Prototype-based Interpretable Model for Glaucoma Detection

Mohana Singh*, B S Vivek*, Jayavardhana Gubbi, Arpan Pal
TCS Research, India
mohana.singh@tcs.com, vivekb.s31@tcs.com, jay.gubbi@tcs.com, arpan.pal@tcs.com

Abstract

Glaucoma remains one of the leading causes of irreversible blindness, its timely detection being imperative to avoiding permanent visual impairment. Deep learning methods offer a solution for early detection of Glaucoma by reducing the need for manual labor at screening stages. Hence, numerous automated methods have been proposed to assist experts in diagnosing Glaucoma from fundus images. However, the sole focus on increasing the accuracy of predictions has resulted in a lack of trust due to the black-box nature of such models. Similar sentiment across multiple high-stakes decision domains has led to a growing demand for replacing black-box models with glass-box ones. In this work, we propose an inherently explainable model that 1.) learns class-specific ***prototypes****, which capture the general characteristics or concepts of the pathology, 2.) uses the actual visualized prototypes in the decision-making process by computing the similarity between them and the query image, as a result revealing the underlying model's reasoning process, 3) is end-to-end optimizable. Moreover, the proposed approach does not require joint training of the classification models with decoders for visualization of the prototypes, simplifying the overall training process. Experimental results demonstrate that our proposed approach achieves comparable performance with its black-box counterparts and outperforms the state-of-the-art baseline, both quantitatively and qualitatively, on the benchmark RIM-ONE DL dataset.*

1. Introduction

Deep learning has revolutionized multiple research areas, with arduous tasks being accomplished in seconds. In the medical imaging community, it has emerged as a promising tool to tackle a multitude of problems. However, the adoption of deep learning-based solutions in clinical settings is slow to fruition, largely due to the black-box nature of these models. In recent years, several attempts have been made to address this issue, such as facilitating model explanation in the form of image attribution methods such as Grad-CAM [34] and Integrated Gradients [38]. However, these methods only provide a localization of the attributes sensitive to the classification models' decisions without shedding light on the models' reasoning processes. Moreover, such saliency-based posthoc visualization methods can oftentimes be misleading [1, 3, 39]. A critical element lacking in these works that could largely benefit the medical imaging community is the intuitive explainability of sensitive 'concepts'. Such high-level features or concepts may be more intuitive to a medical practitioner than a mere localization of sensitive pixels. Recently, a concept attribution method, Gifsplanation, has been proposed in [11], which diminishes the sensitive features to generate new counterfactual images. A string of such counterfactual images is then stitched together into a short video to give a visual understanding of how the sensitive attributes change with changes in the model's predictions. While motivated in the right direction, Gifsplanation [11] being a posthoc explanation technique, lacks *transparency* [15], and the visualized concepts are not explicitly used in the classification task.

In this work, we propose a *prototype*-based [26] design to make black-box models inherently interpretable and inject the models with *transparency* [15]. The proposed method provides a visualization of the *actual* prototypical images of the class, exemplifying the concepts used by the model, and employs the visualized prototypes in the classification task, making the model's reasoning process transparent. This approach aligns with the reasoning process used by domain experts of comparing cases at hand with known prototypical cases to reach conclusions [19]. The proposed model is trained in an end-to-end regime without requiring the joint training of complex components like variational autoencoders, which hinder the training process and put a constraint on the input image resolutions. Additionally, the design can be utilized with any existing classification backbone. We demonstrate the performance of the proposed method on MNIST and a real-world Glaucoma dataset. The proposed method is evaluated by comparison with baseline methods and experimental results show that

[*] Equal contribution

it achieves comparable performance to its black-box counterparts, while also making the models interpretable. The proposed model also performs better than the state-of-the-art baseline [15] in terms of both quantitative metrics as well as prototype visualizations.

The main contributions of this work are outlined as follows:

- We propose a novel prototype-based interpretable network that does not require training in conjunction with decoders.
- To the best of our knowledge, this is the first work exploring an end-to-end trainable approach to achieve both interpretability and diagnostic performance for Glaucoma detection using fundus images.
- We demonstrate the performance of our proposed method on public benchmark datasets and compare it with the state-of-the-art baselines.

2. Related Work

Previous research has focused on explaining black-box models after they have been trained (posthoc visualization) [20, 31, 34, 35, 38, 41]. However, there is a growing need to develop inherently explainable models, especially in high-stakes decision domains [32] such as medicine. One approach to achieving inherent explainability is through prototype networks [26], where the classification of data points depends on their closeness to the prototypical observations in the dataset. *Transparency* of these models is achieved when the learned prototypes are used in downstream classification tasks and these prototypes are visualizable in pixel space [15].

Some methods [4, 9, 22, 28, 33, 40] consider the prototypical observations as specific data points present in the training dataset. They visualize the prototypes using the closest images or image patches in the train set, rather than the learned prototype vectors that are *actually* used in making the predictions. This results in an approximation of model transparency. Whereas in other works [15, 26], the prototypes are not approximated to the nearest training samples but are instead decodable to the input space, resulting in an increased flexibility in capturing the dataset's characteristics. The proposed work follows the second approach to prototype learning and achieves model transparency.

In [26], an autoencoder is trained with a four-part objective loss, optimizing for both the reconstruction of images via the autoencoder, and the classification accuracy through a classifier network. The architecture has a 'prototype layer' that computes distances between the latent representations of the input and those of the prototypes. These distances are then used to compute the final prediction. Gautam *et al.* [15] improve upon [26] by replacing the simple autoencoder with a variational autoencoder, which is known to learn better latent representations, and also include an orthonormality constraint in the loss function, similar to [40], which improves the intra-class prototype diversity. A classification module is trained in tandem, which uses the similarity scores between the prototype vectors and the query image's latent representations to compute the final predictions.

