Glaucoma Precognition Based on Confocal Scanning Laser Ophthalmoscopy Images of the Optic Disc Using Convolutional Neural Network

Krati Gupta[1], Michael Goldbaum[2], and Siamak Yousefi[1,3]

[1]Department of Ophthalmology, University of Tennessee Health Science Center, Memphis, USA
[2]Department of Ophthalmology, University of California, San Diego, USA
[3]Department of Genetics, Genomics, and Informatics, University of Tennessee Health Science Center, Memphis, USA
kgupta3@uthsc.edu, mgoldbaum@ucsd.edu, siamak.yousefi@uthsc.edu

Abstract

We develop an Artificial Intelligence (AI) framework for glaucoma precognition from baseline confocal scanning laser ophthalmoscopy imaging data, using a convolutional neural network (CNN) model. The proposed framework extracts 'deep features' from convolutional layers of the CNN model, which are used as input to the ensemble learning classifier in order to identify patients that will likely convert to glaucoma after few years. The prediction model achieved area under the receiver operating characteristic curve (AUC) of 0.83 using the data from baseline visit. The model predicted the onset of glaucoma more accurately than known glaucoma risk factors, Glaucoma Probability Score (GPS) and Moorfields Regression Analysis (MRA) parameters of the Heidelberg Retinal Tomograph (HRT) software. The proposed AI construct provides a highly specific and sensitive model that can predict the onset of glaucoma from baseline HRT parameters and has the potential to provide clinicians valuable information regarding the onset of glaucoma.

1. Introduction

Glaucoma, the second leading cause of blindness worldwide [33], is a chronic eye disorder, characterized by progressive degeneration of retinal ganglion cells, which in turn, leads to changes in the optic nerve head and characteristics patterns of visual field loss [16, 23, 33]. Affected individuals are typically unaware of the subtle visual decline in early stage glaucoma; therefore, identifying those at risk of developing glaucoma is a major challenge. Due to the lack of reliable markers, clinicians may be unable to identify patients that are likely to progress and develop significant disease. In routine practice, clinicians typically assess the patient's glaucoma status through major risk factors contributing to glaucoma onset and progression including older age, elevated intra-ocular pressure (IOP), African American ethnicity, decreased central corneal thickness (CCT), and increased cup-disk ratio [9, 16, 19], along with visual field tests and evaluation of the optic nerve via ophthalmoscopy and fundus photography. Fundus photographs captured using fundus cameras, scanning laser ophthalmoscopy (SLO) or optical coherence tomography (OCT) provide optic disc images that may help reduce inter and intra-observer variability during clinical examinations of glaucomatous optic neuropathy (GON).

The Confocal Scanning Laser Ophthalmoscopy (CSLO) is a fast, non-invasive and easy-to-use imaging technique that provides a high-resolution 3-D photograph of the optic nerve and surrounding retina for precise observation and documentation of the optic nerve head for glaucoma assessment. The Heidelberg Retinal Tomography (HRT) is a confocal device that uses a diode laser at 670 nm as a light source for imaging, and provides 32 subsequent confocal images [19]. The HRT quantifies CLSO images and provides parameters such as neuro-retinal rim that can aid objective assessment of the optic nerve in glaucoma. The importance of identifying glaucoma-induced structural changes lies in the fact that structural damage may precede functional changes [21, 24], thus allowing identification of early stage glaucoma prior to significant vision loss. However, manual examination of generated fundus photographs is subjective and labor-intensive. Moreover, identifying preclinical signs of glaucoma is even more challenging than identifying clinical signs of glaucoma from fundus photographs. Therefore, development of the methods that can

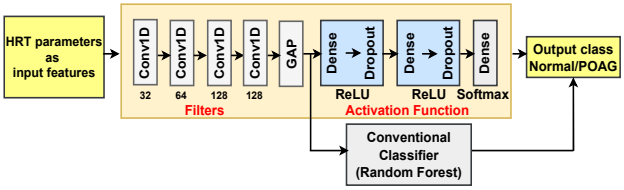


Figure 1. The CNN model along with ensemble learning to predict the onset of primary open angle glaucoma (POAG) from Confocal Scanning Laser Ophthalmoscopy (CSLO) parameters.

automatically quantify, synthesize, and analyze the subtle information existing in imaging data is critical.

To that end, we have developed AI models to recognize subtle pre-clinical signs of glaucoma from CSLO-derived parameters of optic disc collected from the first visit (baseline) and predict the disease before onset. As primary open-angle glaucoma (POAG) is the most prevalent phenotype of glaucoma, the proposed model is developed to identify the likelihood of a subject to develop POAG from baseline data. Technically, the problem of predicting POAG from only baseline data is a harder task than detecting POAG because the latter is based on clinical signs that are already manifested and obvious, while the former is based on subtle pre-clinical signs that are hard to identify.

There are numerous studies investigating the usefulness of HRT parameters for assessing and predicting glaucoma. Gordon et al. [9] used statistical cox proportional hazard models to identify the effectiveness of clinical, demographic, and other factors to predict POAG in eyes with ocular hypertension. The study was conducted on the OHTS dataset and showed baseline factors including age, vertical and horizontal cup-disc ratio, pattern standard deviation, IOP, and CCT as good predictors for development of POAG. Other prior studies also reported the predictive power of glaucoma risk factors [12, 14, 35].

