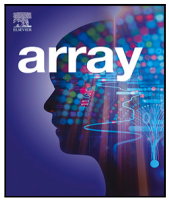
Array 23 (2024) 100359

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array



Evaluating machine learning techniques for enhanced glaucoma screening through Pupillary Light Reflex analysis

Hedenir Monteiro Pinheiro [a,*], Eduardo Nery Rossi Camilo [b,c], Augusto Paranhos Junior [c], Afonso Ueslei Fonseca [a], Gustavo Teodoro Laureano [a], Ronaldo Martins da Costa [a]

[a] Institute of Informatics - Federal University of Goiás, GO, Brazil
[b] Fundação Banco de Olhos de Goiás, GO, Brazil
[c] Escola Paulista de Medicina - Federal University of São Paulo, SP, Brazil

ARTICLE INFO

Keywords:
Glaucoma
Classification
Pupillary Light Reflex
Computer-aided diagnosis
Diagnostic
Machine learning

ABSTRACT

Glaucoma is a leading cause of irreversible visual field degradation, significantly impacting ocular health. Timely identification and diagnosis of this condition are critical to prevent vision loss. A range of diagnostic techniques is employed to achieve this, from traditional methods reliant on expert interpretation to modern, fully computerized diagnostic approaches. The integration of computerized systems designed for the early detection and classification of clinical indicators of glaucoma holds immense potential to enhance the accuracy of disease diagnosis. Pupillary Light Reflex (PLR) analysis emerges as a promising avenue for glaucoma screening, mainly due to its cost-effectiveness compared to exams such as Optical Coherence Tomography (OCT), Humphrey Field Analyzer (HFA), and fundoscopic examinations. The noninvasive nature of PLR testing obviates the need for disposable components and agents for pupil dilation. This facilitates multiple successive administrations of the test and enables the possibility of remote execution. This study aimed to improve the automated diagnosis of glaucoma using PLR data, conducting an extensive comparative analysis incorporating neural networks and machine learning techniques. It also compared the performance of different data processing methods, including filtering techniques, feature extraction, data balancing, feature selection, and their effects on classification. The findings offer insights and guidelines for future methodologies in glaucoma screening utilizing pupillary light response signals.

1. Introduction

Glaucoma encompasses a group of optic neuropathies characterized by progressive degeneration of Retinal Ganglion Cells (RGCs) and their axons, resulting in visual field loss [1–4]. The loss of visual function due to glaucoma is generally an irreversible process and, if not appropriately treated, can progress to visual impairment and even blindness [1–4]. Furthermore, glaucoma treatment becomes more complex, expensive, and challenging the more advanced the disease is. Therefore, early detection and diagnosis of glaucoma are essential to prevent its progression and minimize its damage [2,4].

The most commonly used tests to diagnose glaucoma include:

- **Tonometry:** Standard procedure for measuring Intraocular Pressure (IOP), being an important indicator in assessing the risk of glaucoma, especially when elevated [5]. Tonometry has the disadvantage of being uncomfortable for some patients due to the need to apply anesthetic eye drops and the contact of the tonometer with the cornea.
- **Campimetry:** Diagnostic tool that maps the patient's visual field, essential for identifying regions of visual loss potentially caused by glaucoma [6]. The disadvantage of this test is its subjectivity, as it depends on the patient's response to the perception of visual stimuli. Furthermore, its effectiveness is limited in children and the elderly and only detects significant damage to the optic nerve, since up to 40% of retinal ganglion cells can be lost before any changes in the visual field can be detected [7].
- **Fundoscopy (Ophthalmoscopy):** This exam allows a detailed inspection of the fundus [8]. It can be performed with a direct or indirect ophthalmoscope, fundus camera, or OCT [9], considered the reference technique for the structural detection of glaucoma. However, these exams require specialized equipment and management by an ophthalmologist, which may not be available or accessible to a portion of the population.

The researchers [1,4,10–25] showed that the pupillary response to light in individuals with glaucoma differs from the pupillary response

* Corresponding author.
E-mail address: hedenirmonteiro@inf.ufg.br (H.M. Pinheiro).

https://doi.org/10.1016/j.array.2024.100359
Received 29 April 2024; Received in revised form 18 June 2024; Accepted 23 July 2024
Available online 2 August 2024

of healthy people and can therefore be used as a potential biomarker of this pathology. Chromatic pupillometry presents an additional alternative in detecting glaucoma, expanding the available exam options for diagnosing this condition [26,27]. It consists of evaluating the contraction and dilation movements of the pupil when faced with luminous stimuli with visible light of different colors for specific periods. The advantage of analyzing the pupillary light reflex is that it can be carried out in a practical and non-invasive way. However, the success rates in detecting early primary open-angle glaucoma have not yet become significant enough to make chromatic pupillometry inserted into clinical practice.

Machine learning algorithms, as they can recognize patterns through data analysis, have been widely used to support medical applications in general to diagnose and screen various pathologies. Therefore, these algorithms can also extract information from data obtained through chromatic pupillometry. Currently, there is not enough research on the effectiveness of algorithms that use pupillometry and analysis of the pupillary light reflex to screen, triage, or diagnose glaucoma, especially in detecting primary angle glaucoma at its initial stage. As a result, the success rates of these methods remain relatively unknown.

Several works published in the last five years have only statistical analyses to show PLR's ability to discriminate between healthy and glaucomatous individuals [1,10,11,21–25,28–30].

The studies [31–34] applied machine learning techniques to diagnose glaucoma based on information extracted from the segmentation of fundus images. Our work differs in evaluating machine learning techniques not on fundus images but on data from chromatic pupillometry, which portrays the volunteers' pupillary light reflex. The work of Quan et al. [35] carried out a proof of concept of using PLR to detect glaucoma. His analysis was limited, however, to statistically testing the significance of the characteristics used in the test.