The proposed work differs from both [26] and [15] in that it does not require joint training of autoencoders and classifiers. This facilitates a much simpler training procedure, increases model flexibility, and permits the use of any existing classification backbone. Additionally, the similarity score calculation is done in the classifier's latent space, as opposed to the autoencoder's latent space in [26] and [15].

3. Methodology

Given an image dataset with N data points, $\mathcal{X} : (x_i, y_i)$, for $i \in [1, .., N]$, where for each pair (x_i, y_i), $x_i \in \mathbb{R}^{H\times W\times C}$ is an image sample belonging to K possible classes, and $y_i \in [1, ..., K]$ is the corresponding ground truth class label, the primary aim is to develop inherently interpretable classification models that do not compromise on the prediction accuracy. To accomplish this task, we propose a *transparent* [15] architecture where the model explicitly learns the latent representations of prototypes corresponding to each class, which are further used to make the classification decisions. In addition, the actual learned class prototypes are visualizable in the pixel space, unlike [9] where the nearest training sample is used for visualization. These decoded prototype visualizations provide a global explanation of the concepts the model is sensitive to, along with a clear mechanism to trace the contribution of each prototype in the final decision.

3.1. Architecture

As Figure 1 shows, the proposed architecture is composed of multiple components. The first component is a conditional generative decoder, $\mathcal{D}$, the second component is a set of learnable prototype vectors, $\{\varphi_k\}$, and the third component is a classification module, f. A detailed description of the three components follows next.

The decoder, $\mathcal{D}$, generates images $\hat{x} \in \mathbb{R}^{H\times W\times C}$ conditioned on a specific class y, given latent vectors $z \in \mathbb{R}^d$, and the target class, y, i.e., $\hat{x} = \mathcal{D}(z, y)$. To obtain this decoder, a variational autoencoder [23] is used to learn the underlying distribution of the training dataset, $\mathcal{X}$, where the encoder, $\mathcal{E}$, generates the parameters, σ and μ, of the posterior distribution, instead of synthesizing a latent vector directly, i.e, $\{\sigma_i, \mu_i\} = \mathcal{E}(x_i, y_i)$. Then the reparameterization technique is used to sample the required latent vector, $z_i \sim \mathcal{N}(\sigma_i, \mu_i)$. In addition to this, a class conditioning constraint [37] is applied to learn a set of embeddings for the class labels, which are used to push the decoder to sample only from the target class. With this

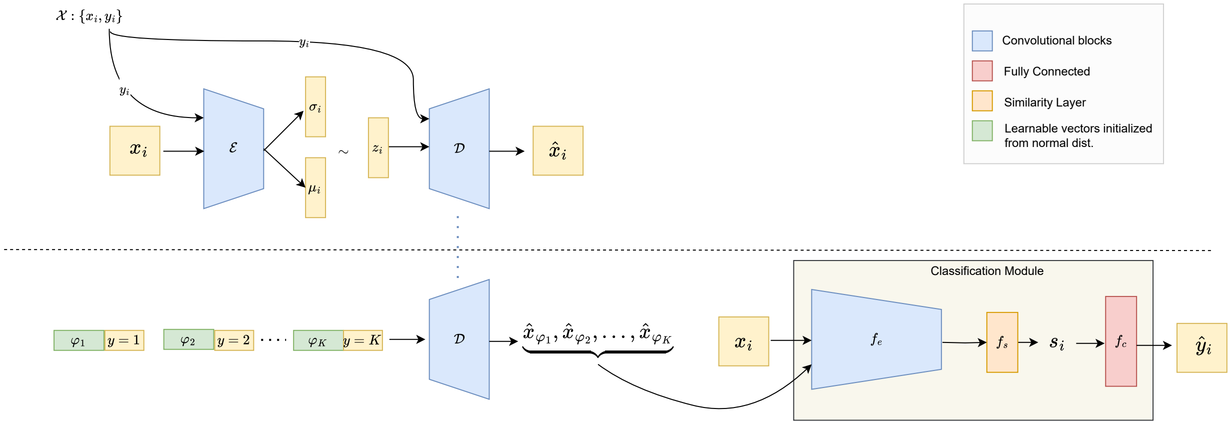

Figure 1. Proposed Architecture. At the top is a one-time trainable VAE model trained for the class-conditional synthesis of images, trained with a perceptual reconstruction loss, a patch-based discriminator loss, and the KL Divergence loss. The decoder obtained from the VAE is extracted and stitched into the proposed pipeline, and its parameters are kept frozen. The bottom part shows the proposed pipeline, with learnable prototype vectors for each class, φ_k, and the classification module with a feature extractor, f_e, a similarity layer, f_s and a fully connected layer, f_c. Refer to Section 3 for a detailed description.

architectural design, the decoder, $\mathcal{D}$, learns to reconstruct the input image samples, x_i, given these latent vectors, z_i, and their class labels, y_i, generating the output reconstruction images, $\hat{x}_i = \mathcal{D}(z_i, y_i) = \mathcal{D}(\mathcal{E}(x_i, y_i))$, where $\hat{x_i} \in \mathbb{R}^{H \times W \times C}$. The traditional objective function optimized by a VAE, the ELBO function, fosters the learning of a well-structured latent space. However, to enforce richer perceptual quality and avoid blurry reconstructions due to the smaller latent dimensions, we use a perceptual reconstruction loss [42] and a patch discriminator loss [21] in addition to the VAE's KL Divergence loss [14]. Hence, the Conditional Variational Autoencoder (CVAE) is trained by optimizing for $\mathcal{L}_{vae} = \mathcal{L}_{per} + \mathcal{L}_{disc} + \mathcal{L}_{KL}$, where $\mathcal{L}_{per}$ is the perceptual reconstruction loss, $\mathcal{L}_{disc}$ is the patch-based discriminator loss, and $\mathcal{L}_{KL}$ is the KL Divergence loss. This model is trained only once to synthesize samples of $\mathcal{X}$ and need not be retrained for changes in other components of the proposed model. The proposed model then uses the decoder of such a trained CVAE to achieve faithful reconstructions without further optimizing for the parameters of the decoder.