Mikelberg et al. [19] discussed different aspects of HRT and proposed a discriminant analysis to analyze various HRT parameters in detecting patients with early glaucomatous visual field loss, using a stepwise discriminant analysis. Bowd et al. [4] used both baseline CSLO parameters (structural) and standard automated perimetry (SAP) visual field parameters (functional) for prediction of glaucoma progression, using relevance vector machines (RVM). The relationship between SAP and CSLO parameters were studied and investigators identified that CSLO-based parameters detect glaucomatous loss earlier than visual field-based parameters [26]. In another study, Stefano et al. [18] investigated the discriminative power of visual field examination and CSLO-HRT imaging in identifying glaucomatous changes. Their findings suggest that CSLO-HRT parameters provide less sensitivity and specificity compared to visual field measurements, using multivariate discriminant analysis and cumulative frequency distribution.

Most of the approaches in the literature develop conventional statistical models for glaucoma onset prediction [1, 34] and the results are usually compared with the Glaucoma Prediction Score (GPS) and Moorefields Regression Analysis (MRA) indices of HRT software. Zangwill et al. [39] studied the association between optic disc topography parameters and clinical factors, and showed that the HRT parameters are strongly associated with baseline stereophotographic estimates of horizontal and vertical cup-disc ratio. Therefore, HRT parameters may also be predictive of POAG development, as also highlighted by Gordon et al. [9] as well. Another related study by the same team [34] focused on examining the predictive ability of baseline GPS and MRA, along with other topographic and stereophotograph based parameters such as, cup-to-disc ratio to predict the onset of glaucoma.

Alencar et al. [1] conducted a comparative study of GPS and subjective stereophotograph assessment for prediction of glaucoma progression using a cohort of 223 patients. They concluded that GPS parameters were highly predictive of POAG progression in patients, and GPS parameters can potentially replace other stereophotograph-based parameters in predicting progression. Similarly, Salvetat et al. [28] evaluated the ability of baseline clinical, morphological, and functional factors to predict POAG conversion using 116 participants with ocular hypertension. In addition to studies discussing POAG prediction before onset, other studies have focused on identifying glaucoma after onset (diagnosis) using CSLO-HRT parameters [15, 16, 31, 38]. Here, we also focus on predicting glaucoma before onset using only baseline measurements.

2. Dataset & Pre-processing

The dataset was acquired from the Ocular Hypertension Treatment Study (OHTS), a multi-center glaucoma clinical trial aimed to delay or prevent the onset of disease in ocular hypertensive participants with moderate risk for developing POAG [9, 39]. The study included comprehensive eye exams and measurements captured by several instruments. In prospective CSLO ancillary study, HRT was used to capture 10-degree images from the optic nerve of both eyes of participants to study the glaucoma-induced structural changes [39].

We included 873 eyes of 438 participants with reliable baseline data and used 175 optic disc topographic and stereo-metric parameters that were quantified from CSLO images by the HRT software. The HRT parameters included different optic disc topographic and stereographic features such as cup area, cup volume, disc area, disc volume, rim area, rim volume, cup depth in the temporal, superior, inferior, nasal, temporal inferior, temporal superior, nasal inferior and nasal superior sectors. Out of 873 non-glaucoma

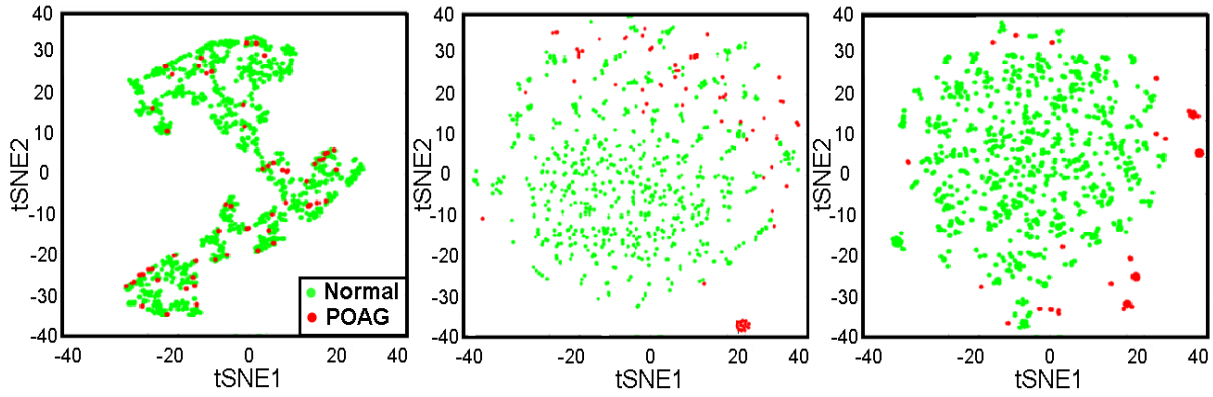


Figure 2. Features visualization using t-distributed stochastic neighbor embedding (t-SNE) (tSNE1, tSNE2: two virtual dimensions): (a) raw features (HRT parameters) of normal and POAG samples overlap significantly, (b) deep feature representation for the first dataset, and (c) deep feature representation for the second dataset.

eyes at the baseline, 59 eyes (from 44 participants) eventually developed glaucoma. Hence, this subset included 814 and 59 samples from the baseline visit of non-glaucoma eyes and eyes that eventually developed glaucoma, respectively. As the smaller number of samples included in this dataset may limit the generalizability of the CNN models (described in next section), we generated another subset. The second subset included 1621 and 115 samples (from 44 participants) from the baseline and the immediate next visit of non-glaucoma eyes and eyes that developed glaucoma, respectively. As this subset includes a greater number of samples, we will retest the models using this subset as well to assure generalizability of the CNN model.

