Breaking The Black Box: Explainable-AI (XAI) Model to Classify Glaucoma Using Transfer Learning

Amira Raina1, a) and Sabaa Ahmed Yahya Al-Galal 1, 2, b)

1Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Malaysia

2Centre for Image and Vision Computing (CIVC), Multimedia University, 63100 Cyberjaya, Malaysia

b) Corresponding author: algalalsabaa@mmu.edu.my  
a) 1211307539@student.mmu.edu.my

**Abstract.** Glaucoma is a primary factor of irreversible blindness worldwide. Early detection is critical to prepare for vision loss, although traditional diagnostic methods are slow and heavily reliant on the expertise of ophthalmologists. Recently, Artificial Intelligence (AI) and deep learning (DL), such as Convolutional Neural Network (CNN), emerged as powerful tools to automate glaucoma detection from fundus imaSges. These models autonomously extract features at a large scale in exchange for being computationally complex, making its decision-making processes opaque and difficult to trust in a high-stake clinical setting - essentially, creating a black box model. This project addresses the challenge by demonstrating Explainable AI (XAI) of a prototypical deep learning model, ProtoPNet, on a subset of the EyePACS AIROGS fundus dataset for glaucoma classification. The proposed framework uses transfer learning to compare pretrained CNNs (VGG16, VGG19, ResNet50) on ImageNet as the feature extractor in the architecture. VGG-based models, especially VGG16, had the best performance with an accuracy of 92.6% in classifying glaucoma using the case-based approach of ProtoPNet. By offering clinicians clear, interpretable insights into how the model makes its predictions, this research aims to build trust in AI-driven diagnostic tools. Ultimately, enabling earlier and more reliable detection of glaucoma in clinical practice.

# Introduction

World Health Organization 2023 statistic reveals that at least 2.2 billion people around the world have had vision impairment [1]. Among this, 7.7 million people had glaucoma, making this eye disease one of the primary causes of blindness. Glaucoma is an asymptomatic, irreversible blindness from the buildup of aqueous fluid called intraocular pressure. This internal high pressure gradually harms the delicate optic nerves vital in sending visual information from the eye to the brain. Early detection is critical to prevent and prepare for significant vision loss, although traditional diagnostic methods are time-consuming and heavily reliant on the expertise of ophthalmologists. Then, in recent years, Artificial Intelligence (AI) and deep learning (DL), such as Convolutional Neural Network (CNN), were developed to automate diagnosis from medical images; supplementing clinical assistance to the medical world.

However, the adoption of data-driven AI solutions in healthcare remains slow despite the extensive research of deep learning in the field of medical imaging. The black-box nature of deep-learning models concealing their internal workings yields an unclear justification behind the model’s diagnostical reasoning, especially for clinicians and patients without technical expertise to comprehend [2]. In response to this issue, the term ‘Explainable Artificial Intelligence’ (XAI) emerges to provide either visual, textual, or example-based techniques that convey a clear, human-understandable representation of the machine’s thinking process [3].

Nowadays, the trend relies on visual explanations, especially Grad-CAM, in medical image analysis. Ophthalmologists from University College London found the role of saliency mapping to be insufficient in providing interpretability. Stating this visual explanation is being misused to help validate accuracy due to its post-hoc nature of reasonings [4]. This issue brings to light the clinicians’ distrust towards the lack of transparency and accountability of Artificial Intelligence models.

Hence, this project aims to test the accuracy of an example-based XAI deep-learning model to improve interpretability in diagnosing ophthalmological disease. While visual-based explanations are best used to debug models and enhance their performance as post-hoc explanations, example-based techniques provide example cases called ‘prototype’ to mimic the way humans know an object because it has seen the object before. Thus, explaining the reasoning process behind a model’s choice of classification.