Our study investigates glaucoma detection methods using machine learning techniques, focusing on improving the analysis of pupil behavior. We explore how strategies for (1) filtering and denoising the pupillary signal, (2) extracting relevant features, (3) performing feature selection, (4) data balancing, and (5) the appropriate choice of a classifier can enhance the analysis of pupillary light reaction in the process of screening, tracking and diagnosing glaucoma.

2. Materials and methods

In this study, we assess various machine learning techniques to determine their efficacy in aiding the diagnosis of glaucoma. The dataset employed comprises examinations from individuals without glaucoma (control group) and from patients with varying degrees of glaucoma. Consequently, we categorize diagnostic scenarios into binary and multi-class classifications to encompass the range of clinically relevant situations.

In our binary classification approach, we organize the data into distinct pairs for targeted analysis:

1. Control vs. Pathological: This classification focuses on distinguishing the control group from the pathological group without considering the pathology's severity.
2. Control vs. Initial Stage: Here, the objective is to precisely differentiate the control group from those in the initial stages of glaucoma.
3. Control vs. Moderate Stage: In this category, we aim to identify differences between the control group and those with moderate glaucoma.
4. Control vs. Severe Stage: This classification is dedicated to contrasting the control group with individuals in the severe stages of glaucoma.

Each category is designed to refine our understanding and detection of glaucoma at various stages of its progression.

In the multi-class classification segment, we categorize the data to discern among multiple classes simultaneously:

5. Control vs. Initial vs. Moderate vs. Severe: This classification aims to differentiate among all possible labels in the database in a singular analysis. It determines whether a patient is healthy or pathological and, if pathological, identifies the specific severity of the condition.
6. Initial vs. Moderate vs. Severe: This approach is designed for scenarios where it is already established that the patient has a pathological condition, but the precise degree of the pathology needs to be ascertained.

Following establishing these diagnostic interest groupings, we implemented several techniques, including filtering, feature extraction, selection, and data balancing. We then evaluated the efficacy of various classifiers. Fig. 1 presents a schematic of the evaluation flow for the machine learning techniques, detailed in subsequent sections.

Upon concluding the study, we aimed to ascertain the most effective technique for each diagnostic scenario.

Table 1
Brief description of the samples available in the database.

Dataset description					
Groups		N	Videos	Age ($\mu \pm \sigma$)	Male
Control		113	243	42 ± 13	49 (43%)
Glaucomatous	Early	108	217	57 ± 12	62 (57%)
	Moderate	21	38	54 ± 14	10 (48%)
	Severe	8	14	61 ± 8	4 (50%)
	Sub. Total	137	270	57 ± 14	76 (55%)
Total		250	512	53 ± 8	125 (50%)

Number of Volunteers (N).

2.1. Pupillary database

The dataset created for this research consists of 512 videos, each lasting 4 min and 5 s, showcasing the pupillary light reflex of 250 volunteers. Among these participants, 113 were healthy, and 137 had glaucoma. Several volunteers contributed more than one recording per eye. The recording adhered to the protocol outlined in Section 2.2. Eligibility criteria for volunteers included having visual acuity better than 20/100 in both eyes, no eye surgeries within the past three months, and being over 18. Participants were also requested to abstain from caffeine consumption for at least an hour before the recording. During the examination, the participants sat comfortably in a completely dark room. They wore the pupillometer and were instructed to focus on a point inside that acted as a visual reference without stimulating their pupils. All participants outside the control group were diagnosed with open-angle glaucoma, the most common form of this condition. The age and sex distribution of the volunteers is detailed in Table 1. This study received ethical approval from the Hospital de Urgências de Goiânia (HUGO) ethics board, under technical advice number 5.990.785, registered on the Brazil platform.

2.2. Video recording protocol and equipment

The recording protocol adopted in this study details the video recording process for analyzing pupillary light reactions. It specifies critical elements such as the initial adaptation time to darkness, the duration, intensity, and color of each light stimulus, the quantity of stimulus applied, the interval of adaptation between stimuli, and which eye will be stimulated and recorded.

The chosen protocol includes an initial dark adaptation period of 10 min, following the guidelines compiled by Pinheiro H. et al. [36] and four stimulation using chromatic LEDs: the first stimulus had a wavelength of 623 nm (red), the second stimulus a wavelength of 466 nm (blue), followed by 517 nm (green) wavelength stimulus, and finishing with white light stimulus. The selection of colors to stimulate pupil

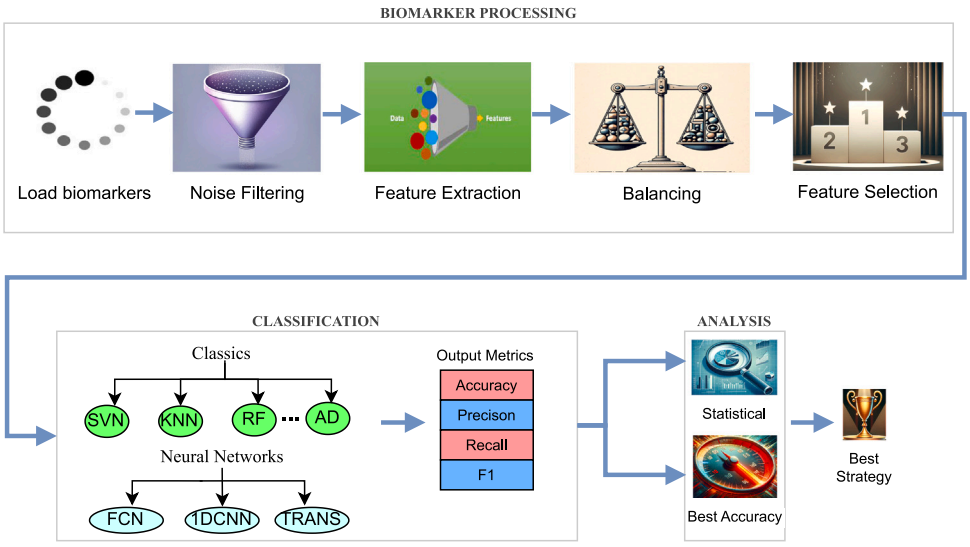


Fig. 1. Proposed method diagram divided into three main steps: biomarkers processing, data classification, and result analysis.

