RESEARCH ARTICLE

A feature agnostic approach for glaucoma detection in OCT volumes

Stefan Maetschke[1]*, Bhavna Antony[1], Hiroshi Ishikawa[2], Gadi Wollstein[2], Joel Schuman[2], Rahil Garnavi[1]

1 IBM Research Australia, Melbourne, VIC, Australia, **2** NYU Langone Eye Center, New York University School of Medicine, New York, NY, United States of America

* stefanrm@au1.ibm.com

Citation: Maetschke S, Antony B, Ishikawa H, Wollstein G, Schuman J, Garnavi R (2019) A feature agnostic approach for glaucoma detection in OCT volumes. PLoS ONE 14(7): e0219126. https://doi.org/10.1371/journal.pone.0219126

Editor: Ireneusz Grulkowski, Nicolaus Copernicus University, POLAND

Received: November 12, 2018

Accepted: June 17, 2019

Published: July 1, 2019

Data Availability Statement: The data underlying this study have been uploaded to Zenodo and are available using the following DOI: 10.5281/zenodo.1481223.

Funding: IBM Research Australia provided support in the form of salaries for authors S.M., B.A, and R. G., but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Abstract

Optical coherence tomography (OCT) based measurements of retinal layer thickness, such as the retinal nerve fibre layer (RNFL) and the ganglion cell with inner plexiform layer (GCIPL) are commonly employed for the diagnosis and monitoring of glaucoma. Previously, machine learning techniques have relied on segmentation-based imaging features such as the peripapillary RNFL thickness and the cup-to-disc ratio. Here, we propose a deep learning technique that classifies eyes as healthy or glaucomatous directly from raw, unsegmented OCT volumes of the optic nerve head (ONH) using a 3D Convolutional Neural Network (CNN). We compared the accuracy of this technique with various feature-based machine learning algorithms and demonstrated the superiority of the proposed deep learning based method. Logistic regression was found to be the best performing classical machine learning technique with an AUC of 0.89. In direct comparison, the deep learning approach achieved a substantially higher AUC of 0.94 with the additional advantage of providing insight into which regions of an OCT volume are important for glaucoma detection. Computing Class Activation Maps (CAM), we found that the CNN identified neuroretinal rim and optic disc cupping as well as the lamina cribrosa (LC) and its surrounding areas as the regions significantly associated with the glaucoma classification. These regions anatomically correspond to the well established and commonly used clinical markers for glaucoma diagnosis such as increased cup volume, cup diameter, and neuroretinal rim thinning at the superior and inferior segments.

Introduction

Glaucoma is a chronic degenerative disease that affects the optic nerve and is one of the leading causes of blindness worldwide. It is characterized by changes to the optic disc, where the neuroretinal rim of the nerve becomes progressively thinner. While the disease is diagnosed using a variety of tests (including pachymetry, tonometry, and visual field tests [1]), imaging techniques such as fundus photography and optical coherence tomography (OCT) have begun to find widespread use in the diagnosis and management of glaucoma.

Competing interests: Authors S.M., B.A, and R.G. are employed by IBM Research Australia. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

OCT [2] is a non-invasive imaging modality using low coherence interferometry to generate high-resolution images of the retina in 3-D. Additionally, this modality allows for the quantification of various retinal structures. In glaucoma, the retinal nerve fiber layer (RNFL) and combined ganglion cell with inner plexiform layer (GCIPL) have been found to be clinically useful biomarkers of glaucoma and begin to thin significantly as the disease progresses [3, 4]. Recently machine learning methods have been employed to automatically detect glaucoma. These methods can be grouped into two categories: classical machine learning applied to features extracted from segmented OCT volumes such as k-Nearest Neighbor, Support Vector Machines, Random Forests and others [5], and deep learning methods such as Convolutional Neural Networks (CNN). Classical machine learning techniques rely on established features such as the peripapillary RNFL thickness and macular GCIPL thickness to differentiate between healthy and glaucomatous eyes. Thus, such techniques require the segmentation and quantification of the relevant retinal structures. CNNs, on the other hand, can directly operate on OCT volumes and are feature-agnostic in the sense that no human-designed disease markers are needed. CNNs have been successfully utilized for a variety of computer vision problems such as natural image classification [6, 7], and offer a powerful, alternative approach for the identification of glaucoma from OCT data.

Early work by Huang et al. [8] extracted 25 features such as average RNFL thickness, 4 quadrants, 12 clock hours, vertical rim area, horizontal rim area, disc area, cup area, rim area, cup-to-disc area ratio, cup-to-disc horizontal ratio and cup-to-disc vertical ratio extracted from Stratus OCT scans. The data set was composed of 89 patients with glaucoma and 100 health patients. Classical methods such as Linear discriminant analysis, Mahalanobis distance, and Artificial neural network were employed to identify glaucoma patients. The highest AUC of 0.991 was achieved by Mahalanobis distance in combination with Principal Component Analysis.

