EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article EAI.EU

# Glaucoma Classification using Light Vision Transformer

Piyush Bhushan Singh[1, *], Pawan Singh[1], Harsh Dev[2], Anil Tiwari[3], Devanshu Batra[4] and Brijesh Kumar Chaurasia[5]

[1] Department of Computer Science and Engineering, Amity School of Engineering and Technology
Lucknow, Amity University Uttar Pradesh, India
[2] Department of Computer Science and Engineering, Pranveer Singh Institute of Technology, Kanpur, India
[3] Amity School of Engineering and Technology Lucknow, Amity University Uttar Pradesh, India
[4] Department of Information Technology, Pranveer Singh Institute of Technology, Kanpur, India
[5] Department of Computer Science and Engineering, Pranveer Singh Institute of Technology, Kanpur, India

## Abstract

INTRODUCTION: Nowadays one of the primary causes of permanent blindness is glaucoma. Due to the trade-offs, it makes in terms of portability, size, and cost, fundus imaging is the most widely used glaucoma screening technique.
OBJECTIVES:To boost accuracy,focusing on less execution time, and less resources consumption, we have proposed a vision transformer-based model with data pre-processing techniques which fix classification problems.
METHODS: Convolution is a “local” technique used by CNNs that is restricted to a limited area around an image. Self-attention, used by Vision Transformers, is a “global” action since it gathers data from the whole image. This makes it possible for the ViT to successfully collect far-off semantic relevance in an image. Several optimizers, including Adamax, SGD, RMSprop, Adadelta, Adafactor, Nadam, and Adagrad, were studied in this paper. We have trained and tested the Vision Transformer model on the IEEE Fundus image dataset having 1750 Healthy and Glaucoma images. Additionally, the dataset was preprocessed using image resizing, auto-rotation, and auto-adjust contrast by adaptive equalization.
RESULTS: Results also show that the Nadam Optimizer increased accuracy up to 97% in adaptive equalized preprocessing dataset followed by auto rotate and image resizing operations.
CONCLUSION: The experimental findings shows that transformer based classification spurred a revolution in computer vision with reduced time in training and classification.

**Keywords:** Glaucoma, CNN models, Vision Transformer Model, Optimizers, Fundus imaging

Received on 11 July 2023, accepted on 05 September 2023, published on 21 September 2023

doi: 10.4108/eetpht.9.3931

*Corresponding author. Email: Piyush.bhadauria@gmai.com

## 1. Introduction

Glaucoma is a condition of the eyes that damages the optic nerves because of elevated intraocular pressure within the retina [1]. Without early diagnosis, glaucoma progresses to permanent sightlessness of the human eye and eventually becomes irredeemable [2]. It is projected that there will be 111.8 million glaucoma sufferers globally by 2040 [3], which is a grave issue for the elderly because the condition frequently affects older persons. Therefore, priority should be given to disease early- stage detection. The obstruction of the drainage canal in the eye is where the illness starts. The intraocular pressure (IOP) will rise as a result of this obstruction. IOP elevation will eventually damage optic nerve fibers, causing the retinal nerve fiber layer (RNFL) to thicken. The diagnosis of a glaucomatous eye has been made rapid and simple using computer vision and deep learning techniques.

Due to its superior removal capabilities and effective computation, ML and CNN has been the best option for image classification over the past ten years [4-6], [16-18]. AlexNet obtains ground-breaking performance on the

ImageNet image categorization at the beginning of deep learning [7, 8]. In this study, the most recent developments in Vision models using transformer architecture for glaucoma classification utilizing the whole fundus image dataset are examined.

A novel class of neural network is called Transformer based model. It mostly makes use of the self-attention mechanism to extract intrinsic characteristics [9], and it has a lot of promise for being widely applied in AI applications. The Vision Transformer (ViT) attracted a lot of scientific interest in 2020. Before outperforming SOTA CNNs in more difficult tasks like object identification and segmentation, it showed promising results by outperforming them in numerous image recognition tests.

Transformers are particularly effective structures for data that can be modeled as a series (for instance, a sentence is a sequence of words), and they were originally developed for Natural Language Processing jobs. It addresses numerous problems that recurrent neural networks and other sequential models confront. Stacks of transformer bricks make up transformers. These blocks are multilayer networks made up of basic linear layers, feed-forward networks, and self-attention layers as illustrated in Fig. 2.