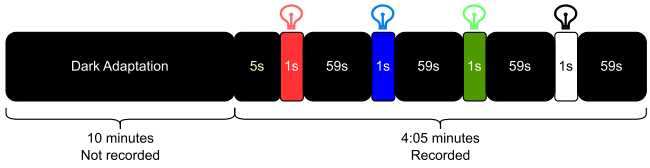


Fig. 2. Illustration of the established protocol, encompassing the initial phase of pupillary adaptation, the color spectrum, and duration of chromatic stimulation, along with the specified intervals for adaptation between each stimulus.

response was guided by the findings presented by Rukmini et al. [37]. They recommended using red and blue light, while Crippa et al. [38] suggested using green light. Both studies highlighted the importance of chromatic pupillometry in assessing the health of photoreceptors in the optic nerve and retina for various diseases. The duration, intensity, and interval of adaptation between each stimulus were set at 1 s, with an intensity of 250 cd/m^2, as established by Park [39]. The adaptation period between stimuli was defined at 59 s, according to the recommendations of Gracitelli et al. [40].

Following the initial adaptation period, the video recording commenced. Five seconds into the recording, the first stimulation was initiated. Each rest and pupil stimulation cycle constituted a signal segment, resulting in four distinct recording phases, as illustrated in Fig. 2. The total duration of each recording was 4 min and 5 s, captured at a rate of 30 frames per second, amounting to a total of 7350 frames.

Stimulation can be applied to the same eye under examination (direct reflex assessment) or to one eye while evaluating the response in the other eye (consensual reflex assessment). In this study, both reflexes were used in the recordings.

The pupillometer used in this study was first introduced by Pinheiro H. et al. [41] and later improved by Silva et al. [42] to incorporate chromatic pupillometry capabilities. This device was chosen based on its ability to meet requirements, such as enabling chromatic stimulation using RGB LEDs, allowing for a customizable duration and intensity of the stimulation, preventing external light from affecting both the recording and the pupillary reflex, providing infrared illumination for recording, which is invisible to the human eye and thus does not impact the pupillary stimulation; offering the flexibility to record and stimulate either eye, enabling the capture of both direct and consensual reflexes.

2.3. Segmentation and preprocessing

In digital image processing and computer vision, image segmentation is partitioning a digital image into multiple segments or objects, often referred to as image regions of interest or pixel sets. The primary goal of segmentation is to simplify the image's representation, making it more meaningful and more straightforward to analyze, as noted by Stockman and Shapiro [43].

Extracting the pupillary signal by measuring the pupil diameter throughout the recording period is essential for Pupillary Light Reflex analysis. This requires accurately locating the pupil and measuring its diameter in each frame. This study used the YOLOv7 [44] convolutional neural network object detector, which was retrained using a dataset of 10,000 hand-labeled pupil images for accurate pupil measurement.

Following the retraining process, the neural network exhibited enhanced proficiency in pupil detection, successfully identifying and delineating the pupil's bounding box in each video frame. The dimensions of this bounding box enabled the calculation of the pupil's pixel diameter: its height representing the vertical diameter and its width the horizontal diameter. Based on the approach recommended by Zandi et al. [45], the larger of the two diameters was designated as the definitive measurement of the pupil's diameter.

Measuring the pupillary diameter in each frame yields a pupillary signal that illustrates the pupil's dynamic behavior and diameter fluctuations over time. Fig. 3 displays a complete pupillary signal from a volunteer.

There are several phases involved in conducting a pupillary examination. First, the volunteer's video is recorded. Then, the video is analyzed to locate the pupil and measure its diameter for each frame. Next, a curve or signal is created to represent the pupillary behavior. Once this is done, the pupillary reflex is assessed. The process of assessing the reflex involves preprocessing, filtering, feature extraction, feature selection, balancing, and signal classification.

In the preprocessing stage, 24 s were excluded from each section's post-stimulation and post-redilation phases to minimize signal redundancy.

2.4. Filtering

The pupillary signal can be affected by noise due to blinks or shifts in the volunteers' gaze direction. Despite instructions to maintain focus on a specific point on the pupillometer and minimize blinking, blinks and gaze deviations were still observed.