Silva et al. [9] trained 10 classical machine learning methods on 20 features such as average RNFL thickness, 4 quadrants, 12 clock hours and visual field test parameters—mean deviation (MD), pattern standard deviation (PSD), glaucoma hemifield test (GHT), extracted from a dataset composed of 62 glaucoma patients and 48 healthy individuals. The highest AUC of 0.946 was obtained by a Random Forest [10] classifier. It is noteworthy that a single feature (PSD) achieved an AUC of 0.915; not significantly (p = 0.37) different from the top AUC of 0.946 based on the complete set of features.

Kim et al. [11] conducted a similar experiment in a larger cohort of 297 glaucomatous eyes and 202 healthy eyes. Seven extracted features such as age, Intraocular pressure (IOP), mean RNFL thickness, corneal thickness, MD, Glaucoma Hemifield Test (GHT) numbers and PSD were used to train four machine learning algorithms (C5.0, Random Forest (RF), Support Vector Machine (SVM) and k-Nearest Neighbor (KNN)) to detect glaucoma. The highest AUC of 0.979 was achieved with RF and C5.0.

While these approaches produced high AUC values, the use of image-based features depends on the accurate segmentation of OCT layers, which is often difficult in advanced glaucoma cases, low quality scans and with co-existing retinal pathologies such as diabetic retinopathy (DR) or age-related macular degeneration (AMD). Furthermore, the use of human-selected disease markers potentially limits the classification accuracy achievable.

Muhammad et al. [12] employed a CNN, utilizing transfer learning based on AlexNet [6] and a Random Forest classifier trained on the features extracted by the CNN to discriminate between 45 healthy eyes and 57 eyes diagnosed with open-angle glaucoma. This method, like the previous approaches, relied on features such as the RNFL and GCIPL thickness extracted from wide-field swept-source OCT scans and furthermore included thickness probability maps. The latter are derived from the thickness distribution of a population of healthy subjects and therefore contain information beyond mere scans or individual patients. The highest AUC score of 0.979 was achieved using RNFL thickness probability maps as input feature.

In this work, we explore CNNs for the detection of glaucomatous eyes directly from unprocessed OCT volumes, thus, by-passing the segmentation steps required to extract features (such as retinal layer thicknesses, rim volume, etc.). The method utilizes optic nerve head (ONH) centered OCT scans only and does not rely on visual field tests or statistical information of healthy subjects such as thickness probability profiles. We compare the classification accuracy of this CNN with classical machine learning methods trained on traditional segmentation-based features extracted from the same dataset of ONH scans.

Material and methods

This study was an observational study that was conducted in accordance with the tenets of the Declaration of Helsinki and the Healthy Insurance Portability and Accountability Act. The Institutional Review Board of New York University and the University of Pittsburgh approved the study, and all subjects gave written consent before participation.

In the following we will distinguish between two approaches: the *feature-based* approach, where machine learning algorithms are trained on established, segmentation-based features extracted from segmented OCT volumes, and the *feature-agnostic* approach, where a CNN is directly trained on raw OCT volumes without the need of segmentation and/or feature selection.

Performance metric

We measured the classification accuracy of the methods based on the Area under the Receiver Operator Characteristic (AUC) curve, which is defined as

$$AUC = \frac{1}{2}\sum_{k=1}^{n}(X_k - X_{k-1})(Y_k + Y_{k-1}) \quad (1)$$

where X_k is the false positive rate and Y_k is the true positive rate for the k-th output in the ranked list of n confidence scores generated by the classifier. AUCs are reported for the validation and the test data.

Data

OCT scans centered on the ONH were acquired from 624 patients on a Cirrus SD-OCT Scanner (Zeiss, Dublin, CA, USA). The scans had physical dimensions of 6x6x2 mm with a corresponding size of 200x200x1024 voxels per volume. Scans with signal strength less than 7 were discarded, resulting in a total of 1110 scans for the experiments. The scans were kept in their original laterality (no flipping of left into right eye). 263 of the 1110 scans were diagnosed as healthy and 847 with primary open angle glaucoma (POAG). Glaucomatous eyes were defined as those with glaucomatous visual field defects (at least 2 consecutive abnormal test results).

Demographical background such as gender and race distribution, and mean values with standard deviations for patient's age, Intraocular Pressure (IOP), Mean Field Defects (MD) and Glaucoma Hemifield Test (GHT) [13] results are provided in Table 1. Note that for some patients demographic data was incomplete and aggregate numbers therefore do not necessarily add up to the data set size. Statistically significant differences ($p < 0.0001$) between the distribution of healthy and patients diagnosed with POAG were found for age, IOP, MD and GHT.

The data set was split into 888 training samples, 112 validation samples and 110 test samples (80%, 10%, 10%). It was ensured that eyes belonging to the same patient were not split across

Table 1. Demographic data: Gender and race distribution, and mean values with standard deviations and ranges for age, IOP, MD and GHT.