Secondly, the proposed architecture comprises of K learnable parameters, denoted $\phi_k \in \mathbb{R}^d$, having the same dimension as the latent vectors, z_i, of the CVAE. These parameters correspond to the latent representations of the prototypes of each of the K classes. Feeding these latent prototype vectors to the decoder, $\mathcal{D}$, produces prototype images, $\hat{x}_{\phi_k} \in \mathbb{R}^{H \times W \times C}$, which are visualized in the pixel space. The conditioning of the VAE on the classes assures that the prototypes correspond to specific classes alone, and thus does not require a "cluster" or "separation" loss as needed in [9].

The final component of the proposed architecture is a classification module composed of a feature extraction network, f_e, a similarity computation layer, f_s, and a fully connected layer, f_c. The feature extraction module, f_e, mimics the convolutional blocks prior to the fully connected layers in conventional classification networks. In contrast to [15], the proposed model can utilize any classification backbone and make existing classification models inherently explainable. The module takes input images, x_i, to extract the features, $f_e(x_i)$, and the decoded prototype images, $\hat{x}_{\phi_k}$, to extract the prototype features $f_e(\hat{x}_{\phi_k})$. This is followed by a similarity layer, f_s, where the conventional inner product operator is replaced by generalized convolution (similarity measure) [16] as done in [9]. This layer calculates the similarity of the input image features, $f_e(x_i)$, with every prototype image feature, $f_e(\hat{x}_{\phi_k})$, to obtain K similarity scores. Same as [9], the similarity function used computes the L_2 distance between the pairs and inverts the distances to obtain similarity scores. For input image x_i, the similarity score $s_i \in \mathbb{R}^K$ is obtained as follows:

$$\begin{aligned} s_i =& f_s(f_e(x_i), f_e(x_{\phi_k}))_{k=1}^{K} \\ =& \log\left(\frac{\|f_e(x_i) - f_e(x_{\phi_k})\|^2 + 1}{\|f_e(x_i) - f_e(x_{\phi_k})\|^2 + \epsilon}\right)_{k=1}^{K} \end{aligned} \tag{1}$$

where, $0 < \epsilon < 1$. Finally, the similarity layer is followed by a fully connected layer, f_c, which takes the similarity scores as input and produces output logits, convertible to probability scores of the input image belonging to each of the K classes. Hence, the final predictions, $\hat{y}_i = f_c(s_i)$ where $\hat{y}_i \in \mathbb{R}^K$, are obtained from a weighted combination of the similarity scores of the input image features and each of the class prototype features. This way the model not only uses the learned prototypes in making the final decisions but also provides the similarity scores for an understanding of the importance of the different prototypes in the classification of a particular image.

3.2. Training Regime

The proposed model is trained in an end-to-end regime, in contrast to the multi-stage training procedure of [9]. While VAEs produce good latent representations, they are also known to be harder to train and prone to blurry reconstructions, especially for higher-resolution images. Unlike [15], our model does not require a VAE to be tied to the training regime, simplifying the training procedure and tackling the restriction on the use of higher-resolution inputs. Hence, the CVAE is trained separately and only once, not requiring retraining for every classification model. As described above, the CVAE is trained by optimizing for $\mathcal{L}_{vae} = \mathcal{L}_{per} + \mathcal{L}_{disc} + \mathcal{L}_{KL}$. The resulting generative decoder is extracted with frozen parameters to aid with image reconstruction in various versions of the proposed model. Once the frozen decoder is obtained, the overall objective function to optimize is given by:

$$\begin{aligned}\mathcal{L}_{ce}(w_{f_e}, w_{f_c}, \Phi) &= \sum_{i=1}^{N} CE(y_i, \hat{y_i}) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} CE(y_i, f_c(f_s(f_e(x_i), f_e(x_{\phi_k}))))\end{aligned} \tag{2}$$

where, CE is the cross entropy loss function, w_{f_e} and w_{f_c} are the parameters of the feature extractor module, f_e, and the final fully connected layer, f_c, respectively, and $\Phi = \phi_1, \phi_2, ..., \phi_K$ are the learnable prototype vectors. The minimization of $\mathcal{L}_{ce}$ penalizes misclassification of the training samples and hence encourages inter-class separation. This loss, in addition to the class conditioning of CVAE, ensures that each of the prototypes corresponds to a particular class.