For the sake of simplicity, we refer to the non-glaucoma group as normal (Class 1) and the "eyes that eventually developed glaucoma" as POAG (Class 2). For each sample, we generated a 1-D array of 175 HRT parameters. Similar optic nerve head parameters such as disc area at different sectors, cup volume and rim volume at different sectors were placed next to each other in the 1-D array to maximize dependency of the neighboring features.

We performed the study in accordance with the ethical standards in the Declaration of Helsinki. Initial OHTS investigators had obtained IRB approval and we signed the respective data use agreements.

3. Proposed Framework

We developed an AI model, by framing the overall problem as a binary classification task, that identifies normal eyes (class 1), and eyes that will likely convert to POAG (class 2). If successful, the AI model may predict those eyes that are at-risk of glaucoma development and future vision loss. Fig. 1 demonstrates the proposed AI framework.

3.1. Convolutional neural network (CNN) to extract deep features

A 1-D CNN model was designed to learn the intrinsic structure of HRT data and discover the latent valuable discriminating information [25]. The filter responses from several convolution layers of the neural network are used to generate deep feature representation of raw HRT parameters. HRT parameters were ordered based on their respective type and sector. For instance, all cup areas and volumes at different sectors were next to each other to mimic the image contextual characteristics in a 1-D array. Deep HRT features are hypothesized to provide higher level representations well-suited for identifying hidden subtle glaucoma-induced patterns, thus may lead to a more specific and sensitive model, for medical domain also [6, 10]. The custom CNN included four convolutional (Conv) layers (with 32, 64, 128, and 128 filters respectively) with three dense layers (Fig. 1). The kernel size in each layer is 3. We divided the data into non-overlapping training and testing subsets and trained the CNN to generate deep features from HRT parameters (raw data).

After training of the CNN model using training subset, the parameters were held fixed and the model is used to generate 128 deep representation features extracted from the CNN. These features are extracted after global average pooling (GAP) of the filter responses acquired from last convolution layer. Fig. 2 (a) shows a 2-D visualization of raw features on a t-distributed stochastic neighbor embedding (t-SNE) space. t-SNE is a technique for feature reduction and visualization of high-dimensional data in lower-dimensional space (e.g., two dimensions here) [32]. As can be seen, the raw HRT parameters of eyes with POAG are highly overlapping with raw HRT parameters of normal eyes highlighting the fact that discriminating normal from

POAG cases is challenging if raw HRT parameters are used. Fig. 2 (b) and (c) show the t-SNE representation of deep HRT features extracted from two prediction datasets. Here, classes are significantly more discriminated in t-SNE space reflecting the fact that CNN has been effective in generating more discriminative deep representations.

3.2. Class-imbalance and small sample size

Like other healthcare problems, the HRT datasets were highly imbalanced with significantly higher number of samples in the normal compared to the POAG class. Therefore, it is likely that the training task is biased towards the class with the majority of samples. To address the data skew issue, we applied Synthetic Minority Oversampling Technique (SMOTE) to augment data [5]. Essentially, SMOTE augments the minority class data by introducing synthetic samples along the line segments joining some of the k-minority class nearest neighbors in the hyperspace, and the number of augmented samples in the POAG class is equal to the samples of normal class. Moreover, to avoid overfitting, we developed the custom CNN with significantly few parameters compared to competing out-of-shelf CNN models as discussed above. We also used dropouts after each dense layer in the network, which randomly sets activations to zero during the training to further avoid over-fitting.

3.3. Classification task

We input 128 deep HRT features into a random forest (RF) classifier [2], which essentially consists of a set of uncorrelated decision-trees, to predict eyes that will likely develop POAG. RF ensemble learning scheme selects a subset of input features randomly and aggregates the response from each decision-tree classifier, operating as a committee, to come up with the final class label for test samples. The low correlation among the learners in RF classifier leads to a less cumulative error and a high generalization ability, which makes it appropriate for most applications. To provide an end-to-end model, we also used the CNN as a classifier. The filter responses extracted from the Conv layer (deep HRT features) were fed to the dense layer to make an end-to-end network. We compared the outcome of this CNN network with combined CNN and RF based classifier and other existing models.

4. Evaluation Measures and Results

4.1. CNN and Ensemble learning parameters

We applied 5-fold cross validation for accuracy evaluation and computed performance metrics i.e., specificity, sensitivity, Area Under the Receiver Operating Characteristic (AUC) curve and Matthews Correlation Coefficient (MCC). AUC provides a robust indication of the degree of separability between normal and POAG classes and MCC generates a balanced evaluation measure [3] by giving equal weights to all components of a confusion matrix, as shown:

Table 1. Baseline Demographics and Clinical Characteristics of the participants (Both in number and % (in bracket)).