	Healthy	POAG
#Female	88	217
#Male	49	215
#White	101	318
#Black	30	154
#Asian	5	12
Age	54.1±15.3 [22.1-88.9]	64.3±12.5 [25.2-93.8]
IOP	13.5±2.4 [9-23]	16.7±5.8 [2-51]
MD	-0.8±1.7 [-9.9-2.8]	-6.8±8.1 [-32.9-2.17]
GHT	1.6±1.0 [1-6]	2.4±0.9 [1-6]

https://doi.org/10.1371/journal.pone.0219126.t001

folds. We performed 5-fold cross-validation and the averaged numbers of healthy and eyes with POAG within these folds are shown in Table 2.

Feature based approach

For the feature-based approach we used a set of 22 measurements computed by the Cirrus OCT scanner. Specifically, for each ONH scan we collected peripapillary RNFL thickness at 12 clock-hours, peripapillary RNFL thickness in the four quadrants, average RNFL thickness, rim area, disc area, average cup-to-disc ratio, vertical cup-to-disc ratio and cup volume [11]. All features were normalized by subtracting the features mean and scaling to unit variance. Normalization parameters were estimated on the training data only and then applied to training, validation and test data. No further pre-processing steps were performed. All features were real valued and contained no missing values.

We then trained the following machine learning algorithms as implemented in the Scikit-learn library [14] on the extracted 22 features: Naïve Bayes (Gaussian) [5], Logistic Regression [15], Support Vector Machine (linear, polynomial, RBF) [16], Random Forest [10], Gradient Boosting [17] and Extra Trees [18].

The hyper-parameters of each classifier were optimized as follows: we selected important hyper-parameters and reasonable ranges (see Table 3), and then uniformly sampled 1000 times for each training fold. The parameters resulting in the highest AUC on the validation set were used to compute the AUC on the test set. This process was repeated 5 times (5-fold cross-validation) and we report mean AUCs with standard deviations (STD) for the validation and test sets.

Feature agnostic approach

The feature-agnostic approach does not extract manually designed features from the OCT volume but operates on the raw data. Apart from down-sampling (linear interpolation) from

Table 2. Average numbers of healthy eyes and eyes with POAG in training, validation and test set.

	Healthy	POAG
Training	216	672
Validation	30	82
Test	17	93

https://doi.org/10.1371/journal.pone.0219126.t002

Table 3. Hyper-parameters and parameter ranges used for parameter tuning on validation set.

Classifier	Parameter ranges
Naïve Bayes	none

Logistic regression	C = $[10^{-1}\ldots10^{1}]$
	penalty = {l1, l2}

Linear SVM	C = $[10^{-3}\ldots10^{3}]$

Polynomial SVM	C = $[10^{-3}\ldots10^{3}]$
	degree = {2, 3}

RBF SVM	C = $[10^{-3}\ldots10^{3}]$ $\gamma = [10^{-3}\ldots10^{3}]$

Random Forest	max_features = [0.1. . .1.0]
	n_estimators = {10, 50, 100, 500, 1000}
	min_samples_split = {2, 4, 6, 8, 10, 20, 40, 60, 100}
	min_samples_leaf = {1, 3, 5, 7, 9}

Gradient Boosting	learning_rate = $[10^{-1}\ldots10^{0}]$
	n_estimators = {100, 200, 500, 1000}
	max_depth = [2. . .10]
	min_samples_split = {2, 4, 6, 8, 10}
	min_samples_leaf = {1, 3, 5, 7, 9}

Extra Trees	max_features = [0.1. . .1.0]
	n_estimators = {10, 50, 100, 500, 1000}
	min_samples_split = {2, 4, 6, 8, 10, 20, 40, 60, 100}
	min_samples_leaf = {1, 3, 5, 7, 9}

https://doi.org/10.1371/journal.pone.0219126.t003

200x200x1024 to volumes with dimensions 64x64x128 voxels due to constraints of the GPU memory (12GB), no other pre-processing or data extraction was performed.

The downsampled volumes were inputted into a CNN [7], depicted in Fig 1. The network is composed of five 3D-convolutional layers with ReLU activation, batch-normalization, filter banks of sizes 32-32-32-32-32, filters of sizes 7-5-3-3-3 and strides 2-1-1-1-1. After the last convolutional layer Global Average Pooling (GAP) [19] is employed and a dense layer to the final softmax output layer is added to enable the prediction of class labels and the computation of CAMs.

An important aspect of the network architecture is the choice of 3D convolutions to allow the computation of 3D Class Activation Maps (CAM) [19]. The input layer of a CNN aggregates input data along the first axis (e.g. color channels). In the case of 2D convolutions the resulting CAM would be 2D and the depth information lost. We therefore employed 3D convolutions, which allowed us to identify regions within the OCT volume that are important for disease classification.