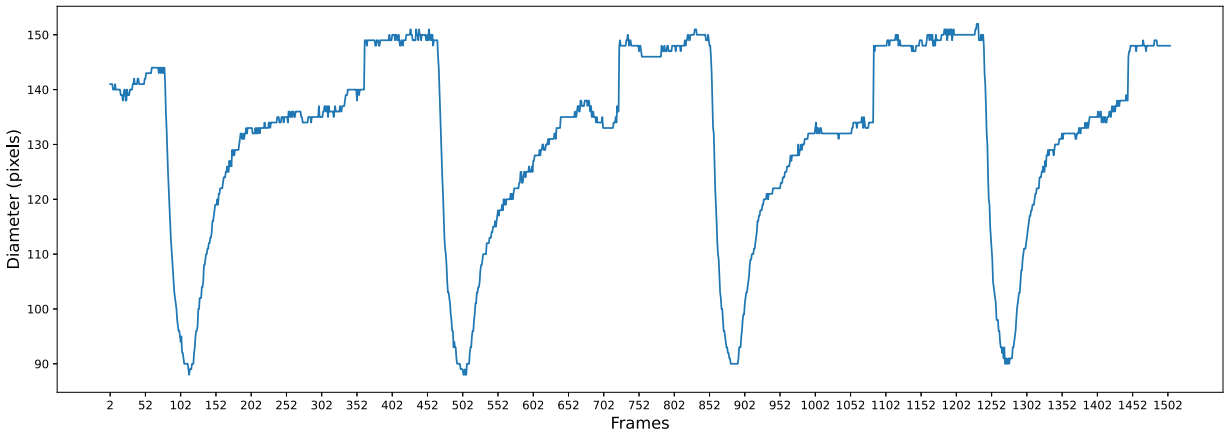


Fig. 3. Segmented pupillary signal, after removing redundant periods.

We conducted a study to determine the effectiveness of three methods for filtering out extraneous noise in signal classification. The first method, No Filtering (NF), analyzes the signal without interference. The second method, called Custom Filtering (CF), uses a custom algorithm to remove outlier readings that exceed the limits established by pupillary physiology and eliminate sudden fluctuations in the signal. The third method, Low-Pass Filtering (LP), attenuates high-frequency noise components. We tested each method to see how well it improved the accuracy and reliability of signal classification.

We excluded the moving average filter from our analysis. It proved ineffective in noise removal with windows smaller than 20 frames and distorted the signal with more extensive window settings.

Custom-designed filter to deal with sudden signal changes that exceed a predetermined threshold. Initially, the threshold is set at a low value of 3, and any changes in the signal that exceed this threshold are removed, except when their removal would result in discontinuity. The filter is programmed to avoid removing more than 20 consecutive frames. If we observe any discontinuity, we adjust the threshold. We incrementally increase the threshold by 1 unit and then reapply the filtering process. We repeat this iterative procedure until we achieve a version of the signal free of sudden changes and discontinuities. You can find more details on this process in the pseudocode outlined in Appendix A.

The low-pass filter is sourced from the SciPy signal library and utilizes the 'filfit' function, which is notable for not inducing phase shifts in the signal. It accomplishes this by applying the filter once forward and once in reverse. This filter was set up with a cutoff frequency of 0.6, an order of 3, and a Nyquist frequency (Nqy) calculated as 0.5 times the recording frame rate, 30 fps in this study.

2.5. Feature extraction

Feature extraction transforms the input data into a more compact subspace, retaining the essential information from the original dataset [46]. This technique is particularly effective in enhancing the performance of classification algorithms, especially compared to using raw data directly.

In our study, we analyzed the pupillary signal in two ways. Firstly, we considered the complete set of diameters without any feature extraction. Secondly, we extracted ten features that have been identified in the literature. These features were derived from the four signal stimulation phases, resulting in 40 features. Additionally, two features were extracted from the complete signal, bringing the total number of features to 42.

The following outlines the extracted features from each signal stimulation section:

1. Initial Diameter (ID): The average pupil diameter at rest before the light stimulus, typically the maximum diameter. Calculated using the average of twenty frames 0.3 s before stimulation.
2. Maximum Contraction (MC): The minor diameter observed in the section, usually during stimulus application. Calculated by identifying the smallest diameter post-stimulus, excluding values under 40 pixels.
3. Time to Maximum Contraction (TMC): The duration is taken to reach maximum contraction post-stimulus onset. Determined by the frame number at which the most minor diameter occurs.
4. Absolute Contraction Amplitude (ACA): The difference between the largest (initial diameter) and smallest (maximum contraction) diameters. Calculated by subtracting MC from ID.
5. Relative Contraction Amplitude (RCA): The smallest to largest diameter ratio. Calculated by dividing MC by ID.
6. Contraction Speed (CS): Amplitude to time ratio for maximum contraction. Calculated by dividing ACA by TMC.
7. Latency (LAT): The time between stimulus onset and contraction effective start. Determined by the first instance where the diameter reduces by 5% from ID.
8. Diameter of Re-dilation after 6 s (DREL6s): Pupil size 6 s post-stimulus. Calculated from the first valid diameter in a tolerance 10-diameter window post-stimulus; −1 assigned if no valid value was found inside the tolerance window.
9. Re-dilation Time (RELT): Time for the pupil to return to 80% of the initial diameter post-stimulus. Calculated by finding the frame position exceeding 80% of ID.
10. Higuchi Fractal Complexity (HFC): A nonlinear measure of fractal dimension in time series. Calculated over 20 frames, 0.3 s before stimulation, discarding invalid values and analyzing up to 5 time series scales (kmax). Refer to Ngo et al. [47] for detailed insights on Higuchi fractal complexity and its significance in differentiating between control and glaucomatous patients.

Below is a description of the characteristics applied to the complete signal:

11. Diameter at the End of the Exam (DEE): represents the pupil size at the end of the exam. We calculated this as the average of the last ten diameters of each curve, discarding diameters smaller than 40 pixels.
12. Higuchi Fractal Complexity in Signal (HFCS): is the Higuchi complexity, in this case, applied across the entire pupillary curve. We calculate the complexity by considering the curve as a whole, disregarding impossible diameters, and, as in the analysis per section, defining the parameter K-max = 5.

With feature extraction from the pupillary signal complete, the following steps involve balancing the samples and selecting the most compelling features for further analysis.

2.6. Balancing

An imbalance in a two-class dataset occurs when the minority class is significantly underrepresented compared to the majority class. Our analyses define an imbalance when the difference in class proportions exceeds 10%. Balancing is deemed unnecessary if the sample difference percentage does not surpass this threshold.