Further, our model can be extended to have any number of prototypes per class to capture intra-class diversity. In the case of such multi-prototype networks, to avoid prototype collapse and encourage disentanglement of the learned prototypes per class, an orthonormal constraint is applied on the prototypes, as described in [15, 40]. This additional penalty can be formulated as:

$$\mathcal{L}_{orth} = \sum_{k=1}^{K} \|\bar{\varphi}_k^T \bar{\varphi}_k - I_M\|_F^2 \tag{3}$$

where, M is the number of prototypes per class, $\varphi_k = \{\phi_{k,1}, \phi_{k,2}, ..., \phi_{k,M}\}$ is the set of prototype vectors corresponding to class k, and $\bar{\varphi}_k$ is a matrix whose column vectors are the differences between the prototypes corresponding to class k and their mean, i.e.,

$$\bar{\varphi}_k = \{\phi_{k,m} - \frac{1}{M}\sum_{m=1}^{M} \phi_{k,m}\} \text{ for } m = 1, ..., M,$$

while $I_M \in \mathbb{R}^{M \times M}$ is an identity matrix, and $\| \cdot \|_F$ is the Frobenius norm. Hence, the overall objective to be optimized is formulated as

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{orth}$$

4. Experiments

The proposed model is initially evaluated on a toy dataset, MNIST [24] for proof of concept. This is followed by an evaluation on a real-world dataset for Glaucoma. The following sections detail the experimental setup and present the results and analysis.

4.1. Datasets

For Glaucoma, we use the Rotterdam EyePACS AIROGS dataset (AIROGS) [12] as well as the Retinal IMage database for Optic Nerve Evaluation for Deep Learning (RIM-ONE DL) dataset [5].

AIROGS The AIROGS dataset consists of 101,442 publicly available color fundus images. Each of the samples is labeled as either Referable Glaucoma or Non-Referable Glaucoma. The original images are available as full fundus images, which are preprocessed to be cropped around the optic disk area as described in [2] since these are the main regions of interest for Glaucoma detection [13, 25]. Despite the higher volume of this dataset compared to most other publicly available Glaucoma datasets [5, 7, 13, 29, 36, 43], it is a highly imbalanced dataset with Referable Glaucoma making up a mere 3.2% of the total dataset. Additionally, the automated cropping adds noise to the dataset. Hence, AIROGS is used only in the pretraining stages of the CVAE training, giving the models a better initialization. Whereas for pre-training of the baselines and the proposed model, we create a subset of AIROGS, denoted AIROGS$_{sub}$, in which we oversample the Referrable Glaucoma samples by repetition, and undersample the Non-Referrable Glaucoma samples, only 27000 samples are kept (see Table 1).

RIM-ONE DL The RIM-ONE DL dataset is specially curated keeping in mind the deep learning paradigm and follows the specifications established in the REFUGE [29] challenge. It consists of a total of 485 retinographies, of which 313 are from healthy individuals and 172 are from Glaucoma patients. The dataset is available in two variants, one partitioned into train and test sets by hospitals and the other partitioned randomly. In this work, the random partition variant is used where the training set has 339 samples while the test set has 146 samples. Each of the samples is

Table 1. Data distribution of datasets used in this work

Dataset	Glaucoma	No Glaucoma
AIROGS [12]	3270	98172
AIROGS$_{sub}$ [12]	9508	27000
RIM-ONE DL [5]	146	313

available cropped around the optic nerve head. The training set is used for finetuning of all models and final results are reported for the test set.

4.2. Baselines

For comparison, we choose four of the best-performing models reported in [5], including VGG16 (with BatchNorm), VGG19 (with BatchNorm), ResNet50, and MobileNetv2. However, we do not use the weights provided by [5] since the results obtained using these do not match the reported numbers and are much lower. Instead, we retrain the models for fair comparison, maintaining the same training paradigm across all the baseline and proposed models. We pretrain all the baseline classification models using AIROGS$_{sub}$ for around 100 epochs and save the weights for the best configuration based on the validation loss. We then finetune the models using the RIM-ONE DL dataset for around 300 epochs. For all these classification models, the medical domain images are resized to 224 × 224. The publicly available implementation of ProtoVAE [15] is used, with the same base architecture used for the CIFAR dataset (since this is the largest resolution dataset (32 × 32) used in [15]). Unlike the black-box classification models, ProtoVAE requires smaller resolution inputs, and hence the medical dataset images are resized to 64 × 64. We also train different versions of ProtoVAE by varying the latent dimension to $16, 32, 64, 128$, and 256. These models are also trained using the same paradigm followed for the rest of the models, i.e. pretraining using AIROGS$_{sub}$ and then finetuning using the RIM-ONE DL dataset. For all the baseline models, at the pretraining stage, no other data preprocessing is performed except data normalization to the $[-1, 1]$ range. At the finetuning stage, since the dataset is small, data augmentation is applied using horizontal flip, vertical flip, random rotation $(-30, 30)$, and random resized crop with scale $(0.8, 1.2)$. For the classification loss, a weighted cross entropy is used to help with class imbalance in RIM-ONE DL.

4.3. Implementation

All the models are trained using the Pytorch framework on an A100 GPU with 30GB RAM.

Decoder $\mathcal{D}$ For the medical domain, we train a VGG-based CVAE. The image samples are resized to 64 × 64 for computational efficiency. No other preprocessing or augmentation is applied to the dataset apart from normalizing to the $[-1, 1]$ range. The latent dimension of the models is varied as $16, 32, 64$, or 128 across different experiments. The Adam optimizer is used with a learning rate of 0.0001. To encourage realistic and sharper reconstructions the model is trained using $\mathcal{L}_{vae}$, composed of a patch-based discriminator loss, a perceptual reconstruction loss, and the KLD loss, as described in Section 3. The coefficients for each component of the total loss are fixed as 1 across all our experiments. We pretrained the models using AIROGS for around 200 epochs with a batch size of 32 and the best model is saved based on the validation loss. This is followed by the finetuning of the models using RIM-ONE DL, which makes the model learn the distribution of the RIM-ONE DL dataset. Similarly for MNIST, the models are trained on the 28 × 28 input images. The CVAE is thus trained only once and the decoder is extracted. It need not be re-trained for every classification model, reducing the training overhead extensively.