To summarize our contributions in this paper

- To the best of our knowledge, very few works pertaining to the glaucoma classification problem utilizing vision transformers on fundus images are available. We are proposing Transformer learning for the classification of Glaucoma diseases.
- We have also analyzed our work on adaptive and without adaptive equalization preprocessing data set.
- We have trained and tested datasets using various optimizers.
- The empirical results can achieve an accuracy of up to 97%.

The remaining sections of the paper are arranged as follows. We'll introduce deep learning methodologies and concepts along with a few transformer learning usages in Section 2. The transformers we employed in our studies are presented in Section 3. The experiments we carried out and the outcomes are displayed in Section 4. The conclusion and future scope are followed by section 5.

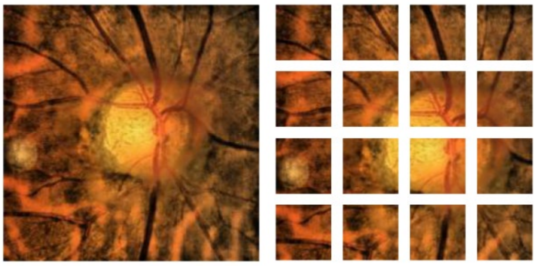
## 2. Related Work

Nowadays, image categorization has found a lot of success in both the academic and industrial worlds. A deep learning model is also used for feature classification and selection. In [10] the authors provide a technique for optimizing a vision transformer (ViT) model using a small lung X-ray dataset. They first fine-tune a ViT pre-training model on famous lung datasets as opposed to directly fine-tuning a ViT pre-trained model. Later, the small dataset is used to retrain the trained model. The precision & accuracy of the R50-B/16 ViT model is 87.57% and 88.02%, respectively. Resnet50, on the other hand, boasts accuracy & precision of 86.57% and 86.51%, respectively. In [11] the lungs image classification task, authors compare the CNN-based and the transformer-baseline model. The outcomes demonstrate that, in terms of accuracy, the suggested strategy was competitive with famous CNN algorithms. In terms of model interpretation, the suggested strategy outperforms CNN methods as well. The backbone network, high/ mid-level vision, low-level vision, and video processing are the primary areas they examine [12]. Implementing Transformers in actual device-based applications, additionally, give effective Transformers approaches. As the foundation of the transformers, the authors also take a quick look at the self-attention process in computer vision. There are various ways to further enhance visual transformers, in addition to the methods already described, such as positional encoding, normalization technique, shortcut connection, and attention removal. They observe that when CNN and transformer are combined, the performance is greater, demonstrating their complementarity through both local and global connections. The goal of fine-grained visual categorization (FGVC) is to identify an object's subclass is discussed in [13]. The FGVC task is extended to use the vision transformer, resulting in advanced performance. These transformer-based line structures, however, ultimately cause harm to the discriminative zones and have a large computational cost. To address these problems, the pooling layer was added to a transformer encoder to build a multi-stage hierarchical structure in the pooling layer-based vision transformer architecture with overlapping patches (OP-ViT). For fine-grained visual categorization, OP-ViT, a multi-phase vision transformer model is suggested. To better maintain local region information, a sliding window with a reduced step size is employed. The transformer is a multi-phase hierarchical structure that is frequently used in CNN due to the pooling layer's architecture demonstrated [13]. This makes it possible for their model to cope with the higher computing demand brought on by overlap patches. In this study, the author utilizes a light weighted module that integrates the information joint of window self-attention with the inductive bias of the Convolution Neural Network model (CNN) to produce a backbone architecture known as the Convolution Lite Transformer model (CLT) based on the ViT model. Additionally, experimental findings demonstrate that CLT outperforms the conventional CNN model in accuracy on ImageNet, achieving 79.21% accuracy with fewer parameters. Light weight deployment than the ViT model and more efficient than the conventional CNN model is presented in [14]. This work introduces a light fusion model for image classification called CLT that is built upon the junction of CNNs and Transformers. It makes use of ViT's information advantage over the world and CNN's effectiveness in local processing. Inductive bias and self-attention are combined to help CLT strike a balance between accuracy and parameter. To develop efficient glaucoma classifiers on the fundus image datasets REFUGE, RIM-ONE DL, and DRISHTI-GS, the most recent Transformer architecture-based Models used in deep learning are fine-tuned and utilizing the appropriate selection of hyperparameters. According to the results, Vision Swin Transformer performs best on the aggregated data set and can be used as a classifier for predicting glaucoma from unidentified fundus images of the Eye. Transformer performs best performance on REFUGE & RIM ONE datasets. The categorization of a fundus picture [19] utilizing an ensemble of vision transformers is investigated. Six complete fundus photos dataset that are freely available were combined into one sizable dataset for the purpose of glaucoma identification. They presented a thorough assessment of more than seven distinct basic Vision Transformers models. The need for a lot of data to effectively train vision transformer models presented the first difficulty. In the second, there were more samples from patients without glaucoma than from those who had the disease as the imbalanced dataset. Swin Transformer exhibits excellent performance both in solo models and in the top-performing ensembles, and it provides excellent snapshot ensemble results.

