Glaucoma diagnosis automation

Machine learning; Data science; Glaucoma

White paper

Version 1.0, February 2024

Abstract

Glaucoma is one of the leading causes of irreversible blindness worldwide. It encompasses a group of conditions that share a common characteristic: optic nerve damage. The optic nerve, responsible for transmitting visual information from the retina to the brain, undergoes structural and functional changes, resulting in a gradual, often unnoticed, loss of vision. These changes are primarily attributed to elevated intraocular pressure (IOP), stressing the optic nerve fibers, impairing their functioning, and causing sudden degeneration in the case of closed-angle glaucoma, or a more progressive loss of vision in the case of open-angle glaucoma, which is the focus of this paper.

The clinical diagnosis of glaucoma involves a combination of qualitative and quantitative assessments, such as visual acuity, visual field testing, and optic nerve head evaluation using imaging techniques like optical coherence tomography (OCT). However, although these conventional diagnostic methods are helpful, they have inherent limitations, such as variability in human interpretation, subjectivity, and reliance on the clinician's expertise.

Moreover, early-stage glaucoma can often go undetected due to the lack of noticeable symptoms, further emphasizing the need for objective and sensitive diagnostic tools. To tackle these issues, the integration of Machine Learning (ML) approaches has emerged as a promising technique to aid in glaucoma detection, either by directing the professional's attention to riskier cases (by ruling out healthy eyes or pointing out those at risk), or simply by acting as a second opinion during the final diagnosis. Due to the algorithms' ability to extract patterns from large amounts of data, this approach has shown promise in various other areas of the medical field.

This paper describes the cognitive challenge and technical approach to implementing a model to automate glaucoma diagnosis. A dataset construction methodology is described, relying on features present in ocular exams of anonymized patients, automatically extracted using Optical Character Recognition (OCR) tools. Following, the use of this data throughout the typical Data Science (DS) / ML stages of exploratory analysis is explained, along with feature engineering, and model training/evaluation. Particular attention is given to explainability as, in real-life situations, it is valuable for healthcare professionals to comprehend what factors lead to a specific decision to determine whether to support it. Finally, some insights on the outcomes achieved and how to move on for an in-field evaluation are provided.

Introduction

Glaucoma, a group of progressive optic pathologies, remains a significant global health concern affecting millions worldwide. This sight-threatening condition is characterized by damage to the optic nerve, which leads to irreversible vision loss when left undiagnosed and untreated. As the leading cause of irreversible blindness worldwide, glaucoma presents an urgent need for accurate and timely detection to mitigate its detrimental effects on affected individuals.

This pathology constitutes a considerable burden on public health, with its prevalence steadily rising as the population ages. It is estimated that over 80 million people globally are affected by glaucoma and is projected to reach about 112 million by 2040. The condition is particularly prevalent among older individuals (60+ years old) and varies among different ethnicities and geographical regions.

The disease's progressive nature often leads to a gradual loss of peripheral and central vision, referred to as visual field loss. When left untreated, glaucoma can progress to complete irreversible blindness, impairing a person's ability to perform daily activities, navigate their environment, and engage in social interactions.

One of the primary challenges in managing glaucoma lies in its early stages' silent and asymptomatic nature. This characteristic makes a timely diagnosis challenging, as individuals may remain unaware of their condition until it is too late and irreversible damage has been done. Additionally, the diagnostic process involves several complexities, including the reliance on subjective assessment, observer variability, and the need for frequent and extensive monitoring. Traditional diagnostic techniques, such as intraocular pressure measurement and visual field testing, generate data with inherent limitations in terms of reliability. Furthermore, the scarcity of skilled ophthalmologists and the increasing global glaucoma burden requires the development of innovative and accessible diagnostic tools.

In terms of the exams, those that are most relied upon when checking for glaucoma are:

Optical Coherence Tomography (OCT)

Non-invasive imaging technology used to get a detailed mapping of the inside of the eye, by following the logic of an ultrasound, but using invisible light instead of sound waves.

Visual Field Test (VFT)

Non-invasive exam to assess the state of the patient's peripheral vision by asking him to focus on a central spot and press a button whenever he notices lights flicker around the fixation light. The exam is performed one eye at a time, with the other eye covered.

The result of the OCT is a set of high-resolution images of different parts of the eye. These images are often summarized using metrics that are aggregated and presented in reports, relative to the specific part of the eye being evaluated. Examples of those reports include:

- **Biometry:** captures several measurements to understand different characteristics such as the axial length, corneal curvatures, and angle and thickness of various parts of the eye;

- **Anterion:** like biometry, it takes several OCT measurements of the patient's eye, with a focus on the frontal part. It also measures the angle of the anterior chamber;

- **Retinal Nerve Fiber Layer (RNFL):** measures the thickness of the nerve layers, providing a quantitative and qualitative assessment, useful for obtaining a reliable estimate of the damage to the optic nerve if it is found to be decreasing in thickness;

- **Glaucoma Overview (GO):** measures retina thickness, asymmetry, and ganglion cell layer thickness;