Various aspects of the network architecture such as the number of layers, number of filter banks per layer, filter sizes, strides and the use of batch normalization were optimized by

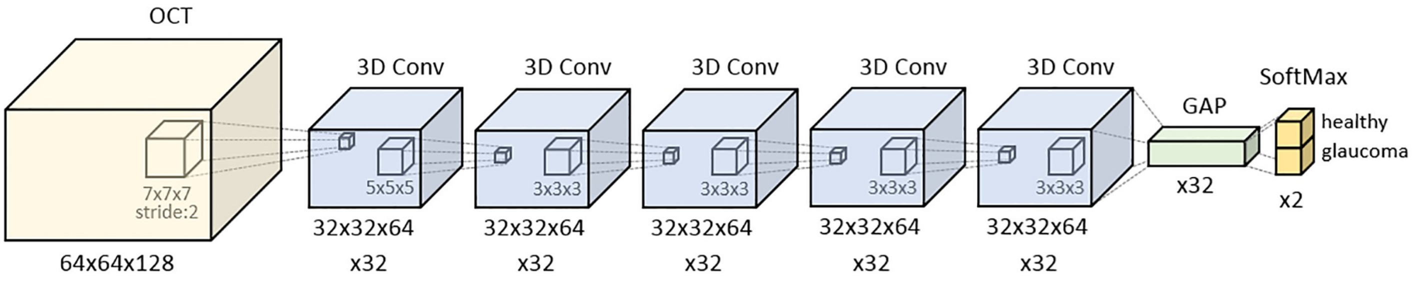

Fig 1. Network architecture.

https://doi.org/10.1371/journal.pone.0219126.g001

random hyper-parameter exploration; similar to the hyper-parameter optimization performed for the feature-based approached. The AUC achieved by the network was used to select the best network. We excluded max-pooling from the network architecture search since it can be replaced by strided convolutions [20]. We also did not explore different activation functions but used ReLU as proposed for CAM generation. However, we studied the impact of different gradient based learning algorithms [21], namely RMSProp, Adam, NAdam and Stochastic Gradient Decent (SGD) [22], and found NAdam to perform best.

The CNN was implemented in Keras [23] with Tensorflow [24] as the backend. Data splitting, stratification and pre-processing was performed with nuts-flow/ml [25]. Training was performed on a single K80 GPU using NAdam with a learning rate of $1e - 4$ over 100 epochs. Data was stratified per epoch via down-sampling. Training data was augmented by random occlusions, translations, left-right eye flipping, small rotations (±10 degrees) along the enface axis, and mixup [26]. However, we also trained the network without any augmentation and report the corresponding AUC. The network with the highest validation AUC during training was saved (early stopping). Accuracies reported are AUCs on the independent test set and the validation set.

CAMs were computed following [19], resized and overlayed on the input OCT scan. Note that CAMs are computed for smaller input OCTs 64x64x128 and then mapped back to scans with the original dimensions of 200x200x1024.

Results

In the following section we first report the prediction accuracies of the feature-based methods and the feature-agnostic CNN, before analyzing a selection of the CAMs generated by the CNN.

Disease detection

The prediction accuracies of the classical, feature-based machine learning methods on the validation and the test data is shown in Table 4. Logistic regression achieved the highest test AUC of 0.89 closely followed by linear SVM. Differences between validation and test AUCs were small for low-capacity classifiers such as Logistic Regression, Naive Bayes and linear SVM. Tree based algorithms, such as Random Forest, Extra Trees and Gradient Boosting tended to overfit—likely due to the larger capacity, the large number of hyper-parameters and the extensive hyper-parameter optimization.

Using the Extra Trees classifier, we evaluated the importance of individual features [14]. We observed large variations in the importance of features and therefore performed 100-fold

Table 4. 5-fold cross-validated prediction performance (mean AUC) of feature-based methods on validation set (AUC_{val}) and test set (AUC_{test}) with standard deviation. Last column shows the differences between test and validation AUCs.

Algorithm	AUC_{val}	AUC_{test}	$AUC_{val-test}$
Logistic Regression	0.88±0.035	**0.89**±0.028	-0.013
SVM (linear)	0.89±0.044	0.88±0.038	0.007
SVM (rbf)	0.90±0.045	0.86±0.039	0.033
Random Forest	0.91±0.034	0.86±0.027	0.043
Extra Trees	0.90±0.038	0.86±0.046	0.043
Naive Bayes	0.87±0.033	0.86±0.029	0.015
Gradient Boosting	0.87±0.033	0.82±0.043	0.049
SVM (poly)	0.85±0.030	0.82±0.033	0.034