This work seeks to demonstrate the potential and accuracy of using concept-based reasoning models through ProtoPNet architecture in detecting glaucoma from a publicly available dataset of retinal images: EyePACS AIROGS. By prioritizing reliability and transparency in the model’s interpretability, this research attempts to foster trust between medical experts and AI to empower future human-machine collaboration in advancing ophthalmological care.

# lITERATURE review

Tariq et al. explored the use of transfer learning and traditional classifiers such as SVM and DenseNet121 for diabetic retinopathy detection on the Kaggle EyePACS dataset [5]. These models, however, operate as black-box classifiers without offering interpretability. To combat this black box issue, studies nowadays use visual-based explanations to justify confidence in the model’s accuracy. T. Shahzad et al. [6] and Velpula et al. [7]utilized both Grad-CAM and Local Interpretable Model-Agnostic Explanations (LIME) to uphold trust in their model’s predictions. Another deep learning model called ‘Diabetic Retinopathy Convolutional Neural Network’ (Dia CNN) was specifically designed for medical image classification tasks [8]. One of its key interpretability techniques is its saliency maps highlighting critical regions the model uses for classifying on top of the annotations from ophthalmologists. By using Dia CNN, Shoaib et al. achieved as high as 99.6% accuracy in detecting diabetic retinopathy from the retinal fundus images [9].

The heat maps highlighting the regions of interest (ROI) provide surface-level visualizations of the inner working of AI models. However, as mentioned previously, ophthalmologists raised concerns about their reliability due to the insufficiency of the model’s medical expertise in making such decisions [4]. For example, highlighting certain regions while disregarding other important areas. Thus, the shortcomings of these visual-based techniques can be made up by combining additional XAI techniques, such as textual-based techniques from medical expert annotations, i.e. Dia CNN. However, in the case of the lack of manpower, example-based explanation exists.

Example-based techniques, like the phrase itself, use examples from relatable data to be used as reference; mimicking human reasoning by giving past similar examples. It extends Convolutional Neural Network (CNN) with addition of making ‘prototype’ to provide human-understandable justifications for classification. In another words, this technique follows the concept of Case-Based Reasoning (CBR); a method where new problems are solved by referring to previous past cases. For instance, the model ‘Interpretable and Accurate Image-based Artificial Intelligence’ (IAIA-BL) applies case-based deep learning to classify mass lesions in digital mammography [9].

In ophthalmology, Singh et al. used prototypes to classify glaucoma through a model architecture called ‘Prototypical Variational Auto Encoder’ (ProtoVAE), achieving accuracy of more than 90% [10]. Besides ProtoVAE, another prototypical architecture called ‘Prototypical Part Network’ (ProtoPNet) was used to classify bird species [11]. While ProtoVAE focuses on understanding the distribution of data, ProtoPNet compares spatial image features. Hence, inspired by the work of Singh et al. utilizing ProtoVAE, this project adopts ProtoPNet instead to evaluate the accuracy of prototype-based reasoning using spatial comparison for glaucoma classification on fundus images.

# Research Methodology

This project uses Google Collab Pro as the primary development environment due to the cloud-based infrastructure without the need for dedicated local hardware. Collab Pro offers increased GPU availability, longer runtime durations, and higher memory limits, which were essential for training deep learning models such as ProtoPNet on high-resolution retinal image datasets.

The methodology involves setting up the environment in Google Collab Pro, initializing with data collection from Kaggle and augmentation, training and testing ProtoPNet model with 3 different feature extractor backbones (VGG16, VGG19, Resnet50; all pre-trained on ImageNet), showcasing XAI techniques visually through prototypes, and finally, evaluating the model using classification metrics to assess the model’s performance.

As for the backbones, these feature extractors were chosen based on the top 3 highest accuracy backbones from Singh et al. experiments [9]. It is noteworthy to mention the authors of ProtoPNet have thoughtfully designed the architecture to allow easy configuration by centralizing key parameters in the file settings.py. These include the base architecture (e.g., 'vgg16', 'vgg19', or 'resnet50'), prototype shape, image size, and the number of classes (set to 2 for binary classification). After training, the models were evaluated using accuracy, precision, recall, F1-score.