- **Deviation Map Single Report (DMSR):** quantitative and qualitative evaluations of the macular retina and macular ganglion cell layer.
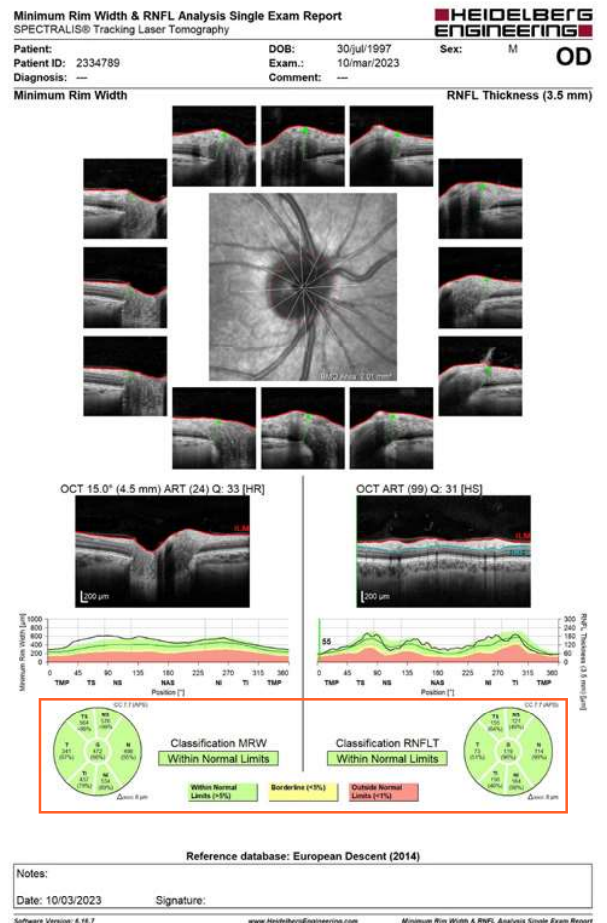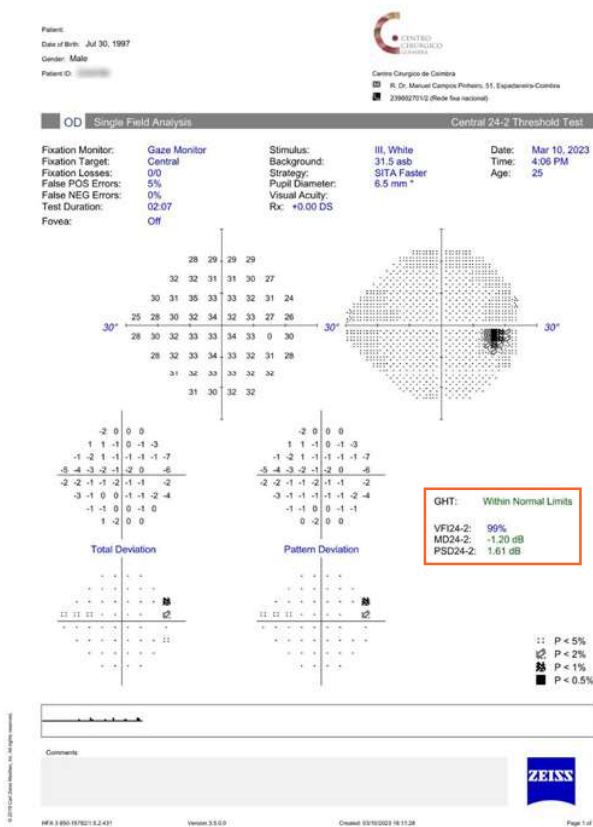

Figure 1 – Example of the VFT and RNFL exams, highlighting the most relevant features

As far as the VFT is concerned, the report contains a map of the eye's blind spots as well as other measurements that compare the state of the eye with itself in the past (follow-up), as well as with the general population. It is worth noting that the VFT is prone to errors since it is somewhat subjective and exposed to many outside factors that can influence the results directly, such as false positives (when the patient presses the button for some reason despite no light being shown) and false negatives (when the patient doesn't press the button when the light is turned on although it was expected to have pressed, based upon earlier responses).

In recent years, advancements in ML techniques have shown promising potential for improving the detection and monitoring of various health conditions, namely glaucoma. By leveraging large datasets, ML algorithms can extract intricate patterns and features from clinical data, enabling more accurate and efficient identification of possible cases of glaucoma.

With this work, we hope to contribute to the ongoing efforts to develop robust and reliable ML solutions that can aid clinicians in the early detection, monitoring, and management of glaucoma. Ultimately, the integration of ML techniques into clinical practice presents immense potential to revolutionize glaucoma detection, leading to better outcomes regarding the patients' ocular health in the future.