https://doi.org/10.1371/journal.pone.0219126.t004

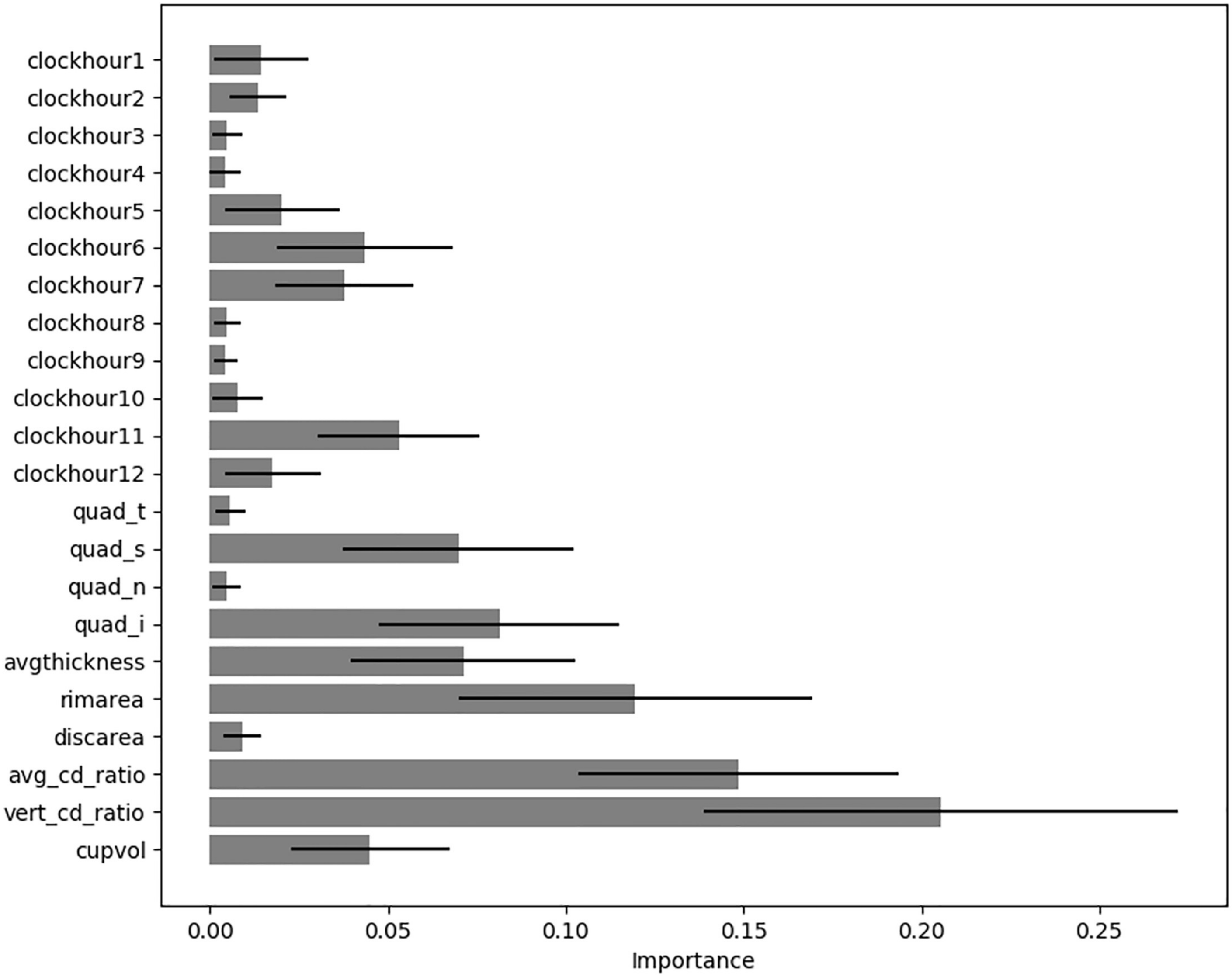

Fig 2. Importance of individual features for glaucoma classification. Error bars show standard deviation. Features are peripapillary RNFL thickness at 12 clock-hours (clockhour1..clockhour12), peripapillary RNFL thickness in the four quadrants (quad_t..quad_i), average RNFL thickness (avgthickness), rim area (rimeara), disc area (discarea), average cup-to-disc ratio avg_cd_ratio), vertical cup-to-disc ratio (vert_cd_ratio) and cup volume (cupvol).

https://doi.org/10.1371/journal.pone.0219126.g002

cross-validation to achieve stable results. Hyper-parameters for the Extra Trees classifier were optimized on the validation set by random search over 100 trials.

The bar plot in Fig 2 shows the mean importance with standard deviations of all features used for glaucoma classification. We find the well known indicators for glaucoma such as 6 and 11 o'clock clock-hours, inferior and superior quadrant and vertical cup-to-disc ratio having the largest importance.

Table 5 lists the 5-fold cross-validation accuracies of the CNN on the OCT data set. The feature-agnostic based approach achieved a peak test AUC of 0.94, which is substantially higher ($p < 0.05$) than the best classical machine learning method (AUC of 0.89) on segmentation-based features. We found that the extensive augmentation of training data had very little effect on test or validation accuracy but training was considerably faster without augmentation.

Table 5. 5-fold cross-validated prediction performance (mean AUC) of feature-agnostic CNN on validation set (AUC_{val}) and test set (AUC_{test}) with standard deviation. Last column shows the differences between test and validation AUCs. Results are reported for training with and without augmentation.