Proposed Model The trained decoder, $\mathcal{D}$, is extracted and stitched into the proposed pipeline with its parameters frozen. The learnable prototype vectors are initialized from the normal distribution. For initial experiments, the number of prototypes is fixed to one per class. As described in Section 3, for the feature extractor, f_e, convolutional layers of different existing classification networks are used. The classification networks used in our experiments include VGG16, VGG19, ResNet50, and MobileNetv2, which are initialized with the weights of the trained baseline models, showcasing the ability of the proposed design to utilize any existing classification backbone. We also experiment with the encoder used in the ProtoVAE [15] baselines. For experiments on the toy dataset, the classification networks used are variations of LeNet. Again, we experiment with different versions of the model with varying prototype dimensions of $16, 32, 64$ and 128. The models are pre-trained using AIROGS$_{sub}$ for about 200 epochs with a batch size of 32 for the medical datasets, and a batch size of 128 for MNIST. For finetuning, RIM-ONE DL is used with a batch size of 4. For the medical domain, the input images are resized to 224 × 224 and fed to the classifier modules, while the output of the frozen decoder is appropriately upsampled for input to this module. The models are trained using the Adam optimizer and the learning rate is kept at 0.0001. The preprocessing and data augmentation performed are the same as those for the baseline models.

4.4. Evaluation

First, as a proof of concept, we evaluate the proposed model on the MNIST dataset. For this set of experiments, the number of prototypes chosen per class is one and the evaluation metric is accuracy. Table 2 shows the comparison with black-box models in terms of Accuracy. Figure 2 shows the prototypes learned for each class. These can be compared

Table 2. MNIST results for black-box LeNet and proposed method using LeNet backbone and a latent dimension of 64. Accuracy is in %.

Model	Test Accuracy
LeNet	99.51
Proposed(LeNet-64)	99.17

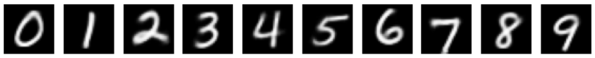

Figure 2. Prototypes learned for each class of MNIST by the proposed method using LeNet backbone and a latent dimension of 64.

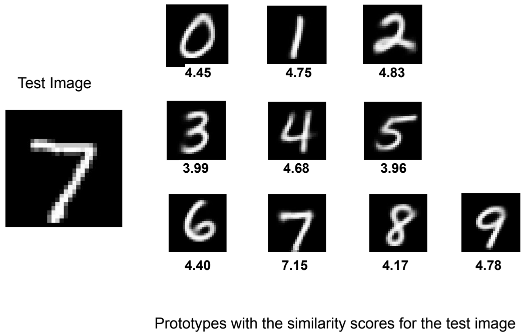

Figure 3. A test image along with the similarity scores it received for each of the class prototypes.

with the prototypes of a model trained using the decoder of a VAE without any class conditioning constraint. As shown in Figure 4, without the class conditioning constraint, the cross entropy loss alone is not enough to enforce the inter-class diversity of the prototypes and the correspondence of each prototype to a specific class. There seems to be an entangling of the prototypes of classes 2 and 3, classes 7 and 9, and classes 3, 5 and 8. Hence, the class conditioning obtained by using CVAE helps ensure that the learned prototypes correspond to a particular class. Figure 3 shows the similarity scores obtained for a correctly classified test sample. These scores show that the model uses the correct prototype for classification. This can be confirmed by visually looking at the prototype, which indeed looks like a prototypical '7'. Additionally, we are assured that the prototype was also sampled from the distribution of class '7' due to the class conditioning constraint which further strengthens the confidence on the model's predictions.

Table 3. Comparison of the proposed model with baselines for Glaucoma detection on the RIM-ONE DL [5] dataset. Sen. is the Sensitivity and Acc. is the Accuracy. All values are in %. The proposed model variations are named as **Proposed(***classification backbone-latent dimension***)**.

Model	**Acc.**	**AUC**	**Sen.**
VGG16 [5]	**95.9**	**95.9**	92.3
VGG19 [5]	95.2	94.6	94.2
Resnet50 [5]	95.2	95	92.3
MobileNetV2 [5]	94.5	94.4	90.4
ProtoVAE-16 [15]	92.47	90.7	84.6
ProtoVAE-32 [15]	91.1	91.4	92.3
ProtoVAE-64 [15]	92.47	91.6	88.5
ProtoVAE-128 [15]	93.15	93	92.3
ProtoVAE-256 [15]	91.78	91	88.5
Proposed(VGG16-64)	**95.21**	**95.85**	98.08
Proposed(VGG19-64)	94.52	94.89	96.15
Proposed(Resnet50-64)	93.15	94.68	**100**
Proposed(MobileNetv2-64)	**95.21**	94.56	92.31
Proposed(ProtoVAE-16)	92.47	91.57	88.46
Proposed(ProtoVAE-32)	93.15	92.96	92.31
Proposed(ProtoVAE-64)	93.15	94.25	98.08
Proposed(ProtoVAE-128)	94.52	94.46	94.23

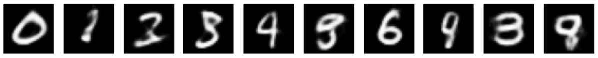

Figure 4. Prototypes learned by a version of the proposed method which uses the decoder of a VAE without any class conditioning. A test accuracy of 99.13% is obtained.