In existing approaches [15-18], the impact of data pre-processing on learning and result accuracy is an issue. Therefore, accurate glaucoma disease detection and data pre-processing is essential. We have addressed all issues using transformer models along with data preprocessing to increase accuracy. Additionally, our suggested model can lower computation costs in terms of resources and time which can be used in mobile devices as well.

## 3. Proposed Methodology

The ViT architecture is relatively simple, and all of its calculations may be summed up as follows: From the input picture, the first layer of ViTs extracts a predetermined number of patches.



a) Before patching b) After Patching

**Figure 1.** The Patching on IEEE Dataset [20]

A special class token vector is then added to the series of embedding vectors after the patches are projected to linear embedding. The sequence is then transferred into the transformer blocks after the vectors holding positional information have been added to the embedding and the class token. The final classification is produced by an MLP head once the class token vector has been taken from the output of the last transformer block.

A transformer block is made up of many layers. Layer Normalization is implemented at the top layer. The vital centre of ViTs' multi-head attention, which is in charge of ViTs' performance, comes next. Two arrows may be seen on the transformer block illustration in Fig. 2. These are the so-called residual skip connections. After Layer Normalization, the results of the multi-head attention are examined once again. Finally, an MLP that was initially developed using the GeLU activation function serves as the output layer.

The convolutional kernel in the case of CNNs would be an inductive bias. By eliminating their inductive biases, ViTs moved in the opposite direction of CNNs and LSTMs to become more broad architectures. Since MLPs do not alter their weights for different inputs after training, a ViT can be thought of as a generalization of MLPs. The attention weights of ViTs, on the other hand, are computed "runtime" based on the specific input.

The preprocessing phases of the proposed model are discussed as follows:

### *IEEE Dataset and Preprocessing*

This set contains 1450 fundus images with 899 glaucoma data and 551 normal data [20]. This dataset is distributed in train, test, and valid dataset. We have tested the model on test dataset which have 291 images as healthy and Glaucoma images Fig.3.

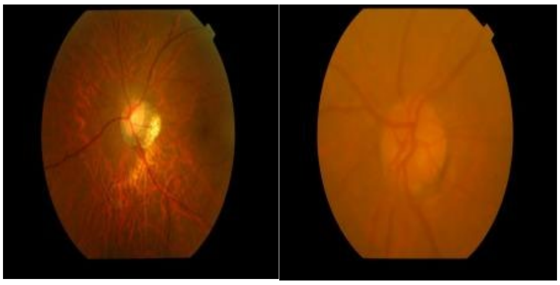


**Fig. 3a** IEEE original (Glaucoma) **Fig. 3b** IEEE original (Healthy)

The steps of preprocessing are as follows

1. Image resizing: The image will be stretched to fit your screen while being resized to 224 × 224.
2. Auto Rotate: This step enables a device's screen orientation to adjust automatically. We utilize the Portrait Mode, which automatically converts any photographs that are not in the correct rotation to portrait mode after rotating them in accordance with the auto-rotation chosen here. To display the pixels and rotate them by 90 or 180 degrees while presenting the image, it simply flips one bit to the viewer.
3. Auto Adjust Contrast: In this case, I'm utilizing adaptive equalization, which adapts the photos from the data set. Sample images after preprocessing are displayed in Fig. 4.