Algorithm	augmentation	AUC_{val}	AUC_{test}	$AUC_{val-test}$
CNN	no	0.93±0.015	**0.94**±0.036	-0.003
CNN	yes	0.95±0.018	0.92±0.046	0.027

https://doi.org/10.1371/journal.pone.0219126.t005

Visualizing CNN's attention

We computed Class Activation Maps (CAMs) to identify the regions in an OCT volume the CNN deems to be important for the classification decision. Fig 3 shows two representative CAMs, one for a healthy eye (Fig 3a and 3b) and one for an eye with POAG (Fig 3c and 3d). Note that aspect ratios of scans do not reflect physical dimensions of OCT volumes.

For healthy eyes the network tends to focus on a section across all layers but usually ignores the optic cup/rim and the lamina cribrosa. In contrast, for POAG eyes the CAMs generally highlight the optic disc cupping and neuroretinal rims as well as the lamina cribrosa and its surrounding regions. These regions agree well with the established clinical markers for glaucoma diagnosis (e.g. cup diameter/volume and rim area/volume).

The visualization software for CAM results with some example volumes is freely available at https://zenodo.org/record/1344287#.W3EN3dUzbmE.

Discussion

Huang et al. [8], Kim et al. [11] and Silva et al. [9] used machine learning based on segmentation-based OCT and other features to detect glaucoma. They report considerably higher peak AUCs between 0.95 and 0.99 than the test AUC of 0.89 we measured for classical machine learning algorithms on our data set. There are several likely reasons for these large differences in performance. Firstly, Kim et al. [11] and Silva et al. [9] utilized datasets that were 2 to 5 times smaller than our own. Over-fitting to smaller datasets is a commonly encountered issue in machine learning. Furthermore, some of these methods were not evaluated on a hold-out test set with additional steps such as a feature selection being performed on the validation set. The further incorporation of IOP measurements and visual field tests (MD, PSD and GHT), that are highly correlated with glaucoma, likely contributed to their higher prediction accuracy. Finally, and most importantly, our data set was not cleaned for this experiment, and arguably represents the challenge as it exists in the clinic today. The signal strength threshold in this experiment was 7, while many studies typically exclude scans with SS < = 8. While strict exclusion criteria such as visual field defect thresholds and low corrected vision [8] are common, our cohort did not exclude such patients and was quite varied and challenging.

The work of Muhammad et al. [12] is most similar to our work, in that they employ a CNN for glaucoma detection. It is, however, important to note that their method is still based on features extracted from segmented volumes such as thickness maps. Other differences are specific inclusion criteria for their cohort, the use of wide-field swept source OCT data and specific design choices. While transfer learning has the advantage of not requiring a large dataset, the architecture of the base network can be a severe limitation. AlexNet, is a 2D CNN and training on thickness and probability maps that does not permit the computation of CAMs for OCT volumes. Our approach, of training a 3D CNN from OCT volumes enables the computation of CAMs in volumes. In addition to common disease markers such as increased cup volume, cup diameter, thinning of neuroretinal rim at the superior and inferior segment, CAMs also consistently highlighted changes at the lamina cribrosa and the surrounding areas (see Fig 3d). In recent glaucoma studies [27, 28] the lamina cribrosa has become a focus as a potentially useful