Next, we report the results for the Glaucoma dataset. Table 3 presents the results of the models for classification of Glaucoma on the RIM-ONE DL test set. The metrics used for comparison are Accuracy, AUC (Area Under the Receiver Operating Characteristic Curve), and Sensitivity. For the proposed model, we experiment across four latent dimensions, $16, 32, 64, 128$, and across different classification backbones including VGG16, VGG19, ResNet50, MobileNetV2, and the encoder backbone used for the ProtoVAE baselines. Accordingly, the different versions of the model are named Proposed(*classification backbone-latent dimension*). Similarly, ProtoVAE [15] results are generated for different versions of the model using five different latent dimensions of $16, 32, 64, 128, 256$. These are named accordingly as ProtoVAE-*latent dimension*. As shown in the table, the proposed models achieve comparable performance to their black-box counterparts in terms of all three metrics. The proposed model achieves a better sensitivity than all the blac-box classification backbones. Again, the quantitative comparison with corresponding versions of ProtoVAE shows the superior performance of the proposed models, especially in terms of the AUC and sensitivity metrics. It should be noted that while the results for all the models are reported for an input size of 224×224, the ProtoVAE model results are for an input size of 64×64.

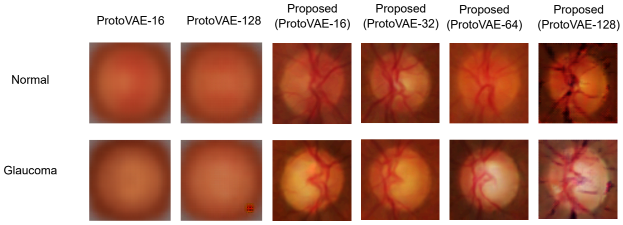

Figure 5. Visual comparison of prototypes learnt by ProtoVAE [15] and the proposed model using the same classification backbone as [15] for different latent dimensions.

Qualitative comparison in terms of the global explanations provided by the prototype images is shown in Figure 5. Across all five latent dimensions, ProtoVAE is unable to produce meaningful prototypes due to the need to optimize for both the VAE losses and the classification losses. Since all the ProtoVAE models' prototypes are extremely blurry, we only show the prototypes for two models, having latent dimensions 16 and 128. Whereas the proposed models, using the same encoder backbones as ProtoVAE, learn meaningful prototype images as shown in Figure 5. Both the CVAE in the proposed model and the VAE in ProtoVAE are trained using 64 × 64 sized inputs, however, the additional perceptual and discriminator losses help CVAE beat the blurring issue faced by ProtoVAE. Since the training regime is kept almost identical for both ProtoVAE and the proposed model, model capacities are kept similar, and a fair chance is given to ProtoVAE with smaller input dimensions, these results show how it is easier to train the proposed model compared to ProtoVAE. For good sample generation using a VAE, a balance needs to be attained between the reconstruction loss and the KL Divergence loss. Adding a classification loss into the mix and the complexity of real-world datasets complicates the training procedure significantly. Overall, the prototypes of the proposed model shown in Figure 5 exhibit an enlarged optic cup for the Glaucoma samples compared to the Normal samples, which is one of the primary factors motivating Glaucoma detection [13]. However, the prototypes of Proposed(ProtoVAE-128) show some visual artifacts around the retinal vessels despite having better quantitative performance, which should help experts qualify whether the classification backbone used is acceptable or needs to be discarded.

Further, a visual comparison of the prototypes learned by different classification backbones is shown in Figure 6. For brevity, the results are shown only for a latent dimension of 64 across all the backbones. Again, the prototypes for Glaucoma exhibit an enlarged optic cup area compared to the prototypes learned for the Normal class. Additionally, the retinal vessels can be noted as having a higher curvature in the prototypes of Glaucoma compared to those of Normal, indicating a focus on retinal vessel concepts as well.

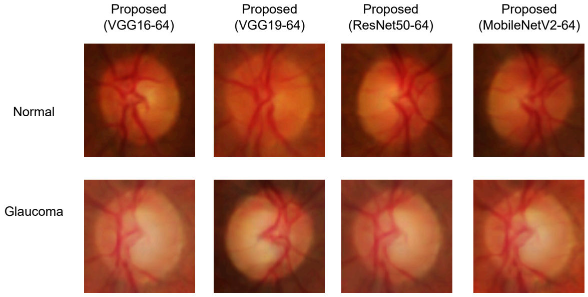

Figure 6. Visual comparison of prototypes learned by the proposed model for different classification backbones and latent dimension 64.

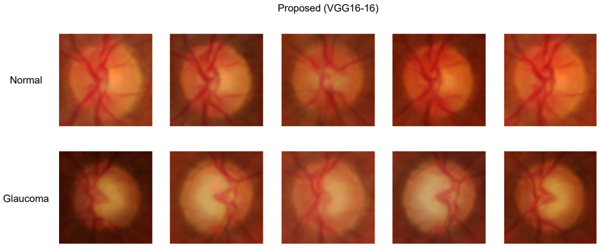

Figure 7. Multiple prototypes per class learned by the proposed model for a VGG16 classification backbone with latent dimension 16. The number of prototypes per class is 5.