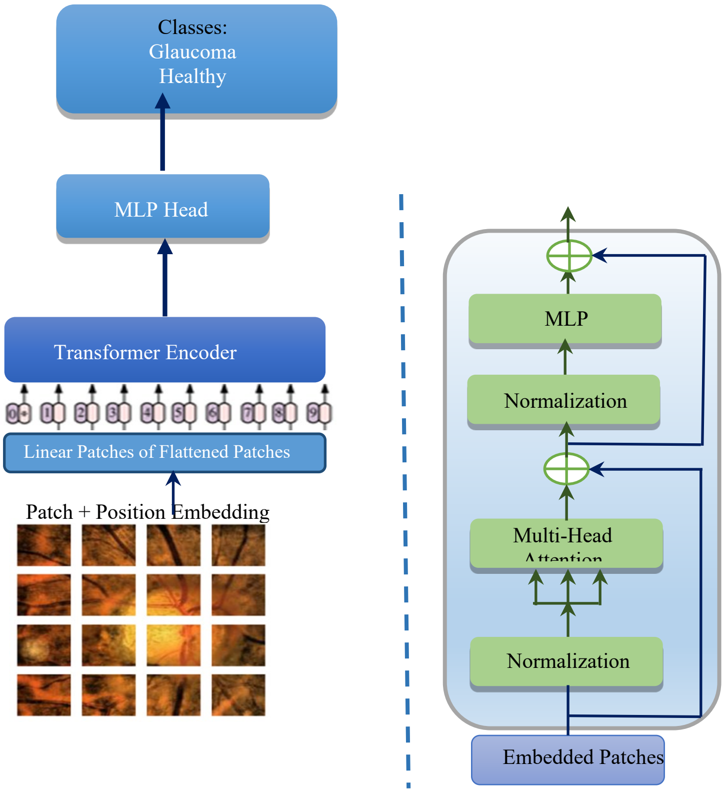


**Figure 2.** The Workflow of the Proposed Model

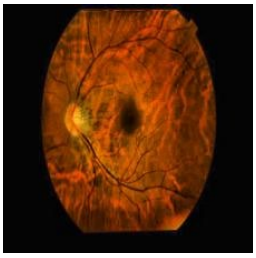


**Fig. 4a** IEEE Adaptive Equalization dataset image after (Glaucoma)

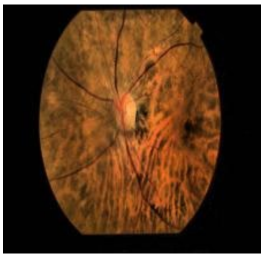
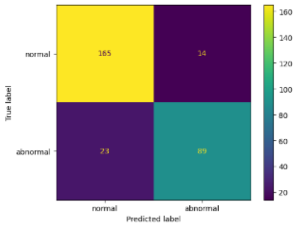


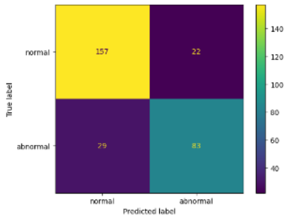
**Fig. 4b IEEE** Adaptive Equalization (Healthy)

## 4. Results and Analysis

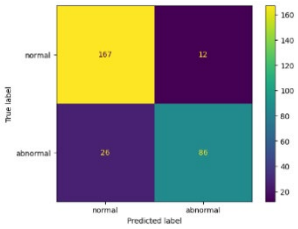
In this section, we evaluate the glaucoma classification of proposed Transformer Learning with different optimizers. We train and validate the model on the IEEE dataset [20]. Fig. 5 shows the confusion matrix of five optimizers on the IEEE dataset without adaptive. The best performance of the confusion matrix shown in Fig. 5(d) is of Nadam optimizer. It is observed that the images correctly classified as Glaucoma are 168 and images correctly classified as healthy are 90. On the other hand, the number of images misclassified as Glaucoma is 11, and healthy eye images are 22.