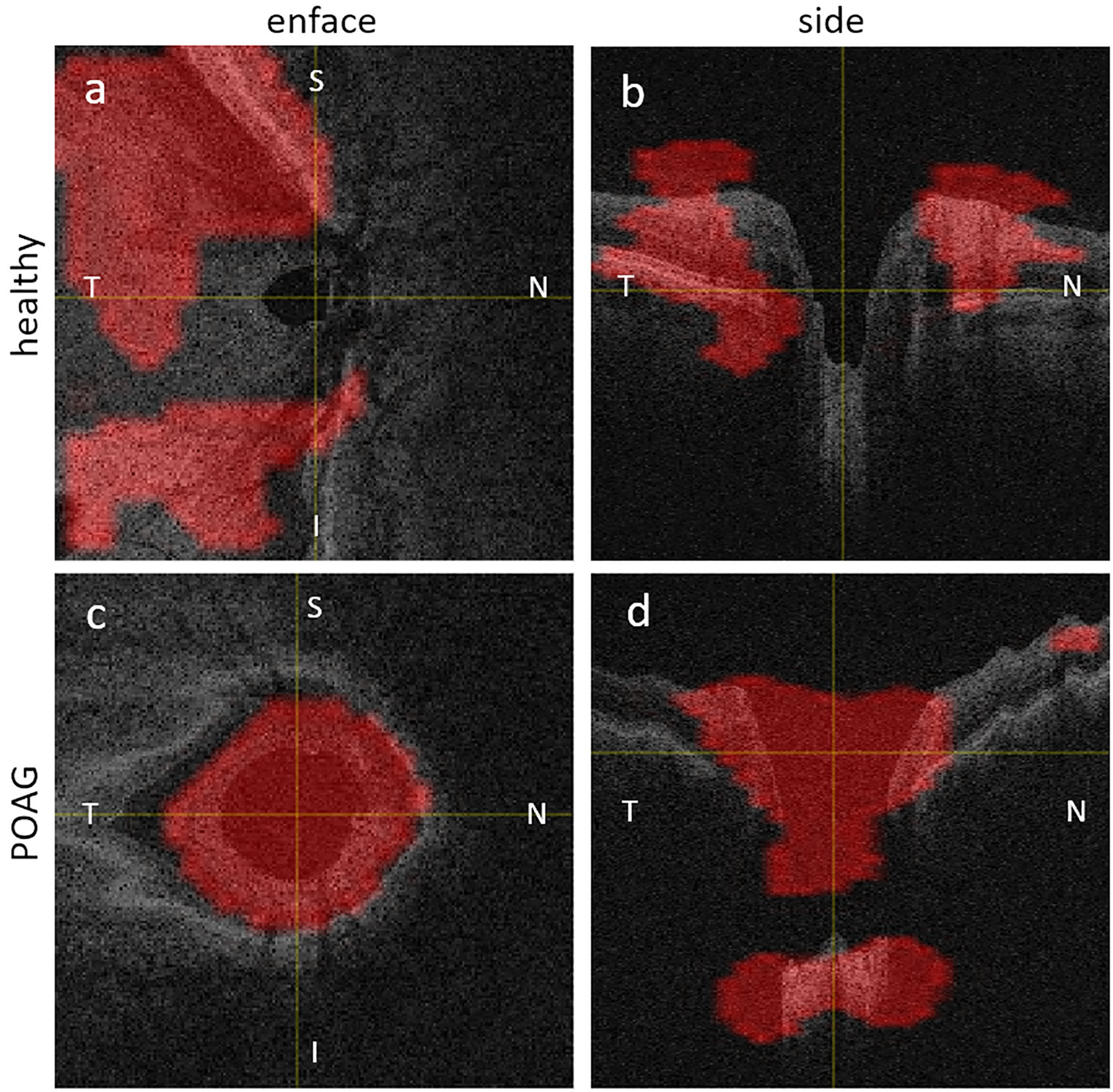

Fig 3. CAMs of a healthy and a POAG eye. Top row shows enface (a) and side (b) view of healthy eye. Bottom row shows enface (c) and side (d) view of POAG eye. (N:Nasal, T:Temporal, S:Superior, I:Inferior).

https://doi.org/10.1371/journal.pone.0219126.g003

structure that can be directly visualized and quantified in vivo and may provide new clinical biomarkers for glaucoma assessment.

The present CAM outcome implies a potential of establishing such biomarkers. However, the usefulness of CAMs depends to a large degree on the network architecture. Since CAMs are derived from the global-averaging-pooling (GAP) layer their resolution depends on the number of max-pooling operations or strided convolutions performed in earlier layers. For instance, an input volume of 128x128x128 will be reduced to a tiny CAM of size 4x4x4 pixels after five convolutions with stride 2 (128 / 2x2x2x2x2), resulting in blurry CAMs that fail to highlight distinct regions when mapped back to the input volume. We therefore chose a CNN architecture with good classification accuracy but small strides and filters sizes. The large size of an OCT volume and the limited GPU memory (12GB) also forced us pick a comparatively shallow network with five layers. Higher classification accuracies may be achieved with deeper networks of different architecture.

It is noteworthy, that during our empirical exploration of hyper-parameters we did not identify any specific network properties of importance for good classification performance apart from batch normalization and learning algorithm (NAdam performed best). All other parameters such as number of filter banks, filter sizes, strides or learning rate showed no correlation with prediction accuracy. On the contrary, very different architectures achieved very similar validation AUCs. Even attempts to flatten and crop the retinal layers in order to normalize OCT scans had little effect on classification accuracy.

Finally, our data set showed statistically significant differences between healthy and glaucoma patients for age, IOP, MD and GHT. While the IOP and visual function measurements are expected to differ between the two groups, the inclusion of age might influence the performance of the CNN. For visual function measurements such as MD and GHT these differences are expected and aimed for. Similarly, differences in IOP between healthy and eyes with glaucoma are expected. Age can be inferred from OCT, e.g. due to progressive layer thinning with advancing age, and while age was not directly included as a feature the CNN potentially takes advantage of it.

Conclusions

In this work we demonstrated that the detection of glaucoma from raw OCT volumes is achievable with an accuracy comparable or better than traditional, feature-based approaches that rely on manually designed features extracted from segmented OCTs. The feature-agnostic approach potentially widens the range of application and improves detection accuracy, since OCT scans of older patients or extreme cases of glaucoma are often difficult to segment accurately.

Manually designed features have the advantage of human interpretability. We employed CAMs with similar purpose and result. They allowed us to identify OCT regions important for glaucoma classification and potentially are helpful for the discovery of novel or more robust disease markers.