	Participants at POAG endpoint	
	Yes	**No**
Race	**Number (%)**	**Number (%)**
Native-American/Alaskan	0	1 (100)
Asian	1 (20)	4 (80)
African American	11 (15)	63 (85)
Hispanic	3 (14)	18 (86)
White, non-Hispanic	29 (9)	301 (91)
Other	0	7 (100)
Gender	**Number (%)**	**Number (%)**
Male	28 (15)	154 (85)
Female	16 (6)	240 (94)
Ocular Factors	**Mean (SD)**	**Mean (SD)**
Age (Years)	58.8 (9.8)	54.4 (7.3)
IOP (mmHg)	25.4 (2.6)	24.9 (2.4)
CCT (μm)	565.3 (36.7)	577.4 (37.5)
Disc Area (mm^2)	1.87 (0.38)	1.98 (0.43)
Horizontal cup-disc ratio	0.54 (0.21)	0.49 (0.21)
Vertical cup-disc ratio	0.45 (0.21)	0.38 (0.21)

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{((TP+FP)(TP+FN)(TN+FP)(TN+FN))}} \quad (1)$$

Here, TP, TN, FP, and FN refer to the true positive, true negative, false positive, and false negative samples, respectively. We used Python Keras libraries to implement both deep features and classification tasks. For RF, we used 10 decision-tree classifiers with the maximum depth of 5. All the hyper-parameters of the classifiers are selected empirically. The entire feature extraction and classification task is done on NVIDIA TITAN RTX GPU device (64 GB RAM).

4.2. Moorfields Regression Analysis (MRA) and Glaucoma Probability Score (GPS)

Moorfields Regression Analysis (MRA) and Glaucoma Probability Score (GPS) are two algorithms integrated in the HRT instrument to improve the diagnostic characteristics of the HRT [13, 20]. Essentially, MRA and GPS focus on capturing the differences in the optic disc area during the quantitative evaluation of the rim area. For MRA estimation, a contour line is drawn around the optic nerve head (ONH) by the operator, in order to compare the overall neuro-retinal rim area, for prediction in healthy individuals. On the other hand, GPS is not dependent upon manually drawn contour, hence it estimates the participant's ONH structure, via 3-D modeling of optic disc and peripapillary retinal nerve fiber layer (RNFL). Since both these parameters are considered as important estimations regarding the structural aberrations in healthy and diseased eyes, the current study also analyzes the effectiveness of the proposed pipeline with respect to the MRA and GPS values.

Moreover, we also evaluated the cumulative probability of developing POAG in all participant's eyes, using the output of our framework, along with the MRA and GPS, via a Kaplan-Meier survival curve [27, 34]. This curve provides an estimate of the fraction of participants survival for a certain amount of time after treatment. Here, the X-axis of the curve denotes the time of the sample acquisition in the complete OHTS study, defined in months and Y-axis shows the proportion of participant's eyes that developed glaucoma.

4.3. Results

About 30% of the 1,636 OHTS participants were included in the CSLO ancillary study, with same demographic characteristics as of the overall participants of the study. The average age of the participants in the normal group was 54 years (±7.38) and the average age of the participants in the POAG group was 59 (±9.84). The average time to POAG onset for the eyes in the first dataset was 7 years and the average time to POAG onset for the eyes in the second dataset was 6 years. Table 1 shows the demographic and clinical characteristics of the first data set in our study.

The Table 2 shows the accuracy of different models using RF classifier with both raw and deep features, in terms of sensitivity, specificity, MCC, and AUC. The time required for execution of each classification experiment is also provided. Deep feature extraction takes only about 0.4 second. It is worth mentioning that the accuracy of the end-to-end CNN network was lower than the CNN with RF model, and the AUCs were 0.73 and 0.81 on the raw HRT features derived from the first and second subsets, respectively.

Table 2. Classification performance of raw and deep HRT features with random forest (RF) classifier.

Features	Cases	Sensitivity	Specificity	MCC	AUC	Time (sec)
Raw (RF)	First dataset	0.39	0.87	0.21	0.71	5
Raw (RF)	Second dataset	0.33	0.95	0.31	0.77	6
Deep (RF)	First dataset	0.62	0.98	0.64	0.83	5.5
Deep (RF)	Second dataset	0.90	0.99	0.94	0.94	6