## Data Collection

The dataset is a subset of EyePACS AIROGS made publicly available by Riler Kiefer in Kaggle [12]. Originally, there were 113893 raw high-quality fundus images which required long downloads and large storage space. Hence, Kiefer created this subset, splitting the images into 8000 training images (containing 4000 glaucomatous named ‘RG’ and 4000 non-glaucomatous cases named ‘NRG’), 700 testing set (385 RG and 385 NRG), and 700 validations set (385 RG and 385 NRG). In total, there are 9540 pre-processed fundus images available for model training and testing.

### Initial Data Pre-Processing

The initial pre-processing of fundus images uses the CROP methodology, proposed and processed by Kiefer et al., which aims to focus on the relevant retinal regions [13]. This method involves segmenting the foreground (the retina) from the black background, identifying the center of the foreground, and then cropping the image around this center. The goal is to minimize the influence of the black background and ensure that the optic disc and surrounding structures are centrally located. Figure 1 shows the sample of the fundus pre-processed image.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

**Figure 1.** Sample fundus images from the dataset labeled (a) Original fundus image from EyePACS and (b) Processed image after applying CROP methodology which removes the black background to primarily focus on retinal features [12]. This preprocessing step was applied to all 9540 images in the dataset, which were subsequently resized to 512×512 pixels and split into training (80%), validation (10%), and testing (10%) sets

### Data Augmentation

To reduce overfitting of the ProtoPNet model, the training dataset underwent data augmentation strategy. The data augmentation pipeline used in this project follows the instruction provided by the authors of ProtoPNet. The process is defined in the img\_aug.py file, where only the dataset paths needed to be modified to integrate a new glaucoma dataset. The techniques include either random horizontal flipping, slight rotations, translations, and brightness adjustments. The loop is also removed to control the variants per image; ensuring only 2 augmentations for each fundus image to reduce storage space. Sample post-augmented data can be seen in Figure 2.

|  |  |  |
| --- | --- | --- |
|  |  |  |
| (a) | (b) | (c) |

**FIGURE 2.** Data sample after augmentation showing (a) original image, (b) horizontally flipped, (c) slightly rotated image

### Exploratory Data Analysis

The final dataset comprises 24,000 fundus images, evenly distributed between 12,000 Recordable Glaucoma (RG) cases and 12,000 No Recordable Glaucoma (NRG) cases. In total, there are 25400 images ready-to-use for deep learning applications.

## Model Training and Testing

### Model Training, Testing, and Evaluation

In the model training process, several parameters are tuned to optimize performance while managing computational resources (the GPU units in Collab), especially when encountering memory limitations during training. Furthermore, the batch size is decreased to 32 (instead of the default 80) to reduce usage of GPU memory. Due to these limited resources, the epochs have also been reduced from 300 to 50 as advised by the original authors on GitHub [14]. There will be 10 prototypes following Singh et al.’s experiment [10].

In this study, no early stopping or validation-based tuning methods were applied during training. Instead, a fixed number of 50 training epochs was used for all ProtoPNet models. This was done to ensure a fair and consistent comparison across different backbone architectures (VGG16, VGG19, and ResNet50).

By default, ProtoPNet computes several evaluation metrics after each epoch to assess both classification performance and interpretability. This includes cross-entropy loss, cluster cost, separation cost, accuracy, L1 regularization loss, and prototype distance pairwise.

# Results And discussion

The key interpretability of ProtoPNet is its justification in classifying an image through prototypes, as shown in the 10 prototypes in Figure 3, learned during the model’s training. ProtoPNet slides each prototype across the feature map of the test image, then measures similarity between the prototype and the patch at every location. The final total points, or logit values, are computed according to the formulation proven in the original ProtoPNet paper where each prototype contributes positively or negatively to the final class calculation [10]. This prototype-based reasoning allows ProtoPNet to provide a transparent explanation of why an image is classified into a certain category, enhancing the model’s trustworthiness. Figure 4 further demonstrates how ProtoPNet classifies a test image to its class.