Our results are based on the largest OCT glaucoma data set so far but were limited to ONH scans only. Including Macula scans and other readily available features such as IOP and visual test measurements are likely to increase the accuracy of the method further.

Author Contributions

Conceptualization: Stefan Maetschke, Bhavna Antony, Hiroshi Ishikawa.

Resources: Gadi Wollstein, Joel Schuman.

Writing – original draft: Stefan Maetschke.

Writing – review & editing: Bhavna Antony, Hiroshi Ishikawa, Rahil Garnavi.

References

1. NICE. Glaucoma: diagnosis and management. National Institute for Health and Care Excellence: Clinical Guidelines. 2017;.
2. Huang D, Swanson EA, Lin CP, Schuman JS, Stinson WG, Chang W, et al. Optical coherence tomography. science. 1991; 254(5035):1178–1181.
3. Medeiros FA, Zangwill LM, Alencar LM, Bowd C, Sample PA, Susanna R, et al. Detection of glaucoma progression with stratus OCT retinal nerve fiber layer, optic nerve head, and macular thickness measurements. Investigative ophthalmology & visual science. 2009; 50(12):5741–5748. https://doi.org/10.1167/iovs.09-3715
4. Lucy KA, Wollstein G. Structural and functional evaluations for the early detection of glaucoma. Expert review of ophthalmology. 2016; 11(5):367–376. https://doi.org/10.1080/17469899.2016.1229599 PMID: 28603546
5. Russell SJ, Norvig P. Artificial Intelligence—A Modern Approach (3. internat. ed.). Pearson Education; 2010.
6. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.
7. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015; 521(7553):436. https://doi.org/10.1038/nature14539 PMID: 26017442
8. Huang ML, Chen HY. Development and comparison of automated classifiers for glaucoma diagnosis using Stratus optical coherence tomography. Investigative ophthalmology & visual science. 2005; 46 (11):4121–4129. https://doi.org/10.1167/iovs.05-0069
9. Silva FR, Vidotti VG, Cremasco F, Dias M, Gomi ES, Costa VP. Sensitivity and specificity of machine learning classifiers for glaucoma diagnosis using Spectral Domain OCT and standard automated perimetry. Arquivos brasileiros de oftalmologia. 2013; 76(3):170–174. https://doi.org/10.1590/S0004-27492013000300008 PMID: 23929078
10. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324
11. Kim JS, Ishikawa H, Sung KR, Xu J, Wollstein G, Bilonick RA, et al. Retinal nerve fibre layer thickness measurement reproducibility improved with spectral domain optical coherence tomography. British Journal of Ophthalmology. 2009; 93(8):1057–1063. https://doi.org/10.1136/bjo.2009.157875 PMID: 19429591
12. Muhammad H, Fuchs TJ, De Cuir N, De Moraes CG, Blumberg DM, Liebmann JM, et al. Hybrid Deep Learning on Single Wide-field Optical Coherence tomography Scans Accurately Classifies Glaucoma Suspects. Journal of glaucoma. 2017; 26(12):1086–1094. https://doi.org/10.1097/IJG.0000000000000765 PMID: 29045329
13. Asman P. Glaucoma Hemifield Test. Archives of Ophthalmology. 1992; 110(6):812. https://doi.org/10.1001/archopht.1992.01080180084033 PMID: 1596230
14. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011; 12(Oct):2825–2830.
15. Cox DR. The regression analysis of binary sequences. Journal of the Royal Statistical Society Series B (Methodological). 1958; p. 215–242. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x
16. Scholkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press; 2001.
17. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in neurorobotics. 2013; 7:21. https://doi.org/10.3389/fnbot.2013.00021 PMID: 24409142
18. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine learning. 2006; 63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1
19. Zhou B, Khosla A, Lapedriza À, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. CoRR. 2015;abs/1512.04150.
20. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:14126806. 2014;.
21. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. nature. 1986; 323(6088):533. https://doi.org/10.1038/323533a0
22. Ruder S. An overview of gradient descent optimization algorithms. CoRR. 2016;abs/1609.04747.
23. Chollet F. Keras; 2017. Available from: https://github.com/fchollet/keras.
24. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467. 2016;.

25. Maetschke S, Tennakoon RB, Vecchiola C, Garnavi R. nuts-flow/ml: data pre-processing for deep learning. CoRR. 2017;abs/1708.06046.

26. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:171009412. 2017;.

27. Downs JC, Girkin CA. Lamina cribrosa in glaucoma. Current opinion in ophthalmology. 2017; 28 (2):113. https://doi.org/10.1097/ICU.0000000000000354 PMID: 27898470

28. Abe RY, Gracitelli CP, Diniz-Filho A, Tatham AJ, Medeiros FA. Lamina cribrosa in glaucoma: Diagnosis and monitoring. Current ophthalmology reports. 2015; 3(2):74–84. https://doi.org/10.1007/s40135-015-0067-7 PMID: 26052477