Relevant works

The use of ML for glaucoma diagnosis in the literature has been gaining popularity in the last few decades, following two main approaches: one that uses features already extracted from medical exams in a tabular approach, while the other uses high-quality medical scans of the eye and Deep Learning (DL) to derive its own features. The feature engineering and the tabular nature of the data used in the first approach usually lead to more explainable models from which we can extract relevant insights. On the other hand, DL often leads to better classification results at the cost of interpretability. **Table 1** presents a summary of most relevant works.

Maetchke et al. [1] used RNFL thickness values, rim area and disc area features, among others, to predict glaucoma in a dataset with 263 healthy and 847 glaucomatous eyes, achieving Area Under the Curve (AUC) scores between 0.82 and 0.89, while a DL approach using a Convolutional Neural Network (CNN) to predict glaucoma from fundus images[1] achieved a better AUC score of 0.94. Although important, these results significantly depend on the size of the dataset used to compute them.

Most of the published results on this subject rely on significantly smaller datasets, especially when it comes to the glaucoma class. Kim SJ et al. relied on tabular data from multiple exams to predict glaucoma on a dataset with 178 glaucomatous and 164 healthy eyes with classic ML algorithms, achieving high accuracy scores around 0.97.

Muhammad et al. [2], similarly, worked with both summary metrics from OCT scans as well as the raw exams by using a CNN to extract features from the scans. They worked with 57 and 45 glaucoma and non-glaucoma (healthy or suspected glaucoma cases), respectively, obtaining significantly better results - 93% accuracy - when using the features extracted using CNN, compared to the 64% accuracy using OCT summary metrics and visual field test features, following a classic ML approach.

Asaoka et al. [3] relied on the visual field exam alone to distinguish between 53 glaucomatous and 108 healthy eyes comparing the use of DL and classic ML algorithms, where Deep Neural Networks (DNN) achieved a top AUC score of 0.926 while the alternatives all had scores below 0.80.

The more standard approach with DL in this use case uses other types of imageology exams, mainly fundus and sometimes raw OCT scans (instead of summary metrics), rather than visual field tests. A meta-analysis by Jo Hsuan Wu et al. [4] revealed that approaches relying on fundus achieved on average a 0.97 AUC while OCT's AUC was 0.96.

Paper	Features	Data	Results
Maetchke et al.	RNFL	Glaucoma: 847 Non-Glaucoma: 263	CNN: 0.94 AUC ML: 0.82 – 0.89 AUC
Muhammad et al.	OCT (summary and raw)	Glaucoma: 57 non-Glaucoma: 45	CNN: 0.93 Accuracy ML: 0.64 Accuracy
Asaoka et al.	Visual field exams	Glaucoma: 53 Non-Glaucoma: 108	DNN: 0.926 AUC ML: <0.80 AUC

Table 1 – Summary of relevant works

---

1 Fundus imaging is a noninvasive technique used to capture a detailed image of every aspect of the retina, including the size, shape, and color of significant regions such as the optic disc (OD), optic cup (OC), blood vessels, neuro-retinal rim, and fovea, that is used to indicate glaucoma presence.

Methods

Data sources and ingestion

The work described in this paper has been developed in partnership between Altice Labs and the Centro Cirúrgico de Coimbra (CCC) that provided different kinds of exams for healthy, suspect, and glaucomatous eyes. They were fundamental with the necessary expertise for the selection of the most relevant and reliabile of the expected outcomes, which was of paramount importance for its validation.

Besides the anonymized exams' data, a diagnosis was also provided, indicating whether the case had been diagnosed as non-glaucoma, suspect (of glaucoma), or glaucoma. Information regarding additional pathologies the patient might suffer, which is useful in ensuring that the model distinguishes eyes with glaucoma vs. non-glaucoma, instead of healthy vs. unhealthy was also included.

The exams were sent in an image format and scripts had to be created, specifically tailored for each exam type, to extract the relevant information using OCR technology, more specifically, Tesseract [5]. Given the nature of the exams/reports, we opted for the tabular and classic ML approach previously mentioned to process the data.

Features and dataset creation

For the development of our dataset, besides the features directly extracted from the exams, we took also into consideration features derived from them. The final feature selection process evolved as an iterative approach, and considered:

The feedback from CCC, which provided insights into the reliability of the exams as well as the priority features to look at when making a diagnosis

Correlation studies among the features themselves and between the features and the target, along with statistical analysis

The results of the feature importance assessments based on Principal Component Analysis (PCA)

Examples of derived features include:

- In RNFL, from the initial values (both absolute and relative to some reference set) for the thickness of the retinal nerve fiber layers in different areas of the optic nerve (represented by the letters: I-inferior, S-superior, N-nasal, T-temporal, G-general), shown in **Figure 2**; a new set of custom features was derived that summarizes the entire sector in terms of maximum and minimum absolute/relative values, for example (see **Figure 2**).
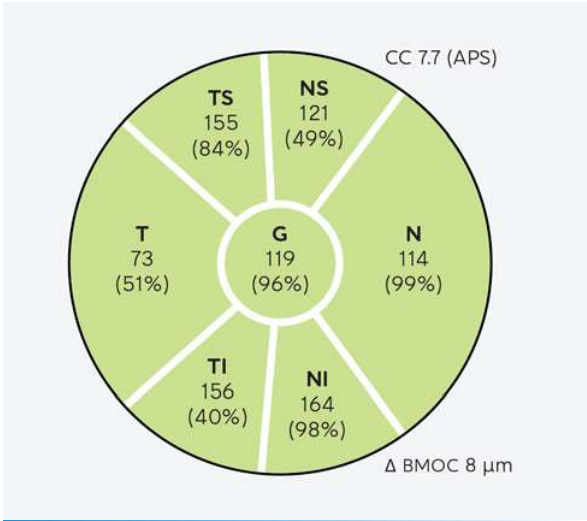

