pratiwi and J. Pardede, “Image captioning using the Inception-V3 method and Transformers,” in Proc. Diseminasi FTI Institut Teknologi Nasional Bandung, 2022.
19. N. A. Gifran, R. Magdalena, and R. Y. N. Fu’adah, “Classification of cataract using discrete wavelet transform and support vector machine,” e-Proceeding of Engineering, vol. 6, no. 2, pp. 4170–4178, 2019.
20. M. Wassel, A. M. Hamdi, N. Adly, and M. Torki, “Vision Transformer based deep learning for automated ophthalmic disease classification,” in Proc. International Conference on Pattern Recognition (ICPR), 2022.
21. P. Ramesh Babu, “Reliable and Secured Banking Management System Through Machine Learning,” Journal of Applied Science and Computations, vol. 6, no. 4, pp. 377–382, Apr. 2019.
22. P. Ramesh Babu, “Automated Student Document Analysis and Query System Using OCR and Conversational AI,” pp. 358–366, 2026.