We employed both undersampling and oversampling techniques to achieve balance in cases where class imbalance was identified. Undersampling involves randomly removing samples from the majority class to align their size with that of the minority class. Conversely, oversampling involves augmenting the minority class with synthetic samples (created from existing samples) until the class sizes are equivalent. For oversampling, we used the SMOTE algorithm (Synthetic Minority Over-sampling Technique), setting the number of neighbors k to 5. SMOTE is a data augmentation method that resamples by considering the neighborhood of a sample and generating a new instance based on this proximity [48].

2.7. Feature selection

Feature selection (FS) is a process aimed at identifying pertinent features while eliminating irrelevant, redundant, or noisy data. Irrelevant features fail to contribute meaningful information, whereas redundant features offer no additional insights beyond what is already obtained from the selected features [49]. The purpose of feature selection is threefold: (1) to enhance classification accuracy by preventing overfitting, (2) to develop more streamlined models, and (3) to increase the interpretability of models for human understanding. Feature selection methods fall into three main categories: filter, wrapper, and embedded methods. Filter methods, such as SelectKBest, utilize general criteria like correlation to discard irrelevant features independently of machine learning algorithms. Wrapper methods use classifiers to gauge performance and choose the optimal feature combination. Embedded methods integrate feature selection directly into the machine learning process.

In our research, we utilized three different feature selection methods.

Firstly, we used the SelectKBest filter-based method, which evaluates the individual relationship between each feature and the output variable using a specific statistical test (analysis of variance—ANOVA in this study) to select the top k attributes.

Secondly, we employed Recursive Feature Elimination (RFE), a wrapper-based technique. The RFE method is an iterative approach that starts with all features and progressively removes the least significant ones based on model performance. Random Forest is used in this study until the targeted number of features is achieved.

Finally, we used LassoCV, an embedded method. LassoCV applies the Lasso linear model, which utilizes iterative fitting along a regularization path during training. It employs L1 (Lasso) regularization to identify and select significant features effectively, enhancing model performance and interpretability.

The feature selection process was crafted to identify the top twenty features. This was explicitly implemented for the SelectKBest and RFE methods. Unlike these, the LassoCV selector was set with a tolerance level of 0.01, enabling it to select a flexible number of features. The twenty best coefficients were used in the second step to choose the 20 best features. Following the completion of feature selection from the pupillary signal data, the subsequent phase involves tackling the challenge of sample balancing.

2.8. Crop signal

In signal processing, “cropping a signal” refers to identifying and isolating a specific segment from a complete signal for further analysis or use. This is similar to cropping an image, but instead of selecting a portion of an image, we choose a portion of a time series or waveform. This can isolate a specific time interval or remove unwanted signal parts for analysis or processing. In this research, we will divide the signal into four segments corresponding to the color of stimulation: red, blue, green, and white, as defined in the recording protocol.

This approach aimed to reduce the total number of features by treating this division as a type of feature selection. As a result, the number of features without extraction decreased from 1500 to 375, and with feature extraction, it was further reduced to just 10. This analysis excluded the DEE and HFCS features, which refer to the complete signal.

2.9. Classifiers

To evaluate the effectiveness of previously utilized techniques in glaucoma classification, we employed the following classifiers:

(A) Linear Classifier:
Linear Discriminant Analysis (LDA): Identifies a linear combination of features that best differentiates two or more classes.

(B) Neighbor-based Classifier:
K-Nearest Neighbors (KNN): This method classifies an input by considering its closest neighbors, where “K” denotes the number of neighbors used.

(C) Tree-based Classifiers:
Decision Tree (DT): Employs a decision tree structure for classification.
Random Forest (RF): Utilizes an ensemble of decision trees.
Extra Trees (ET): A variant of Random Forest, where splits at each tree node are chosen randomly rather than for maximal purity.

(D) Boosting-based Classifiers:
AdaBoost (AB): Integrates multiple ‘weak’ classifiers to enhance classification strength.
Gradient Boosting (GB): Improves classification by adjusting for the residuals of previous models.

(E) Probability-based Classifier:
Naive Bayes (NB): Applies Bayes’ theorem with the feature independence assumption.

(F) Support Vector Based Classifier:
Support Vector Machine (SVM): This technique utilizes kernel functions (linear, polynomial, RBF, sigmoid, etc.) to identify optimal class separation boundaries.

(G) Neural Network-based Classifiers:
Fully Connected Neural Networks (FCN): These networks feature neurons in each layer that are fully connected to all neurons in the subsequent layer. Due to the complete interlayer connections, they are known as dense.
One-dimensional Convolutional Neural Networks (1D-CNN) are designed for data with a grid-like structure, like time series (1D). They effectively identify local patterns in datasets like audio, financial time series, and genetic sequencing.
Transformers Neural Networks (TRANS): Introduced by Vaswani et al. in “Attention is All You Need” (2017), these networks use ‘self-attention’ allowing each part of a sequence to interact with every other part, capturing complex interrelationships. Their primary innovation is the attention mechanism, which assesses the relevance of different words or features in the context of a specific word or feature.

Traditional classifiers were sourced from the Python scikit-learn library, while artificial neural networks were developed using the TensorFlow framework. The architectural diagram representations of the neural networks are shown in Appendix B.

Table 2
Main parameters used by the classifiers in the pupil signal classification process.