Figure 2 – Example of a RNFLT sector of a RNFL exam, showing the thickness of the nerve layers in various areas

- In GO, the report contained an 8x8 matrix (**Figure 3**) displaying the asymmetries in retinal thickness between the top and bottom hemisphere of the eye, where darker entries indicate a larger difference between the thickness in its position and the symmetric position of the lower hemisphere. To capture this without including 64 extra features, we extract metrics such as min, max, median, and mean for each row and column instead (though the use of convolutional methods could be explored here).
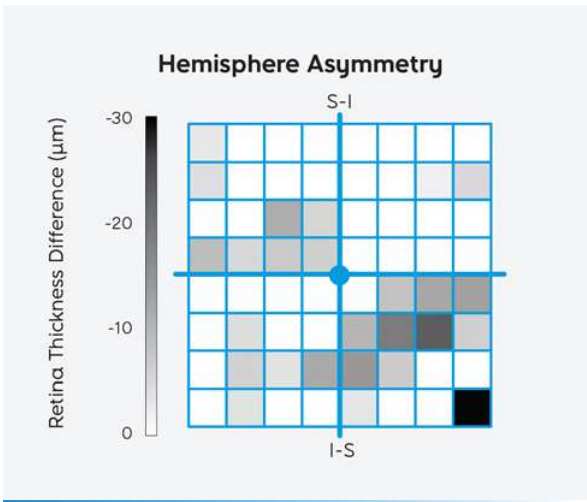

Figure 3 – Example of a hemisphere asymmetry matrix from a GO exam, displaying the difference in thickness between the top and bottom hemispheres of the retina

Training

ML models

Throughout our experiments, we relied mostly on classic ML models such as Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM), with the most out of the ordinary models being the Light Gradient Boosting Model (LGBM) [6] and the Explainable Boosting Machine (EBM) from the InterpretML library [7]. The choice of the first three algorithms (LR, RF, and SVM) was primarily based on their common occurrence in the relevant literature, while LGBM was chosen since this type of gradient-boosted trees tend to achieve state-of-the-art performance on classification tasks with tabular datasets and it did perform the best. As for EBM, it was chosen because it allows for better explainability, but we found that our explainability tool (ExplainerDashboard [8]) was enough to satisfy our needs in a model-agnostic manner. All this considered, LGBM was selected as the final model to apply.

Setup

In preparing our machine learning model, one of the initial steps involved addressing missing values within the dataset. These missing values resulted from either the absence specifically from the corresponding exam, or the absence of the whole exam itself. To manage this, we opted to:

- Exclude the cases where an entire exam was missing;

- Use the KNN imputer method, which works by finding the nearest neighbors to the entry with a missing value and imputing the missing value as a weighted average of its neighbors; this imputer was fitted to avoid information leakage.

The resulting class distribution of the number of eyes was the following:

- Non-Glaucoma: 623 (64%)

- Suspect: 153 (15.7%)

- Glaucoma: 198 (20.3%)

We divided this dataset into train/test and validation subsets (70%/30%) ensuring a similar class distribution in both sets. After that, we also divided the train/test set into train and test subsets (80%/20%).

Regarding the targets, we opted for a multi-class scenario to make full use of the diagnosis information available. In early experiments, we adopted a binary classification scenario (labeling suspicion of glaucoma as non-glaucoma) but we found that according to the resulting metrics, including a separate class for the suspects helped in differentiating glaucoma from non-glaucoma.
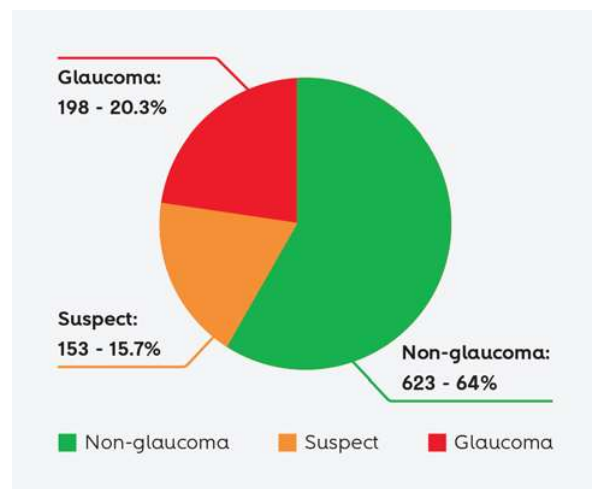

Figure 4 – Class distribution of the full dataset

The current feature set includes features extracted from the RNFL, Computerized Static Perimetry (PEC), DMSR, and GO exams, meaning that to make a prediction, the model evaluates the thickness of the retinal nerve fibers, the visual field, the macular retina, and the macular ganglion cell layer, as well as the asymmetry found in thicknesses between the top and bottom hemispheres of the eye.