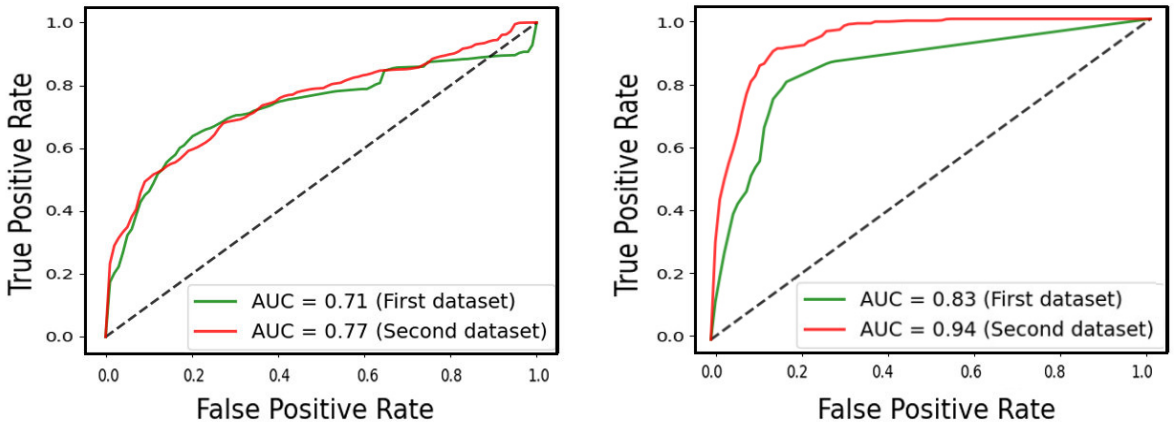
5. Discussion

Our study is an attempt to evaluate the effectiveness of HRT parameters in glaucoma precognition. The work focuses on extracting a high-level feature representation, acquired via a custom-designed CNN, followed by a random forest classifier. Importantly, the study is designed to predict the likelihood of occurrence of POAG in participants, using only baseline data, consisting of the first two visits only. Our findings confirm the effectiveness of the deep feature extraction, along with ensemble learning, to identify eyes that will likely convert to POAG.

The model applied to raw HRT parameters generated a high specificity but relatively low sensitivity (Table 2) indicating that while raw HRT parameters (features) capture the characteristics of normal eyes effectively, these parameters are unable to distinguish eyes that will eventually develop POAG. Nevertheless, this model outperforms the reported statistical models in the literature using a similar HRT dataset of the OHTS study [1, 39]. These findings motivated us to seek even more sensitive models.

Deep learning models have received extraordinary attention over the past few years and have been applied in numerous ocular conditions [11, 17, 22, 29, 30, 36, 37]. We thus developed CNN models to generate deep HRT features for identifying eyes that likely convert to POAG rather than using raw features only. In fact, CNN was able to capture the latent low- and high-level discriminating information from raw HRT parameters and provided a significant improvement in identifying those suspected eyes that eventually converted to glaucoma using two datasets (Fig. 3 (b)). We chose to use DNNs, based on their ability to learn the intrinsic structure of the data and discover the latent valuable discriminating information [25]. CNNs use both local and global information from the data, representing the interrelationships among data points. The custom CNN model was trained using data from baseline visits, in order to bypass use of any pre-trained model built using irrelevant image datasets, such as ImageNet [8]. We observed a significant improvement in the performance of the proposed AI model using the second dataset with AUC of 0.94 (95% CI: 0.92-0.95), compared to the first dataset with AUC of 0.83 (95% CI: 0.80-0.86); (Table 2 and Fig. 3). There could be several reasons for such improvement. First, there was a greater number of eyes in the second dataset compared to the first dataset (about two-fold), therefore, the AI model may have been exposed to a more representative dataset leading to improved accuracy. Second, the larger sample size in the second dataset may have prevented the AI model from becoming biased or somehow overfitted. Third, as the average time to POAG onset for the eyes in the first dataset was about 7 years and the time to POAG onset for the eyes in the second dataset was about 6 years (about one year less), this likely lead to greater manifested glaucoma-induced signs thus making prediction easier. Nevertheless, the improved performance of the AI model using the dataset with a greater number of samples may also indicate the critical role of samples size in deep learning models.

The AI construct composed of CNN and RF led to the highest accuracy compared to a RF model alone without deep feature representation (Table 2). Also, the proposed model outperformed an end-to-end artificial neural network (ANN) model. An ANN model was also developed to compare with the proposed hybrid model, which achieved AUCs of 0.65 and 0.69 based on the first and second datasets, respectively. The superiority of the proposed construct may be explained by the involvement of significantly fewer parameters compared with end-to-end ANN or end-to-end CNN models, thus providing improved optimization com-



(a) (b)

Figure 3. ROC curves of AI models. (a) represents the ROC curves of ensemble learning applied on raw HRT, (b) ROC curves of ensemble learning applied on deep HRT features.

pared with those models. Hence, the AI construct with deep HRT features along with ensemble learning achieved the highest accuracy for glaucoma precognition based on baseline HRT data.

There are some studies that have used baseline HRT data to predict PAOG [1, 4, 28, 34, 39]. Most of these models are statistical predictive models such as GPS and MRA, available in the HRT software [20]. Hence, we also compared the proposed AI model with GPS and MRA parameters in the HRT software (Table 3). We obtained sensitivity and specificity of 0.16 and 0.96 using MRA, respectively. This is in agreement with MRA outcome reported in the literature (with 0.30 sensitivity and 0.78 specificity, using a subset of OHTS dataset) [34]. Similarly, the estimated sensitivity and specificity of GPS using the first dataset were 0.22, and 0.89, which is in agreement with the previously reported findings [34]. However, the AI construct led to a sensitivity and specificity of 0.62 and 0.98 (using the first dataset as is the case in most previous studies), respectively, which significantly outperforms the conventional statistical models in the literature. The AUC of the AI construct using the first dataset was 0.83 while the highest reported AUC in the literature using same dataset is about 0.73 [1]. Hence, we showed that the proposed methodology outperforms the built-in measures of the HRT software, i.e., GPS, and MRA. We also compared our model with model proposed by Bowd et al. [4], that applied RVM classifier on raw HRT parameters for glaucoma prediction and achieved an AUC of 0.64.