Classifiers parameters	
Classifier	Parameters
LDA	solver: {'svd', 'lsqr', 'eigen'}, used='svd', n_components: int, used=None
KNN	n_neighbors: int, used=5, weights: {'uniform', 'distance'}, used='uniform', algorithm: {'auto', 'ball_tree', 'kd_tree', 'brute'}, used='auto', metric: string, used='minkowski'
DT	criterion: {'gini', 'entropy'}, used='gini', splitter: {'best', 'random'}, used='best', max_depth: int, used=None
RF	n_estimators: int, used=100, criterion: {'gini', 'entropy'}, used='gini', max_depth: int, used=None
ET	n_estimators: int, used=100, criterion: {'gini', 'entropy'}, used='gini' max_depth: int, used=None
AB	base_estimator: object, used=None, n_estimators: int, used=50 learning_rate: float, used=1.0
GBM	loss: {'deviance', 'exponential'}, used='deviance' learning_rate: float, used=0.1, n_estimators: int, used=100
NB	–
SVM	C: float, used=1.0, kernel: {'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'}, used='rbf', degree: int, used=3, gamma: {'scale', 'auto'}, used='scale'
FCN	epochs=100, batch_size=8, dropout=0.25, callbacks=early_stop
1D-CNN	epochs=100, batch_size=8, filters=64, dropout=0.25, callbacks=early_stop
TRANS	epochs=100, batch_size=8 num_heads=4, head_size=64 ff_dim=32, num_transformer_blocks=4 mlp_units={64} dropout=0.25, mlp_dropout=0.4, callbacks=early_stop

Linear Discriminant Analysis (LDA). k-Nearest Neighbors (KNN). Decision Tree (DT). Random Forest (RF). Extra Tree (ET). AdaBoost (AB). Gradient Boosting Machine (GBM). Naive Bayes (NB). Support Vector Machine (SVM). Fully Connected Neural Network (FCN). One Dimensional Convolutional Neural Network (1D-CNN). Transformer Neural Network (TRANS).

2.10. Classification

The study involved a thorough evaluation that classified six approaches to grouping data, detailed in Section 2. Three filtering methods and two feature extraction strategies (with and without feature extraction) were also evaluated, which resulted in 36 separate classification scenarios. Four distinct feature selection techniques were examined for each scenario: No Feature Selection, SelectKBest, Recursive Feature Elimination, and Lasso.

We then applied three different filtering approaches for each comparison type: no filter, a customized filter, and a low-pass filter. Initially, we analyzed the entire pupillometry curve diameters without feature extraction. Later, we extracted 42 features from the signal for further analysis. Two data balancing methods, namely undersampling and oversampling, were utilized.

This diverse approach was designed to uncover complex patterns in pupillometry data, guiding toward techniques that produce high accuracy in glaucoma classification. Table 2 presents the main parameters used by each classifier evaluated in the research.

2.11. Validation

Cross-validation using the K-fold partitioning method (CV) is a widely recognized technique for assessing the generalizability of machine learning models. This approach involves dividing the dataset into K roughly equal, mutually exclusive subsets (folds). These folds are then utilized in K rounds of model training and evaluation. K−1 subsets are used for training each round, while the remaining subset serves as the test set. This cycle is repeated until each subset has been the test set once. Upon completion of these iterations, the performance results from each fold are aggregated to yield an average measure of model effectiveness [50].

In our study, we opted for a k-value of 5 in the cross-validation process for classifiers. This choice strikes a balance between the size of the training and test data sets while offering reduced computational demands compared to larger k-values. To ensure data integrity and prevent information leakage between the training and testing phases, we grouped user videos such that all videos from a single volunteer were exclusively assigned to either the training or the testing group but never to both simultaneously.

2.12. Evaluation of results

To discern the impact of different methodologies on classification outcomes and identify the most effective approach, we conducted two evaluations for each classifier: a statistical assessment and an analysis based on optimal accuracy.

2.12.1. Statistical assessment

Upon completing the classification phase, we performed statistical analyses of the results yielded by the classifiers. For determining the normality of result distributions, we utilized the Shapiro–Wilk test, which is particularly effective for small datasets like ours, as opposed to the Kolmogorov–Smirnov test, which is better suited for larger datasets [51]. When the classifier results followed a normal distribution, we applied the paired Student's t-test [52] for analysis. Conversely, for non-normally distributed results, we employed the Wilcoxon test [53], with a significance level of 0.05.

2.12.2. Evaluation by best accuracy

While statistical analysis provides an average performance overview of various classifiers, it may not fully recognize the effectiveness of a singular approach that performs exceptionally well, especially if other classifiers do not yield similarly impressive results.

Thus, we supplemented the statistical analysis with an assessment highlighting the classifier that achieved the highest accuracy, aiming to pinpoint the most effective model in each employed approach.

This dual analysis method allows us to evaluate the consistency across multiple classifiers and spotlight the top-performing one. Our goal is to offer a comprehensive insight into the machine learning techniques used in this study.

2.13. Evaluation metrics

To provide a comprehensive evaluation, we considered metrics: *Accuracy (ACC)*, reflecting the proportion of correctly identified cases out of the total cases examined; *sensitivity (True Positive Rate — TPR)*, which quantifies the proportion of actual positive cases (individuals with glaucoma) correctly identified as such; *Specificity (True Negative Rate — TNR)*, measuring the percentage of actual negative cases (healthy control patients) correctly identified; and *F1-score*, a metric that combines sensitivity and precision into a single measure, representing the harmonic mean of sensitivity and specificity.

Table 3
Statistical analysis of the filtering methodologies highlighting the most effective approach or identifying the equivalents for each grouping.