Regarding the hyperparameter tuning, we followed different approaches throughout the stages of the work, from grid search to randomized search and to genetic algorithms, opting for the first in the chosen model. The main goal of this tuning was to avoid overfitting, by adapting the model's hyperparameters, since this was a problem identified during earlier iterations.

Results

We've incorporated diverse classification metrics, including Accuracy, Precision, Recall, F1-Score, and ROC-AUC (see **Table 2**), along with a thorough analysis of confusion matrices, to comprehensively assess model performance [10]. Given the imbalance in class distribution, special importance was given to the F1-score metric, especially of the positive class (Glaucoma), so, when deciding on certain parameters or features, this is what we tried to maximize.

Class	Precision	Recall	F1-score	ROC-AUC	Accuracy
Glaucoma	0.84	0.93	0.88	0.99	
Suspect	0.64	0.32	0.42	0.86	0.84
Non-Glaucoma	0.86	0.94	0.90	0.94	

Table 2 – Summary of relevant works

The F1-scores and confusion matrices (**Figure 5**) obtained from the final model, suggest that it is capable of distinguishing eyes with glaucoma and those of the non-glaucoma class while being consistently worse at classifying "suspect" ones. This might be explained by the distributions of the feature values in the non-glaucoma and suspect classes, which tend to be more similar to each other than to the glaucoma class feature value distribution. However, it is also limited by the CCC's criteria of what constitutes a suspect eye, as it lies in a "gray zone". The results achieved with our evaluation metrics seem to be in line with those found in similar works, when compared with
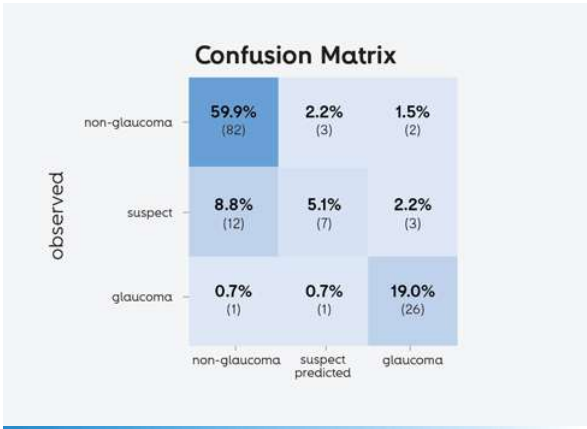

Figure 5 – Confusion matrix of the final model

the metrics that are most commonly reported (ROC-AUC), despite our focus lying on optimizing F1-scores, which were also good for the non-glaucoma and glaucoma classes. However, one must consider that a comparison with the relevant works isn't as simple as looking at the metrics, since this isn't a binary problem as is the case in most of the papers we found. Despite bringing up some classification difficulties, we still consider the inclusion of the suspect class to be the correct approach, since it is valuable information that otherwise would be wasted, and it is useful to distinguish sure cases from those more borderline in a real-life scenario.

In this scenario, the most critical errors involve misclassifying eyes, with a suspicion/actual glaucoma being labeled as non-glaucoma, potentially leading to patients not receiving necessary care. Other errors, though undesirable, are more tolerable, as they still flag non-glaucoma patients for monitoring.