In Figure 5, it seems that reducing the training epochs from 100 to 50 didn’t affect the model’s performance in this study. All three ProtoPNet models already reached good accuracy and converge before hitting 50 epochs. Meaning the model has learned what it can from the data. Training longer could potentially induce overfitting; hence, using 50 epochs was enough to get acceptable accuracy while saving time and computing resources.

The results in Table 1 summarize the performance across the 3 backbones where VGG-based backbones outperformed ResNet-based backbones. Among them, ProtoPNet with a VGG16 backbone had the highest accuracy at 92.60%, along with strong precision (92.68%), recall (92.60%), and F1-score (92.59%). ProtoPNet with VGG19 also performed competitively, with an accuracy of 92.08% meanwhile ResNet50 exhibited the lowest performance at 88.18%.

As comparisons, Tariq et al. reported an accuracy of 52% using an SVM and 48% with DenseNet121 on the same dataset [5]. In contrast, the proposed ProtoPNet-based framework not only enhances predictive accuracy but also provides visual prototypes to support clinical trust and understanding.

A collage of images of the eye

AI-generated content may be incorrect.

**FIGURE 3.** These 10 images illustrate the 10 prototypes, or 'examples’, learned by VGG16 from the training set images. The top 5 are non-glaucomatous type and bottom 5 are glaucomatous type. Note how the first 3 images in NRG class are the same which allows different important regions in the same image to be used

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |

**FIGURE 4.** An example of how ProtoPNet classifies a fundus image. (a) For NRG, ProtoPNet compares patches of the test image with learned prototypes and aggregates the points toward the NRG class. (b) For the RG class, ProtoPNet similarly calculates the contribution of each prototype, resulting in a lower total score. In the end, it classifies the image based on the higher total points which are in the NRG class

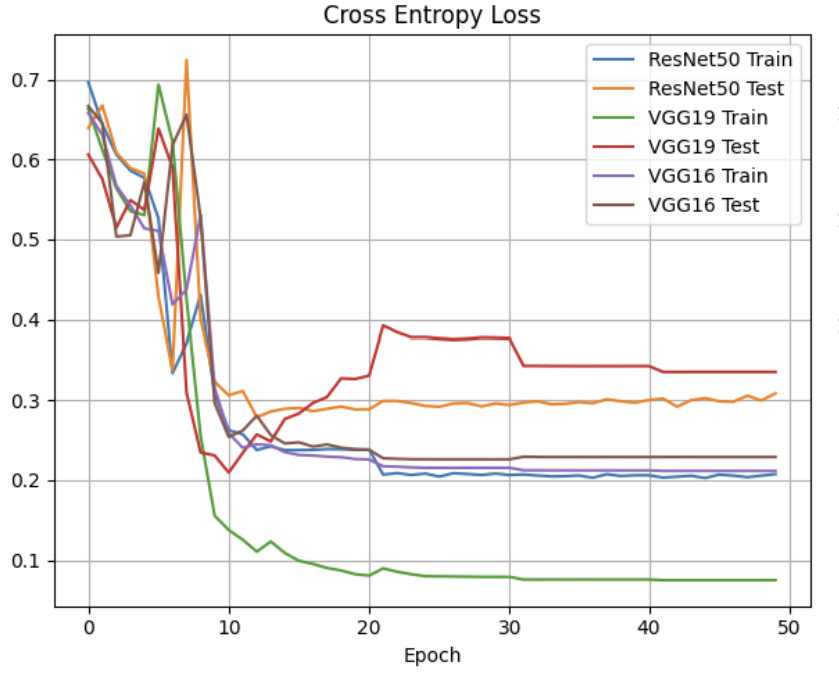
|  |  |  |
| --- | --- | --- |
| A graph with blue and orange lines  AI-generated content may be incorrect. | A graph with blue and orange lines  AI-generated content may be incorrect. | A graph of a graph  AI-generated content may be incorrect. |
| 1. ProtoPNet-VGG16 | 1. ProtoPNet-VGG19 | 1. ProtoPNet-ResNet50 |