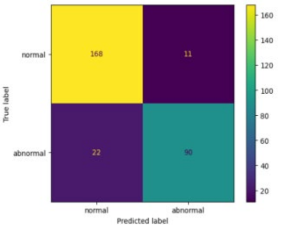
(a) Confusion Matrix of Adam Optimizer



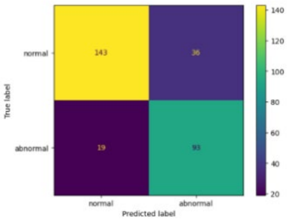
(b) Confusion Matrix of Adamax Optimizer



(c) Confusion Matrix of AdamW Optimizer



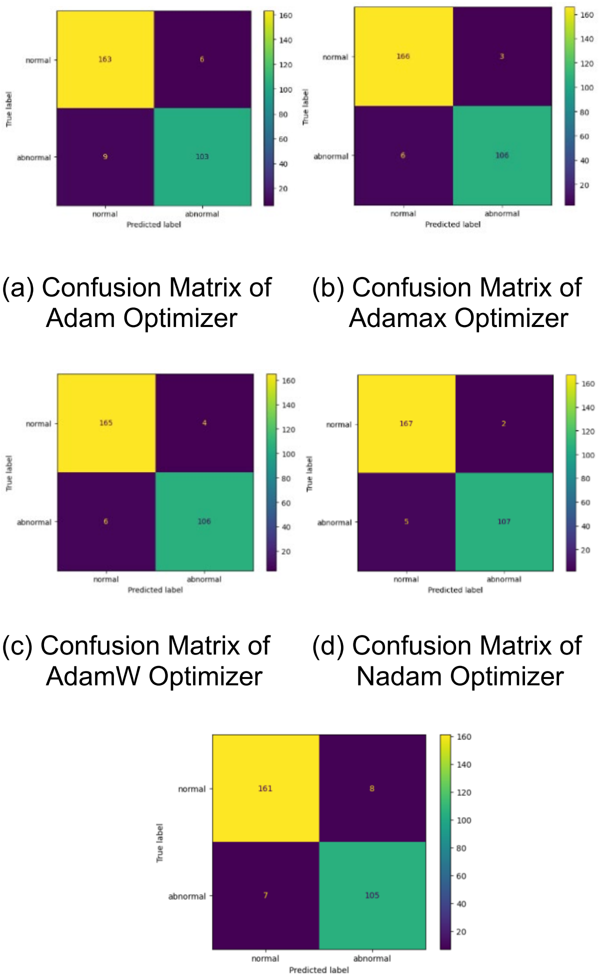
(d) Confusion Matrix of Nadam Optimizer



(e) Confusion Matrix of RMSProp Optimizer

Figure 5. The Confusion Matrix of Optimizers on IEEE Dataset (Without Adaptive equalized)

The confusion matrix of five optimizers on the IEEE dataset with adaptive is shown in Fig. 6. The Nadam optimizer performs the confusion matrix in Fig. 6(d) with the best results. It has been noted that 167 images were appropriately identified as glaucoma, while 107 images were correctly identified as healthy. On the other side, there were 2 images were incorrectly labeled as glaucoma and 5 images of healthy eyes.



(a) Confusion Matrix of Adam Optimizer

(b) Confusion Matrix of Adamax Optimizer

(c) Confusion Matrix of AdamW Optimizer

(d) Confusion Matrix of Nadam Optimizer

(e) Confusion Matrix of RMSProp Optimizer

**Figure 6.** The Confusion Matrix of Optimizers on IEEE Dataset (Adaptive Equalized)

Figs. 7 and 8. Show the performance of the proposed Transformer model using different parameters on the IEEE Dataset without and with Adaptive equalization respectively. The results of the proposed model have achieved accuracy up to 97%, however, existing work is able to achieve only up to 93% [2]. In addition to the same accuracy is achieved in [10] over chest X-ray images.

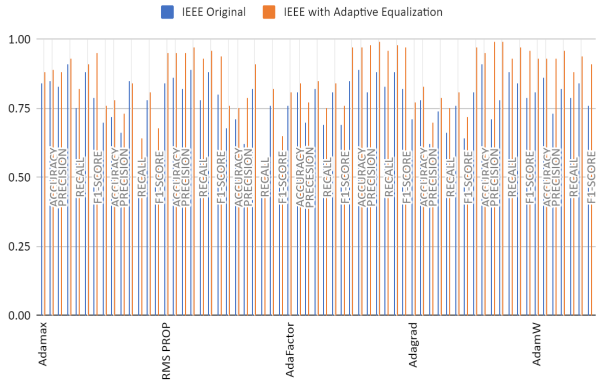


**Figure 7.** The performance of the proposed Transformer Learning model on various parameters over IEEE Dataset without and with Adaptive respectively.