Though most of the Glaucoma [8] detection literature using deep learning focuses on the cup-to-disc (C/D) ratio for Glaucoma detection [13], there are many other factors that experts use for clinical diagnoses, such as the presence of disc hemorrhage [6], thinning of the neuroretinal rim and rim that does not obey the ISNT rule [17, 30], bayoneting or the disappearance of vessels near the optic cup as they bend with a sharp turn [10] and the vanishing of the nerve fiber layer [6]. Considering that diagnosing Glaucoma is a complex process and requires examining multiple concepts, it justifies the need for more than one prototype per class to capture the diverse concepts. Figure 7 shows the prototypes for the proposed model with a VGG16 backbone and a latent dimension of 16. The models are trained to learn 5 prototypes per class. While the observation of a bigger optic cup remains consistent across the Glaucoma prototypes, there is also some variety captured in terms of color and brightness. The vessels are observed to have a higher curvature in the Glaucoma prototypes compared to the normal ones. While multiple prototypes per class can give a better idea of the concepts being focused on by the model [27], since there is no explicit effort in the architecture design to ground the various concepts, the models may not be motivated to look at the subtle signs. Hence, this motivates the need to design models that explicitly capture the concepts used by domain experts, and with such models, the

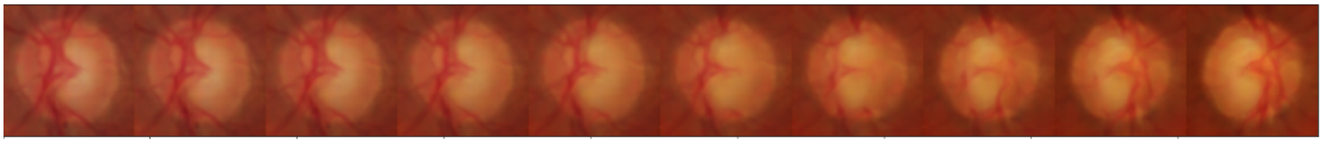

Figure 8. Interpolation between prototypes of Normal and Glaucoma learned by the proposed model with a VGG16 encoder backbone and latent dimension of 32.

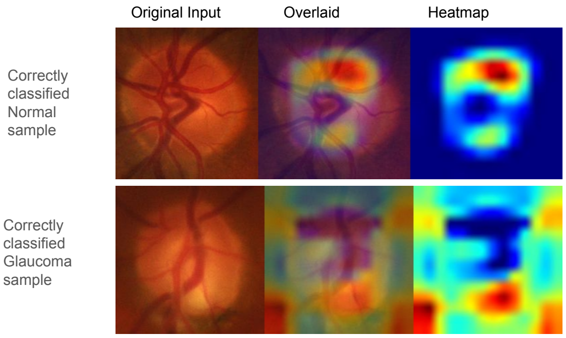
Original Input
Overlaid
Heatmap
Correctly classified Normal sample
Correctly classified Glaucoma sample

Figure 9. 2D attribution maps obtained using Grad-CAM [34] on a black-box VGG16 classification network.

proposed design can be used to further confirm if the targeted concepts are actually being used and if so, quantify the amount of contribution of each such concept.

Figure 8 shows the interpolation between prototypes of the Normal and Glaucoma classes learned by the proposed model with a VGG16 encoder backbone and a latent dimension of 32. The smooth transition discloses the model's transparent latent space. The latent movement from Normal to Glaucoma shows the thinning of vessels and gradual enlargement of the optic cup, also known as cupping.

Traditional attribution methods were used, including Grad-CAM [34] and Integrated Gradients [38], for the black-box classification models trained on the RIM-ONE DL dataset. The resulting maps for the Normal class are consistent and show that the models are sensitive to pixels at two regions in the fundus image, around the superior and inferior rim areas. Whereas for the Glaucoma class, the maps are not consistent across the images, while some show sensitivity around the optic disk, some have sensitive pixels all over the input image, while for a few images, it is sensitive to regions outside of the optic disc as well [18]. These sensitive pixel attribution maps do not help to conclude anything substantial about the model's reasoning process. Figure 9 shows the GradCAM maps for a Normal and a Glaucoma sample. We also trained the Gifsplanation [11] for retinal images and the resulting gifs indicate that the classification models rapidly change the predictions for subtle visual changes in the counterfactual images. However, since these counterfactual images are not actually used in training the models, they may not be loyal to the model's reasoning process. Additionally, the method [11] is limited by the latent representation learned by the autoencoder. Even if the autoencoder learns to reconstruct the input images, which it does for the retinal images, there is no guarantee that it learns to express the features being used by the classifier and hence the visualized concepts may not be trusted.

5. Discussion

A novel prototype-based interpretable model is proposed and its performance is demonstrated for Glaucoma detection. The learned prototypes exhibit cupping in the Glaucoma samples, which complements the hypothesis followed by most of the literature for automated Glaucoma detection using deep learning approaches. This provides a more intuitive explanation to the medical practitioner in comparison to posthoc explanations provided by traditional attribution methods. There is scope to use rich features like annotations of the optic disc and cup, retinal vessels, and other retinal landmarks to explicitly ground these concepts into the network's learning in the lines of [4]. Then the proposed method can be utilized to examine if the model focuses on the correct diagnostic concepts. Automated medical diagnosis is a complex problem, requiring analysis of multiple concepts and at varied scales. The proposed method can be extended to focus on such multi-scale features, similar to 'part-prototypes' in [9].