**FIGURE 5.** The graph illustrates the accuracy progression of VGG16, VGG19, and ResNet50 over 50 training epochs. All models began to generalize well after epoch 10, as indicated by the stable and consistently high accuracy curves that flatten into straight lines

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TABLE 1.** Summarizes the performance of different feature extractors (VGG16, VGG19, ResNet50) in training ProtoPNet | | | | | |
|  | **Models** | **Accuracy** | **Precision** | **Recall** | **F1** |
|  | ProtoPNet-VGG16 | **92.60** | 92.68 | 92.60 | 92.59 |
| Proposed | ProtoPNet-VGG19 | 92.08 | 92.09 | 92.08 | 92.08 |
|  | ProtoPNet-ResNet50 | 88.18 | 88.26 | 88.18 | 88.18 |
| M. Tariq et al. [5] | DenseNet121 | 48.00 | - | - | - |
| Support Vector Machines | 52.57 | - | - | - |

Figure 6 shows all models converge early around epoch 10. Among them, VGG16 has the most stable and reliable learning pattern by having a flat and consistent cross-entropy loss throughout the 50 epochs. Although VGG19 had the lowest cross-entropy loss during training (with its loss curve dropping to nearly zero), its cross-entropy loss fluctuated during testing, suggesting signs of overfitting. Similarly, ResNet50 showed a pattern where its training loss was very low, but its cross-entropy loss increased during testing. In the end, ResNet50 recorded the lowest accuracy of them all.

Table 2 reveals VGG16-based models had better accuracy in both the proposed ProtoPNet model and Singh et al. ProtoVAE model [10]. It seems the simpler VGG’s sequential architecture is more ideal for prototypical matching, making it easier for the models to match test images with prototype images, making it the most effective backbone for this study.



**FIGURE 6.** A loss graph comparing ProtoPNet models using different backbone architectures (VGG16, VGG19, ResNet50) across training and testing phases over 50 training epochs

|  |  |  |
| --- | --- | --- |
| **TABLE 2.** shows both the VGG16-based prototypical networks, ProtoPNet and ProtoVAE, demonstrate superior accuracy in the classification task | | |
|  | **Models** | **Accuracy** |
|  | **ProtoPNet-VGG16** | **92.60** |
| Proposed Model | ProtoPNet-VGG19 | 92.08 |
|  | ProtoPNet-ResNet50 | 88.18 |
|  | **ProtoVAE-VGG16** | **95.21** |
| Singh et al. Model [10] | ProtoVAE-VGG19 | 94.52 |
|  | ProtoVAE-ResNet50 | 93.15 |

# CONCLUSION AND FUTURE WORK

This study demonstrates that ProtoPNet can achieve acceptable classification accuracy while maintaining interpretability through prototype-based reasoning. ProtoPNet interpretability comes from its ability to classify images by comparing regions of the input image with predefined prototypes, providing a transparent, step-by-step justification for its predictions. Each prototype’s point contributes to the final classification explicitly shown through these computed calculations: the network calculates the similarity score between a prototype and patches of the input image and then translates these scores into positive or negative points toward each class. This detailed calculation process allows users to trace precisely how each prototype influences the model's decision.

As shown in the experimental results, VGG-based feature extractors, particularly VGG16, outperformed ResNet-based, suggesting that simpler feature extractors may be more effective for this task since VGG’s shallower depth and straightforward structure may have contributed to more stable training and better generalization. Although the tested architectures showed similar performance within a narrow accuracy range (88%–92%), VGG16 achieved the highest accuracy of 92.60%. These results demonstrate ProtoPNet potential in domains where interpretability is critical, especially in clinical settings. Thus, breaking the black box of unknown reasoning processes in classifying such high-stake decision-making involving human lives.