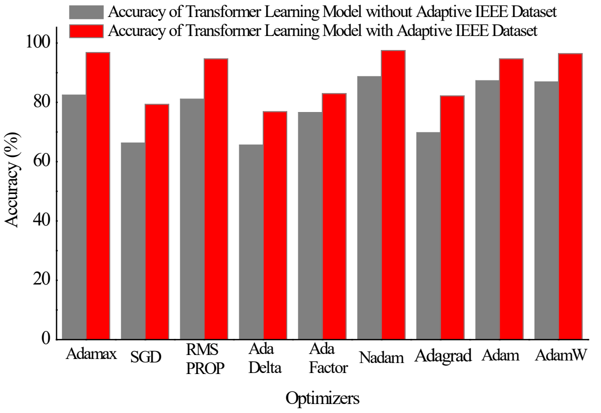
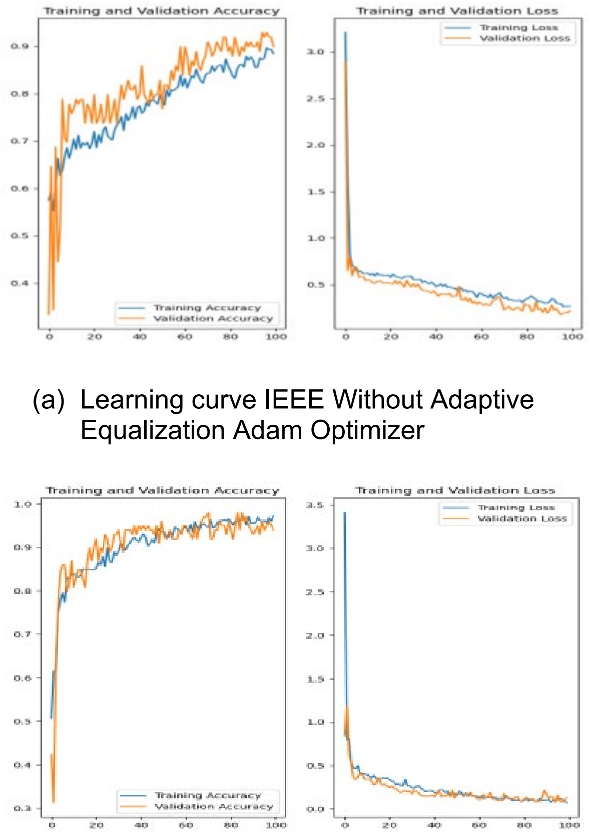


**Figure 8.** The Performance of Transformer Learning on IEEE Dataset

Figs. 9 (a) & (b) show that the training and validation accuracy and validation loss of the proposed NAdam optimizer over adaptive IEEE dataset and without adaptive. We have computed our proposed study on up to 100 epochs.



(a) Learning curve IEEE Without Adaptive Equalization Adam Optimizer

(b) Learning curve IEEE With Adaptive Equalization NAdam Optimizer

**Figure 9.** Learning Curve with the best accuracy on IEEE Dataset with and without Preprocessing as adaptive equalization

## 5. Conclusion and Future Work

In this work, transformer-based architectures for glaucoma eye classification are studied. We have analyzed the several optimizers on the dataset, and it has been presented that all optimizers provide better results in the preprocessed dataset with adaptive equalization and others as compared to the original dataset. Empirical results proved that the vision transformer works best on the Nadam optimizer, which improves the accuracy to 97.6%; however, CNN models are achieved up to 88% [7] using learning. In future work, we will explore hybrid CNN with machine learning and ViT on large private dataset.

## References

[1] Weinreb R N, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. In JAMA, 311(18), 1901–1911)
DOI: https://doi.org/10.1001/jama.2014.3192.