			Filtering							
Grouping	Balancing	Extraction	NS	SB	RFE	LA	Summarization (freq.)			
			Best filtering strategy				NF	CF	LP	EQ
Control vs Pathological	Under	No	EQ	NF	NF	NF	3	0	0	1
		Yes	NF	NF	NF	EQ	3	0	0	1
	Over	No	EQ	NF	NF	CF	2	1	0	1
		Yes	NF	CF	NF	NF	3	1	0	0
	SubTotal:						11	2	0	3
Control vs Early	Under	No	NF	NF	NF	NF	4	0	0	0
		Yes	CF	EQ	EQ	EQ	0	1	0	3
	Over	No	NF	NF	NF	NF	4	0	0	0
		Yes	NF	NF	NF	CF	3	1	0	0
	SubTotal:						11	2	0	3
Control vs Moderate	Under	No	NF	NF	CF	CF	2	2	0	0
		Yes	EQ	NF	NF	LP	2	0	1	1
	Over	No	NF	CF	NF	LP	2	1	1	0
		Yes	LP	LP	LP	LP	4	0	0	0
	SubTotal:						10	3	2	1
Control vs Severe	Under	No	NF	NF	NF	CF	3	1	0	0
		Yes	CF	CF	CF	CF	0	4	0	0
	Over	No	NF	NF	NF	NF	4	0	0	0
		Yes	LP	LP	LP	LP	0	0	4	0
	SubTotal:						7	5	4	0
Control vs Early vs Moderate vs Severe	Under	No	CF	CF	EQ	CF	0	3	0	1
		Yes	NF	NF	NF	NF	4	0	0	0
	Over	No	LP	LP	LP	LP	4	0	0	0
		Yes	NF	NF	LP	LP	2	2	0	0
	SubTotal:						10	5	0	1
Early vs Moderate vs Severe	Under	No	NF	NF	NF	NF	4	0	0	0
		Yes	CF	CF	NF	CF	1	3	0	0
	Over	No	NF	NF	NF	LP	3	1	0	0
		Yes	LP	LP	LP	LP	0	0	4	0
	SubTotal:						8	4	4	0
Total:							**49**	**18**	**20**	**9**

Under (Undersampling), Over (Oversampling), No Feature Selection (NS), SelectKBest (SB), Recursive Feature Elimination (RFE), Lasso Feature Selection (LA), No Filtering (NF), Custom Filtering (CF), Low Pass Filter (LP), Equivalents (EQ), Frequency of best strategy (freq.).

3. Results

In our study, we undertook a comprehensive exploratory analysis of the pupillary signal encompassing (1) three distinct filtering techniques — no filtering, filtering with a customized filter, and using a low pass filter; (2) two methods of data balancing — undersampling and oversampling; (3) two approaches to feature extraction — no feature extraction and extracting forty-two mentioned in the literature features; and (4) four strategies for feature selection — no feature selection, and using SelectKbest, RFE, and LassoCV feature selectors. We then processed the data using nine conventional classifiers: Support Vector Machines, K-Nearest Neighbors, Decision Tree, Random Forest, Extra Tree, Linear Discriminant Analysis, Gaussian Naive Bayes, AdaBoost, and Gradient Boosting Classifier. Additionally, the data was subjected to three advanced classifiers based on neural network architectures, including Fully Connected, Convolutional, and Transformer models.

The findings from our investigation are compiled and analyzed statistically, considering the performance metrics of all classifiers and exclusively focusing on the best accuracy achieved. In the following sections, we will detail the outcomes of this statistical analysis, emphasizing the effects and efficacy of signal filtering, feature extraction, data balancing, and feature selection.

3.1. Analysis of filtering techniques:

The pupillary signal underwent classification through three distinct filtering methodologies: (1) without any filtration of the pupillary signal, (2) using a specific filter designed to eliminate abrupt changes in the pupillary signal, and (3) applying a low-pass filter to the pupillary signal.

Table 3 presents a comparative analysis of these filtering strategies. It outlines instances where the three filtering approaches were significantly equivalent or superior to the others, considering six data grouping methods and both undersampling and oversampling data balancing techniques. The comparison is made for scenarios with and without feature extraction, considering feature selection techniques.

3.2. Evaluating feature extraction efficacy:

This segment of the analysis focused on determining the most effective approach for classifying the pupillary signal: whether extracting specific, literature-identified features from the pupillary signal enhances classification performance or supplying the classifiers with the complete spectrum of pupillary diameters without prior data synthesis.

Table 4 provides a statistical breakdown, based on classification metrics, of instances where utilizing the full range of diameters was

Table 4
Statistical analysis of feature extraction methodologies highlighting the most effective approach or identifying the equivalents for each grouping.

			Extraction					
Grouping	Balancing	Filtering	NS	SB	RFE	LA	Summarization	
			Best extraction strategy				Strategy	Freq.
Control vs Pathological	Under	NF	Yes	No	No	No	No	7
		CF	EQ	EQ	No	No	Yes	2
		LP	Yes	EQ	No	No	EQ	3
	Over	NF	Yes	No	No	No	No	7
		CF	EQ	EQ	No	No	Yes	2
		LP	Yes	EQ	No	No	EQ	3
Control vs Early	Under	NF	Yes	EQ	No	No	No	7
		CF	Yes	No	No	No	Yes	4
		LP	Yes	Yes	No	No	EQ	1
	Over	NF	No	No	No	No	No	7
		CF	EQ	EQ	No	No	Yes	3
		LP	EQ	Yes	No	Yes	EQ	2
Control vs Moderate	Under	NF	Yes	EQ	Yes	EQ	No	1
		CF	EQ	EQ	No	EQ	Yes	6
		LP	Yes	Yes	Yes	Yes	EQ	5
	Over	NF	No	Yes	No	Yes	No	6
		CF	EQ	Yes	No	No	Yes	5
		LP	No	Yes	Yes	No	EQ	1
Control vs Severe	Under	NF	Yes	Yes	EQ	Yes	No	0
		CF	Yes	Yes	Yes	Yes	Yes	10
		LP	Yes	Yes	Yes	EQ	EQ	2
	Over	NF	No	Yes	No	No	No	3
		CF	Yes	Yes	Yes	Yes	Yes	8
		LP	Yes	Yes	EQ	Yes	EQ	1
Control vs Early vs Moderate vs Severe	Under	NF	EQ	No	No	No	No	7
		CF	Yes	Yes	EQ	EQ	Yes	2
		LP	No	No	No	No	EQ	3
	Over	NF	EQ	Yes	No	EQ	No	1
		CF	Yes	Yes	Yes	Yes	Yes	7
		LP	EQ	Yes	EQ	Yes	EQ	4
Early vs Moderate vs Severe	Under	NF	No	EQ	No	No	No	5
		CF	Yes	Yes	Yes	Yes	Yes	5
		LP	EQ	Yes	No	No	EQ	2
	Over	NF	No	Yes	No	Yes	No	2
		CF	Yes	Yes	Yes	Yes	Yes	10
		LP	Yes	Yes	Yes	Yes	EQ	0
Total:							**No**	**53**
							Yes	**64**
							EQ	**27**