To further enhance performance, future work could explore hybrid ProtoPNet models by combining multiple feature extractors, such as ensembles of VGG and ResNet architectures, leveraging the best of these algorithms. Additionally, with better computational resources, increasing the number of prototypes could help the model learn more diverse and representative features, leading to higher accuracy and better interpretability. These strategies could improve classification accuracy while preserving the interpretability that is essential for sensitive domains such as medical imaging.

# References

1. “Vision impairment and blindness,” *Blindness and Vision Impairment* (2023).
2. C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence* **1**(5), 206–215 (2019).
3. B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis,” *Medical Image Analysis* **79**, 102470 (2022).
4. C. Y. T. Wong, F. Antaki, P. Woodward-Court, A. Y. Ong, and P. A. Keane, “The role of saliency maps in enhancing ophthalmologists’ trust in artificial intelligence models,” *Asia-Pacific Journal of Ophthalmology* **13**(4), 100087 (2024).
5. M. Tariq, V. Palade, and Y. Ma, “Transfer learning based classification of diabetic retinopathy on the Kaggle EyePACS dataset,” in *Medical Imaging and Computer-Aided Diagnosis*, edited by R. Su, Y. Zhang, H. Liu, and A. F. Frangi (Springer Nature, Singapore, 2023), pp. 89–99.
6. T. Shahzad, M. Saleem, M. S. Farooq, S. Abbas, M. A. Khan, and K. Ouahada, “Developing a transparent diagnosis model for diabetic retinopathy using explainable AI,” *IEEE Access* **12**, 149700–149709 (2024).
7. V. K. Velpula, D. Sharma, L. D. Sharma, A. Roy, M. K. Bhuyan, S. Alfarhood, and M. Safran, “Glaucoma detection with explainable AI using convolutional neural networks based feature extraction and machine learning classifiers,” *IET Image Processing* **18**(13), 3827–3853 (2024).
8. M. R. Shoaib, H. M. Emara, J. Zhao, W. El-Shafai, N. F. Soliman, A. S. Mubarak, O. A. Omer, F. E. A. El-Samie, and H. Esmaiel, “Deep learning innovations in diagnosing diabetic retinopathy: The potential of transfer learning and the DiaCNN model,” *Computers in Biology and Medicine* **169**, 107834 (2024).
9. A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, and C. Rudin, “IAIA-BL: A case-based interpretable deep learning model for classification of mass lesions in digital mammography” (2021).
10. M. Singh, B. S. Vivek, J. Gubbi, and A. Pal, “Prototype-based interpretable model for glaucoma detection,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (IEEE, 2024), pp. 5056–5065.
11. C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, “This looks like that: Deep learning for interpretable image recognition,” arXiv:1806.10574 (2018).
12. R. Kiefer, “Glaucoma dataset: EyePACS-AIROGS-light-V2” (2024).
13. R. Kiefer, M. Abid, M. R. Ardali, J. Steen, and E. Amjadian, “Automated fundus image standardization using a dynamic global foreground threshold algorithm,” in *2023 8th International Conference on Image, Vision and Computing (ICIVC)* (2023), pp. 460–465.
14. “GitHub – cfchen-duke/ProtoPNet: This code package implements the prototypical part network (ProtoPNet) from the paper ‘This looks like that: deep learning for interpretable image recognition’ (to appear at NeurIPS 2019), by Chaofan Chen (Duke University), Oscar Li (Duke University), Chaofan Tao (Duke University), Alina Jade Barnett (Duke University), Jonathan Su (MIT Lincoln Laboratory), and Cynthia Rudin (Duke University),” GitHub (2019).