[2] Singh PB, Singh P, Dev H. Optimized convolutional neural network for glaucoma detection with improved Optic-Cup segmentation. Advances in Engineering Software 175(2023), 1-13 (2022)
DOI: https://doi.org/10.1016/j.advengsoft.2022.103328

[3] Tham, Y.C., Li, X., Wong, T.Y., Quigley, H. A., Aung, T., Cheng, C.Y. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology,121(11), 2081–2090 (2014)
DOI: 10.1016/j.ophtha.2014.05.013

[4] Bajpai S, Sharma K, Chaurasia BK. Intrusion Detection Framework in IoT Networks. Springer Nature Computer Science Journal, Special Issue on Machine Learning and Smart Systems, 4(350), 1-17 (2023)
DOI: https://doi.org/10.1007/s42979-023-01770-9

[5] Courtie, E., Veenith, T., Logan, A.: Retinal blood flow in critical illness and systemic disease: A Review. Annals of Intensive Care 10(152), 1-18 (2020)
DOI: https://doi.org/10.1186/s13613-020-00768-3

[6] Bajpai S, Sharma K, Chaurasia BK. Intrusion Detection System in IoT Network using ML. In NeuroQuantology 20(13), 3597-3601 (2022)
DOI: 10.14704/nq.2022.20.13.NQ88441

[7] Qummar S, Khan FG, Shah S, Khan A,Shamshirband S, Rehman Z U, Khan IA, Jadoon W (2019) A Deep Learning Ensemble Approach for Diabetic Retinopathy Detection. In *IEEE Access*, 7:150530- 150539
DOI: 10.1109/ACCESS.2019.2947484

[8] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In Communications of the ACM, 60, 84–90 (2017)
DOI:10.1145/3065386.

[9] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need, In Proc. Conf. Neural Informat. Process. Syst., 6000–6010 (2017)

[10] Nguyen MH, Quang KN. A Study of Vision Transformer for Lung Diseases Classification. In 6th International Conference on Green Technology and Sustainable Development (GTSD), 116-121 (2022)
DOI: 10.1109/GTSD54989.2022.9989100

[11] Okolo GI, Katsigiannis S. Ramzan, N. IEViT: An enhanced vision transformer architecture for chest X-ray image classification, In Computer Methods and Programs in Biomedicine, 226 (107141), 1-11 (2022).
DOI:10.1016/j.cmpb.2022.107141

[12] Han K, Wang, Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao AuC, Xu Y, Yang Z, Zhang Y, Tao D. A Survey on Vision Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45 (1), 87-110 (2023)
DOI: 10.1109/TPAMI.2022.3152247

[13] Huang Z, Du Ji-X, Zhang H-Bo. A Multi-Stage Vision Transformer for Fine-grained Image Classification. In 11th International Conference on Information Technology in Medicine and Education (ITME), 191-195 (2021)
DOI: 10.1109/ITME53901.2021.00047

[14] Han X, Wang K, Tu S, Zhou W. Image Classification Based on Convolution and Lite Transformer. 4th Internati-onal Conference on Applied Machine Learning (ICAML), 3-7 (2022)DOI: 10.1109/ICAML57167.2022.00009

[15] Mallick S, Paul J, Sengupta N, Sil J. Study of Different Transformer based Networks for Glaucoma Detection. In IEEE Region 10 Conference (TENCON), 1-6 (2022)
DOI: 10.1109/TENCON55691.2022.9977730

[16] Tripathi A, Misra A, Kumar K, Chaurasia BK.Optimized Machine Learning for classifying colorectal tissues. Springer Nature Computer Science Journal, Special Issue on Machine Learning and Smart Systems, 1-26 (2023)
DOI : 10.1007/s42979-023-01882-2

[17] Chaurasia BK, Raj H, Rathour SS,Singh PB. Transfer Learning driven Ensemble Model for Detection of Diabetic Retinopathy Disease. In Medical & Biological Engineering & Computing, Springer, 1-22 (2023)
DOI : 10.1007/s11517-023-02863-6

[18] Tripathi A, Misra A, Kumar K, Chaurasia BK. Colon Cancer classification using Machine Learning. IEEE ISCON, 1-6 (2023).
DOI: 10.1109/ISCON57294.2023.10112181

[19] Wassel M, Hamdi AM, Adly N, Torki M. Vision Transformers Based Classification for Glaucomatous Eye Condition. In 26th International Conference on Pattern Recognition (ICPR), 5082-5088 (2022)
DOI: 10.1109/TENCON55691.2022.9977730

[20] IEEE Dataset, Online Available at: https://ieee-dataport.org/documents/1450-fundus-images-899-glaucoma-data-and-551-normal-data