We also evaluated the accuracy of several other classical machine learning models and observed that most are not well-suited for discriminating between normal and POAG eyes using HRT parameters. The best performing machine learning classifier based on our study was an RF classifier that achieved an AUC of 0.71 on the raw HRT parameters from the first dataset. In order to compare the AI construct with other state-of-the-art machine learning classifiers, deep HRT features were integrated with an XGBoost classifier [7]. XGBoost, an ensemble of gradient boosted decision trees, is one of the widely used classifiers with competing accuracy in most applications. While XGBoost performed similar to the RF classifier using raw HRT features (AUC of 0.82), its performance was lower than RF using deep HRT features (AUC of 0.83).

Table 3. Comparison of proposed framework with GPS and MRA parameters of HRT software.

S. No.	Approach	Sensitivity	Specificity
1	GPS	0.22	0.89
2	MRA	0.16	0.96
3	Proposed (deep features)	0.62	0.98

Salvetat et al [28] also highlighted the significance of baseline clinical, morphological, and functional factors in prediction of glaucoma. Similarly, we assessed the predictive power of age, CCT, and IOP only and obtained AUCs of 0.60, 0.53, 0.51, respectively. Our findings were in agreement with Gordon et al. [9], which suggests a requirement for employing more appropriate models for combining glaucoma risk factors with other imaging factors to predict glaucoma from baseline data.

We also computed the cumulative probability of developing POAG over time using the Kaplan-Meier survival analysis (Fig. 4). We observed that the predicted output of the proposed model has the most similar pattern to that of the ground truth compared to other models. In fact, the model is both sensitive and specific in identifying POAG and normal eyes, however, MRA tends to be specific but not as sensitive and GPS tends to be sensitive but highly non-specific.

We used a 5-fold non-overlapping cross validation to train the CNN model for extracting deep HRT features. To ensure the robustness of the model and the framework, an-

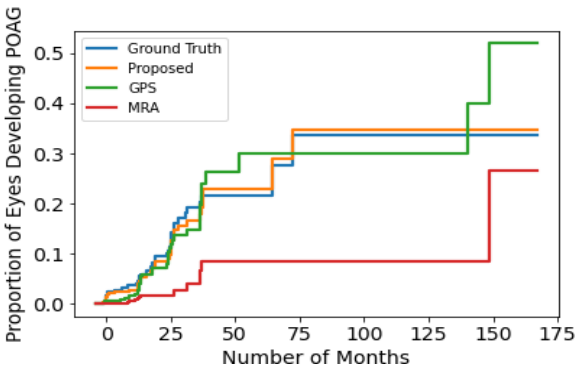


Figure 4. Kaplan-Meier plot of cumulative probability of developing POAG by the proposed model, Moorfields Regression Analysis (MRA), and Glaucoma Probability Score (GPS) along with the ground truth.

other experiment is also performed by setting 20% of samples aside (in the patient level), as a re-test data, while developing the CNN model. As such, 5-fold cross validation is used to select the best classification model. The best model is then further re-tested using the unseen re-test subset. We achieved AUCs of 0.75, and 0.86 using 20% held-out samples from the first and second datasets, respectively. The accuracy using re-test subset was comparable with that of the initial experiment. A slight drop of accuracy could be explained by the relatively smaller sample sizes that we used for training the CNN models. Thus, the AI construct is robust in prediction using unseen data as well. Additionally, from the computational time perspective (Table 2), the proposed AI construct provide the same level of computational complexity while its accuracy exceeds other models.

The clinical novelty of the proposed AI construct is its capability in synthesizing information from different topographical and stereoscopic baseline optic nerve head characteristics quantified by CSLO imaging technology. The technical novelty is that the AI construct provides a more specific and sensitive model for recognizing subtle patterns that may lead to glaucoma development. In other words, the AI construct may likely recognize the pre-clinical signs of the disease that are either not obvious to clinicians or not recognizable based on current routine clinical guidelines.

In addition to several significant aspects, this study also has some limitations. Since about 30% of all OHTS participants had been included in the ancillary study, this led to a smaller sample size compared to the overall OHTS study. While this may not be a major issue for statistical modeling, it may pose a major challenge to machine learning models. However, we generated the second dataset by including the second visit of eyes (provided the visit was before POAG onset) to generate a larger dataset for deep and machine learning models. The other limitation was that the original CSLO images were not accessible to this study, thus we were unable to investigate a full AI model on HRT images in addition to HRT parameters. Follow up studies are desirable to analyze the effectiveness of an end-to-end CNN model with input HRT images. We also used CNN on HRT parameters (features) that may not inherently correlated to each other (as is the case for 2-D images or 1-D signals) but deep features provided the highest accuracy even compared to an ANN model. While the interpretability issue remains as a concern here, this challenge exists for the simple ANN model that how features are combined.