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 1

[2] Ahmed Al-Mahrooqi, Dmitrii Medvedev, Rand Muhtaseb, and Mohammad Yaqub. Gardnet: Robust multi-view network for glaucoma classification in color fundus images. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 152–161. Springer, 2022. 4

[3] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6): e200267, 2021. 1

[4] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao,

Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3(12):1061–1070, 2021. 2, 8

[5] Francisco José Fumero Batista, Tinguaro Diaz-Aleman, Jose Sigut, Silvia Alayon, Rafael Arnay, and Denisse Angel-Pereira. Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis and Stereology*, 39(3):161–167, 2020. 4, 5, 6

[6] Rupert RA Bourne. The optic nerve head in glaucoma. *Community Eye Health*, 19(59):44, 2006. 7

[7] Attila Budai, Rüdiger Bock, Andreas Maier, Joachim Hornegger, Georg Michelson, et al. Robust vessel segmentation in fundus images. *International journal of biomedical imaging*, 2013, 2013. 4

[8] Robert J Casson, Glyn Chidlow, John PM Wood, Jonathan G Crowston, and Ivan Goldberg. Definition of glaucoma: clinical and experimental concepts. *Clinical & experimental ophthalmology*, 40(4):341–349, 2012. 7

[9] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 2, 3, 4, 8

[10] Teresa C Chen. Spectral domain optical coherence tomography in glaucoma: qualitative and quantitative analysis of the optic nerve head and retinal nerve fiber layer (an aos thesis). *Transactions of the American Ophthalmological Society*, 107:254, 2009. 7

[11] Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. In *Medical Imaging with Deep Learning*, pages 74–104. PMLR, 2021. 1, 8

[12] Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, Adrian Galdran, Miguel Ángel González Ballester, Gustavo Carneiro, R. G. Devika, Hrishikesh Panikkasseril Sethumadhavan, Densen Puthussery, Hong Liu, Zekang Yang, Satoshi Kondo, Satoshi Kasai, Edward Wang, Ashritha Durvasula, Jónathan Heras, Miguel Ángel Zapata, Teresa Araújo, Guilherme Aresta, Hrvoje Bogunović, Mustafa Arikan, Yeong Chan Lee, Hyun Bin Cho, Yoon Ho Choi, Abdul Qayyum, Imran Razzak, Bram van Ginneken, Hans G. Lemij, and Clara I. Sánchez. Airogs: Artificial intelligence for robust glaucoma screening challenge. *IEEE Transactions on Medical Imaging*, 43(1):542–557, 2024. 4, 5

[13] Andres Diaz-Pinto, Sandra Morales, Valery Naranjo, Thomas Köhler, Jose M Mossi, and Amparo Navea. Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical engineering online*, 18:1–19, 2019. 4, 7

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[15] Srishti Gautam, Ahcene Boubekki, Stine Hansen, Suaiba Salahuddin, Robert Jenssen, Marina Höhne, and Michael Kampffmeyer. Protovae: A trustworthy self-explainable prototypical variational model. *Advances in Neural Information Processing Systems*, 35:17940–17952, 2022. 1, 2, 3, 4, 5, 6, 7

[16] Kamaledin Ghiasi-Shirazi. Generalizing the convolution operator in convolutional neural networks. *Neural Processing Letters*, 50(3):2627–2646, 2019. 3

[17] Noga Harizman, Cristiano Oliveira, Allen Chiang, Celso Tello, Michael Marmor, Robert Ritch, and Jeffrey M Liebmann. The isnt rule and differentiation of normal from glaucomatous eyes. *Archives of ophthalmology*, 124(11):1579–1583, 2006. 7

[18] Ruben Hemelings, Bart Elen, João Barbosa-Breda, Matthew B Blaschko, Patrick De Boever, and Ingeborg Stalmans. Deep learning on fundus images detects glaucoma beyond the optic disc. *Scientific Reports*, 11(1):20313, 2021. 8

[19] Alec Holt, Isabelle Bichindaritz, Rainer Schmidt, and Petra Perner. Medical applications in case-based reasoning. *The Knowledge Engineering Review*, 20(3):289–292, 2005. 1

[20] George Ioannou, Tasos Papagiannis, Thanos Tagaris, Georgios Alexandridis, and Andreas Stafylopatis. Visual interpretability analysis of deep cnns using an adaptive threshold method on diabetic retinopathy images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 480–486, 2021. 2

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[22] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. Xprotonet: diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15719–15728, 2021. 2

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 4

[25] David A Lee and Eve J Higginbotham. Glaucoma and its treatment: a review. *American journal of health-system pharmacy*, 62(7):691–699, 2005. 4

[26] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1, 2

[27] Chiyu Ma, Brandon Zhao, Chaofan Chen, and Cynthia Rudin. This looks like those: Illuminating prototypical concepts using multiple visualizations. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[28] Meike Nauta, Ron Van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recogni-

tion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14933–14943, 2021. 2

[29] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020. 4

[30] Linda Yi-Chieh Poon, David Solá-Del Valle, Angela V Turalba, Iryna A Falkenstein, Michael Horsley, Julie H Kim, Brian J Song, Hana L Takusagawa, Kaidi Wang, and Teresa C Chen. The isnt rule: how often does it apply to disc photographs and retinal nerve fiber layer measurements in the normal population? *American journal of ophthalmology*, 184:19–27, 2017. 7

[31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 2

[32] Cynthia 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. 2

[33] Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz Zieliński. Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1420–1430, 2021. 2

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 2, 8

[35] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[36] Jayanthi Sivaswamy, SR Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A Ujjwaft Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pages 53–56. IEEE, 2014. 4

[37] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 2

[38] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1, 2, 8

[39] Joseph D Viviano, Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. Saliency is a possible red herring when diagnosing poor generalization. *arXiv preprint arXiv:1910.00199*, 2019. 1

[40] Jiaqi Wang, Huafeng Liu, Xinyue Wang, and Liping Jing. Interpretable image recognition by constructing transparent embedding space. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 895–904, 2021. 2, 4

[41] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014. 2

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3

[43] Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. Origa-light: An online retinal fundus image database for glaucoma analysis and research. In *2010 Annual international conference of the IEEE engineering in medicine and biology*, pages 3065–3068. IEEE, 2010. 4