To ensure that the approach chosen is robust, we've decided to experiment with training the same model using 200 different train/test and validation splits. The distribution of F1-scores per class across all runs is shown in **Figure 6**, where we can see that the suspect class is consistently the hardest to classify, while the model tends to predict non-glaucoma the best, with glaucoma showing good results across different splits.
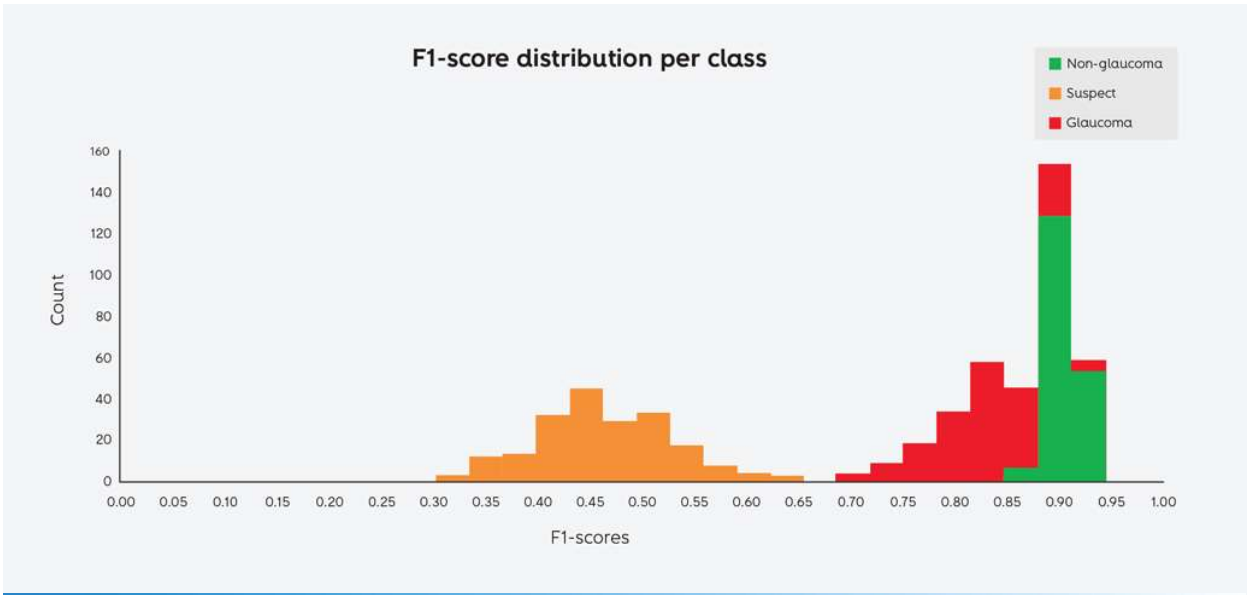

Figure 6 – Distribution of F1-scores per class when training and testing the final model with 200 different data splits

Explainability

Regarding explainability, we relied on a tool called Explainerdashboard [8], which could be of great use in a real-life scenario since it generates visualizations that allow the healthcare professionals to assess how relevant (and towards which diagnosis) each feature was, to the final result for a particular case - a way in which it could be used in a real scenario, since professionals focus their attention on one patient at a time - see **Figure 7**. This tool allows the clinician to measure how the diagnosis would change if some particular values were altered (by letting the clinician customize some parameters).
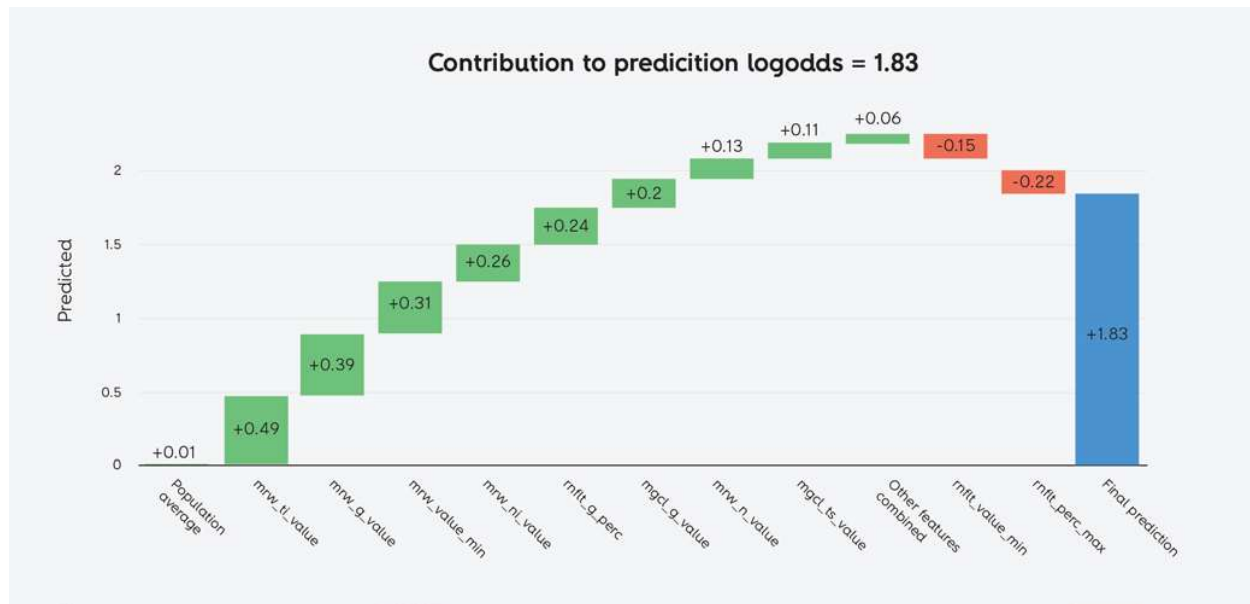

Figure 7 – Example of the use of Explainerdashboard to interpret the decision of one case in particular (an eye with glaucoma that was correctly classified as such in this example)

The Explainerdashboard uses primarily SHAP [9] values to help visualize how each feature influences the model's final decision. SHAP values indicate how much and in which direction a feature's value influences the model's prediction. For example, a large positive value means that a given feature has a significant impact in making the model predict the positive class (glaucoma in this case). Likewise, a large negative value skews the model towards a non-glaucoma diagnosis, while a small value has a minor impact either way.