In summary, this study proposes a hybrid AI construct using both deep learning and conventional machine learning to predict which suspected eyes with ocular hypertension will likely convert to glaucoma using CSLO-derived optic nerve head parameters. We demonstrated that the AI construct could predict those who will convert to glaucoma after about five years with relatively high accuracy. The framework could be used in glaucoma research or clinical practice in order to predict glaucoma development.

6. Conclusion

In this study, we developed an AI construct using a custom designed CNN and a random forest classifier to predict POAG before onset, using baseline HRT-derived imaging parameters of the optic nerve head. The approach extracts deep HRT features from the last convolutional layers of a previously trained CNN model and inputs the features to the random forest classifier to identify eyes that will likely develop POAG in the future. The model significantly outperformed MRA, GPS, and other proposed statistical and machine learning approaches reported in the literature. The model has the potential to be used in glaucoma clinical practice to aid physicians in identifying eyes with higher risk of developing glaucoma. Additional, independent study is desirable to validate and generalize the findings in this study.

7. Acknowledgement

The authors were funded by National Institute of Health (NIH), National Eye Institute (NEI) grants EY030142, EY031725 and in part by Research to Prevent Blindness (RPB), New York, NY. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] L. Alencar, C. Bowd, R. Weinreb, L. Zangwill, P. Sample, and F. Medeiros. Comparison of HRT-3 glaucoma probability score and subjective stereophotograph assessment for prediction of progression in glaucoma. *Investigative ophthalmology visual science*, 49:1898–906, 05 2008. 2, 5, 6

[2] G. Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13:1063–1095, Apr. 2012. 4

[3] S. Boughorbel, F. Jarray, and Md. El-Anbari. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, 12:e0177678, 06 2017. 4

[4] C. Bowd, I. Lee, M. Goldbaum, M. Balasubramanian, F. Medeiros, L. Zangwill, C. Girkin, J. Liebmann, and R. Weinreb. Predicting glaucomatous progression in glaucoma suspect eyes using relevance vector machine classifiers for combined structural and functional measurements. *Investigative ophthalmology visual science*, 53:2382–9, 03 2012. 2, 6

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 4

[6] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani. Deep features learning for medical image analysis with convolutional autoencoder neural network. *IEEE Transactions on Big Data*, pages 1–1, 2017. 3

[7] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794, 2016. 6

[8] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[9] Mae. O. Gordon, J. A. Beiser, J. D. Brandt, D. K. Heuer, E. J. Higginbotham, C. A. Johnson, J. L. Keltner, J. P. Miller, II Parrish, Richard K., M. Roy Wilson, M. A. Kass, and for the Ocular Hypertension Treatment Study Group. The Ocular Hypertension Treatment Study: Baseline Factors That Predict the Onset of Primary Open-Angle Glaucoma. *Archives of Ophthalmology*, 120(6):714–720, 06 2002. 1, 2, 6

[10] K. Gupta, A. Bhavsar, and A. K. Sao. Detecting mitotic cells in HEp-2 images as anomalies via one class classifier. *Computers in Biology and Medicine*, 111:103328, 2019. 3

[11] K. Gupta, A. Thakur, M. Goldbaum, and S. Yousefi. Glaucoma precognition: Recognizing preclinical visual functional signs of glaucoma. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4393–4401, 2020. 5

[12] F. C. Hollows and P. A. Graham. The epidemiology and control of open angle glaucoma: A population-based perspective. *Annual Review of Public Health*, 17(1):121–136, 1996. 2

[13] S. Jindal, T. Dada, S. Vishnubhatla, V. Gupta, R. Sihota, and A. Panda. Comparison of the diagnostic ability of moorfields regression analysis and glaucoma probability score using Heidelberg retinal tomograph III in eyes with primary open angle glaucoma. *Indian Journal of Ophthalmology*, 58, 11 2010. 4

[14] M.C. Leske. The epidemiology of open-angle glaucoma: a review. *American journal of epidemiology*, 118(2):166–191, 1983. 2

[15] C. Mardin, A. Peters, F. Horn, A. Jünemann, and B. Lausen. Improving glaucoma diagnosis by the combination of perimetry and HRT measurements. *Journal of glaucoma*, 15:299–305, 09 2006. 2

[16] J. Maslin, K. Mansouri, and S. Dorairaj. HRT for the diagnosis and detection of glaucoma progression. *The Open Ophthalmology Journal*, 9:58–67, 05 2015. 1, 2

[17] Felipe A. Medeiros. Deep learning in glaucoma: progress, but still lots to do. *The Lancet. Digital health*, 1(4):e151—e152, August 2019. 5

[18] S. Miglior, M. Casula, M. Guareschi, I. Marchetti, M. Iester, and N. Orzalesi. Clinical ability of heidelberg retinal tomograph examination to detect glaucomatous visual field changes. *Ophthalmology*, 108(9):1621–1627, 2001. 2

[19] F S. Mikelberg, C M. Parfitt, N V. Swindale, S L. Graham, S M. Drance, and R Gosine. Ability of the heidelberg retina tomograph to detect early glaucomatous visual field loss. *Journal of Glaucoma*, 4(4):242–247, 08 1995. 1, 2

[20] J. Moreno-Montañés, A. Antón, N. García, Loreto Mendiluce, E. Ayala, and A. Sebastián. Glaucoma probability score vs Moorfields classification in normal, ocular hypertensive, and glaucomatous eyes. *American journal of ophthalmology*, 145 2:360–368, 2008. 4, 6