After analyzing these visualizations, namely **Figure 8**, we came to some conclusions that seem to be corroborated by the medical community, regarding glaucoma diagnosis.

Each line of the graph represents a feature, ranked by impact, and each point represents an eye. Points leaning towards blue denote lower feature values, while those leaning toward red signify higher feature values. The point's position indicates its impact on the model's decision: centered points had no impact, rightward points leaned towards a glaucoma diagnosis, and leftward points leaned toward a non-glaucoma diagnosis.
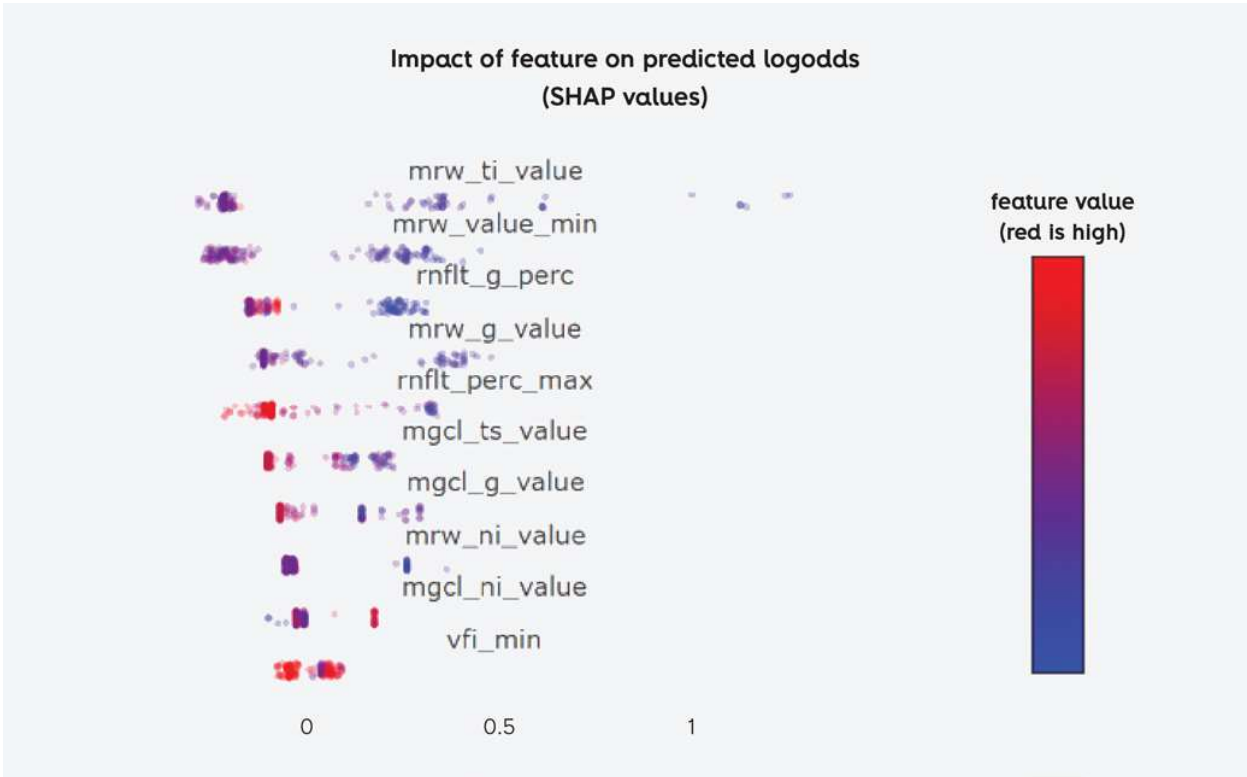

Figure 8 – Analysis (SHAP) of the impact of each feature in the glaucoma diagnosis (ordered from most to least impactful)

Overall, one can see the relevance of the RNFL exam, once its features top the feature importance charts throughout the various iterations, including in the presented iteration, where the top 5 features all belong to the RNFL exam. In many cases, RNFL features would be enough to provide a decently accurate diagnosis by themselves, without relying on any other exams at all.

Conclusion

In conclusion, the results obtained with this work look promising in terms of the ability to detect glaucoma, one of the most widespread and challenging ocular pathologies to diagnose. The results are in line with the best results of the relevant work identified in the reference papers and provide a clear path to the accomplishment of the proposed objectives. It was shown that each iteration offers better results than the previous ones, clearly benefitting from the access to more samples. An additional effort to sanitize sample labels has also contributed to these results and the increasing model stability. Nevertheless, the suspect class has many misclassifications, which means we must look at it differently, adopting new approaches, such as cascade or ensemble models.

One of the main issues in most ML projects is how well the data science team can cope with the specificities of the area under study and, therefore, create domain knowledge – relevant exam fields, non-evident correlations, etc. – to support the model evolution. Close involvement with CCC experts – ophthalmologists, biomedical engineers, and exam technicians – has proven crucial in the process and should be shepherded and intensified along the way.

One additional challenge faced during the development of this project was the unreliability of the OCR process. However, it could be mitigated by an ongoing effort to integrate directly with the medical equipment and, therefore, extract the raw values.

The main goal is to introduce these processes into daily clinical practice. For now, we are envisaging two different approaches, which would require a slightly different model tuning. If the goal is to screen a vast array of patients and to remove those that pose a low risk of developing glaucoma, or point out those at high risk, we should tune it with a particular focus on the precision of the non-glaucoma class or the recall of the other two classes. If it is the case that we are looking to provide a professional who has already evaluated the patient's exams with a second opinion, the current model, along with the explainability tools, should be able to provide that assistance. A field trial is being planned to validate the quality of the models, identify and minimize the burdens of integration with the medical equipment, and, most importantly, collect feedback from professionals and patients to design and implement a diagnosis tool, integrating humanized communication and adequate explainability tools. In addition, the field trial will also allow the collection of more samples.

A natural evolution will be to move from diagnosis to predicting the risk of progression to pathology – to allow patients to receive proper care at the first signs of possible glaucoma development – which will require some additional history of tests and different technical approaches, given that we will be dealing with a use case with different characteristics. This work is already underway.

Aside from that, there are a few main options that could be further explored. One is the inclusion of high-definition imageology exams to apply DL along with the classic ML approach. Another one is the use of Neural Networks (NNs) as classification models. Still, a lot of high-quality samples would be needed, which does not seem realistic at this point without significant data augmentation (which could potentially introduce its fair share of problems as well). Finally, another direction would be to follow a similar methodology to address different ocular pathologies.

References

[1] S. Maetschke et al."A feature agnostic approach for glaucoma detection in OCT volumes," *PLOS ONE*, vol. 14, no. 7, p. e0219126, Jul. 2019, doi: 10.1371/journal.pone.0219126.

[2] H. Muhammad et al., "Hybrid Deep Learning on Single Wide-field Optical Coherence tomography Scans Accurately Classifies Glaucoma Suspects," *J Glaucoma*, vol. 26, pp. 1086–1094, 2017. Accessed: Nov. 2, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:205683952.

[3] R. Asaoka, H. Murata, A. Iwase, and M. Araie, "Detecting Preperimetric Glaucoma with Standard Automated Perimetry Using a Deep Learning Classifier," *Ophthalmology*, vol. 123, no. 9, pp. 1974–1980, 2016, doi: 10.1016/j.ophtha.2016.05.029.

[4] JH Wu, T Nishida, RN Weinreb, JW Lin, "Performances of Machine Learning in Detecting Glaucoma Using Fundus and Retinal Optical Coherence Tomography Images: A Meta-Analysis" *Am J Ophthalmol*, vol. 237, pp. 1–12, 2022, doi: 10.1016/j.ajo.2021.12.008.

[5] Tesseract, "Tesseract doumentation", 2023. Accessed: Nov. 2, 2023. [Online]. Available: https://tesseract-ocr.github.io/.

[6] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. Accessed: Nov. 2, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[7] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "InterpretML: A Unified Framework for Machine Learning Interpretability," *arXiv preprint* arXiv:1909.09223, 2019.

[8] O. Dijk et al., "oegedijk/explainerdashboard: explainerdashboard 0.4.2: dtreeviz v2 compatiblity." *Zenodo*, Feb. 2023. doi: 10.5281/zenodo.7633294.

[9] S. M. Lundberg and SI. Lee, "A unified approach to interpreting model predictions," CoRR, vol. abs/1705.07874, 2017. Accessed: Nov. 2, 2023. [Online]. Available: http://arxiv.org/abs/1705.07874.

[10] Wikipedia, "Evaluation of binary classifiers", 2023. Accessed: Nov. 2, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers.

Acronyms

AUC	Area Under the Curve
CCC	Centro Cirúrgico de Coimbra
CNN	Convolutional Neural Networks
DL	Deep Learning
DMSR	Deviation Map Single Report
DNN	Deep Neural Networks
DS	Data Science
GO	Glaucoma Overview
ML	Machine Learning
NN	Neural Networks
OCR	Optical Character Recognition
OCT	Optical Coherence Tomography
PCA	Principal Component Analysis
PEC	Computerized Static Perimetry
RNFL	Retinal Nerve Fiber Layer
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations

Authors

Rodrigo Ferreira

Junior Data Scientist

Rita Oliveira

Junior Data Scientist   ✉ rita-a-oliveira@alticelabs.com
Altice Labs, Portugal

Maria Manuel Castro

Senior Data Scientist   ✉ maria-m-castro@alticelabs.com
Altice Labs, Portugal   in https://www.linkedin.com/in/mariamanuelcastro/

Luís Cortesão

Head of Cognitive   ✉ luis-m-cortesao@alticelabs.com
Altice Labs, Portugal   in https://www.linkedin.com/in/luis-cortesao/

António Travassos

Ophthalmologist   ✉ oftalexames@ccci.pt
Centro Cirúrgico de Coimbra, Portugal