[21] Sample PA. Glaucoma is present prior to its detection with standard automated perimetry: is it time to change our concepts? *Graefe's Archive for Clinical and Experimental Ophthalmology*, 241(3):168–169, 2003. 1

[22] S. Phene, R. C. Dunn, N. Hammel, Y. Liu, J. Krause, N. Kitade, M. Schaekermann, R. Sayres, D. J. Wu, A. Bora, C. Semturs, A. Misra, A. Huang, A. Spitze, F. Medeiros, A. Maa, M. Gandhi, G. Corrado, L. Peng, and D. R. Webster. Deep learning and glaucoma specialists: The relative importance of optic disc features to predict glaucoma referral in fundus photographs. *Ophthalmology*, 126:1627–1639, 12 2019. 5

[23] Harry A Quigley. Glaucoma. *The Lancet*, 377:1367–1377, 03 2011. 1

[24] Harry A. Quigley, Gregory R. Dunkelberger, and W. Richard Green. Retinal ganglion cell atrophy correlated with automated perimetry in human eyes with glaucoma. *American Journal of Ophthalmology*, 107(5):453–464, 1989. 1

[25] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 512–519, 2014. 3, 5

[26] N. J. Reus and H. G. Lemij. Relationships between standard automated perimetry, HRT confocal scanning laser ophthalmoscopy, and GDx VCC scanning laser polarimetry. *Investigative ophthalmology & visual science*, 46(11):4182–4188, 2005. 2

[27] J. T. Rich, J. G. Neely, R. C. Paniello, Courtney C.J. Voelker, B. Nussenbaum, and E. W. Wang. A practical guide to understanding Kaplan-Meier curves. *Otolaryngology - Head and Neck Surgery*, 143(3):331–336, 2010. 5

[28] M. Salvetat, M. Zeppieri, C. Tosoni, and P. Brusini. Baseline factors predicting the risk of conversion from ocular hypertension to primary open-angle glaucoma during a 10-year follow-up. *Eye (London, England)*, 30:784–795, 05 2016. 2, 6

[29] A. Thakur, M. Goldbaum, and S. Yousefi. Convex representations using deep archetypal analysis for predicting glaucoma. *IEEE Journal of Translational Engineering in Health and Medicine*, 8:1–7, 2020. 5

[30] A. Thakur, M. Goldbaum, and S. Yousefi. Predicting glaucoma before onset using deep learning. *Ophthalmology Glaucoma*, 3(4):262–268, 2020. 5

[31] H. Uchida, H. Saito, A. Iwase, A. Tomidokoro, M. Araie, T. Yamamoto, M. Shirakashi, K. Yaoeda, and Y. Ohno. Improvement of glaucoma diagnosis with heidelberg retina tomography (HRT) in japanese myopic eyes. *Investigative Ophthalmology Visual Science*, 48(13):3316, 2007. 2

[32] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008. 3

[33] R. Weinreb, T. Aung, and F. Medeiros. The pathophysiology and treatment of glaucoma a review. *JAMA : the journal of the American Medical Association*, 311:1901–11, 05 2014. 1

[34] R. N. Weinreb, L. M. Zangwill, S. Jain, L.M. Becerra, K. Dirkes, J.R. Piltz-Seymour, G.A. Cioffi, G.L. Trick, A.L. Coleman, J.D. Brandt, J.M. Liebmann, M.O. Gordon, and M.A. Kass. Predicting the onset of glaucoma: the confocal scanning laser ophthalmoscopy ancillary study to the ocular hypertension treatment study. *Ophthalmology*, 117(9), 2010. 2, 5, 6

[35] M. Wilson. Epidemiology of chronic open-angle glaucoma. *The Glaucomas*, 2:753–768, 1996. 2

[36] S. Yousefi, Louis R. Pasquale, and Michael V. Boland. Artificial intelligence and glaucoma: Illuminating the black box. *Ophthalmology. Glaucoma*, 3(5):311–313, Sept. 2020. 5

[37] S. Yousefi, H. Takahashi, T. Hayashi, H. Tampo, S. Inoda, Y. Arai, H. Tabuchi, and P. Asbell. Predicting the likelihood of need for future keratoplasty intervention using artificial intelligence. *The ocular surface*, 18, 2020. 5

[38] L. M. Zangwill, C. Bowd, C. C. Berry, J. Williams, E. Z. Blumenthal, C. A. Sánchez-Galeana, C. Vasile, and R. N. Weinreb. Discriminating Between Normal and Glaucomatous Eyes Using the Heidelberg Retina Tomograph, GDx Nerve Fiber Analyzer, and Optical Coherence Tomograph. *Archives of Ophthalmology*, 119(7):985–993, 07 2001. 2

[39] L. M. Zangwill, R. N. Weinreb, C. Berry, A. Smith, K. Dirkes, J. M. Liebmann, J. Brandt, G. Trick, G. Cioffi, A. Coleman, et al. The confocal scanning laser ophthalmoscopy ancillary study to the ocular hypertension treatment study: study design and baseline factors. *American journal of ophthalmology*, 137(2):219–227, 2004. 2, 5, 6