Under (Undersampling), Over (Oversampling), No Filtering (NF), Custom Filtering (CF), Low Pass Filtering (LP), No Feature Selection (NS), SelectKBest Feature Selection (SB), Recursive Feature Elimination (RFE), Lasso Feature Selection (LA), Extracting is the best option (Yes), No extraction is the best option (No), Statistical Equivalents (EQ), Frequency of best strategy (Freq.).

either equivalent to or more effective than conducting feature extraction from the pupillary signal. The table concludes with a summary indicating the number of times feature extraction was statistically superior, the number of times not extracting was better, and the number of times the results were equivalent.

3.3. Analysis of balancing techniques:

This part of our research aimed to explore the impact of data balancing on the classification process. We specifically examined how the application of balancing techniques influences the classification outcomes.

Table 5 presents a comprehensive comparison, detailing the instances in which the oversampling and undersampling techniques—applied to data both with and without feature extraction and across various feature selection methods—proved to be equivalent or superior. These findings are essential for providing a benchmark for subsequent comparative analyses.

3.4. Evaluating the impact of feature selection

In this analysis, our objective was to understand the influence of Feature Selection on the classification process. Specifically, we aimed to determine whether performing feature selection before classification enhances results and, if so, which techniques demonstrate superior performance.

Table 6 details instances where a feature selection technique outperformed others, considering both undersampling and oversampling balancing methods and contexts with or without feature extraction.

3.5. Best accuracy analysis

We evaluated the best-performing classifier, which could highlight important performance within a particular methodology. Analyzing the top classifier strengthens conclusions drawn from statistical analysis.

Table 8 presents the best classification accuracy achieved in our study, along with the specific approach that contributed to this optimal outcome.

Table 5
Statistical assessment of oversampling and undersampling balancing techniques highlighting the most effective approach or identifying the equivalents for each grouping.

			Balancing					
Grouping	Extraction	Filtering	NS	SB	RFE	LA	Summarization	
			Best balancing strategy				Strategy	Freq.
Control vs Pathological	No	NF	Over	EQ	Over	Under	Over	4
		EF	Over	EQ	EQ	EQ	Under	1
		LP	Over	Over	Over	EQ	EQ	7
	Yes	NF	EQ	Over	EQ	EQ	Over	4
		EF	Over	Over	Over	Under	Under	2
		LP	EQ	EQ	Over	Under	EQ	5
Control vs Early	No	NF	Over	Over	Over	Under	Over	8
		EF	Over	Over	Over	EQ	Under	1
		LP	Over	EQ	Over	EQ	EQ	3
	Yes	NF	EQ	Over	Over	Over	Over	7
		EF	Over	Over	EQ	Under	Under	1
		LP	Over	Over	EQ	EQ	EQ	4
Control vs Moderate	No	NF	Over	Over	Over	Over	Over	12
		EF	Over	Over	Over	Over	Under	0
		LP	Over	Over	Over	Over	EQ	0
	Yes	NF	Over	Over	Over	Over	Over	12
		EF	Over	Over	Over	Over	Under	0
		LP	Over	Over	Over	Over	EQ	0
Control vs Severe	No	NF	Over	Over	Over	Over	Over	12
		EF	Over	Over	Over	Over	Under	0
		LP	Over	Over	Over	Over	EQ	0
	Yes	NF	Over	Over	Over	Over	Over	12
		EF	Over	Over	Over	Over	Under	0
		LP	Over	Over	Over	Over	EQ	0
Control vs Early vs Moderate vs Severe	No	NF	Over	Over	Over	Over	Over	12
		EF	Over	Over	Over	Over	Under	0
		LP	Over	Over	Over	Over	EQ	0
	Yes	NF	Over	Over	Over	Over	Over	12
		EF	Over	Over	Over	Over	Under	0
		LP	Over	Over	Over	Over	EQ	3
Early vs Moderate vs Severe	No	NF	Over	Over	Over	Over	Over	12
		EF	Over	Over	Over	Over	Under	0
		LP	Over	Over	Over	Over	EQ	0
	Yes	NF	Over	Over	Over	Over	Over	12
		EF	Over	Over	Over	Over	Under	0
		LP	Over	Over	Over	Over	EQ	0
Total:							**Over**	**122**
							Under	**5**
							EQ	**17**

Under (Undersampling), Over (Oversampling), Lasso Feature Selection (LA), Recursive Feature Elimination (RFE), SelectKBest Feature Selection (SB), No Feature Selection (NS), Statistical Equivalents (EQ), Frequency of the best strategy (Freq.).

3.6. Assessing the signal after cropping

This analysis segments the signal into four distinct parts, each corresponding to a stimulation period, to ascertain if a specific signal section is more prominent for classification purposes. Furthermore, the analysis aims to identify the most effective stimulation color for enhancing glaucoma screening. Considering the applied classification methodology, the results for each segment are presented in Table 7.

3.7. Implementing the most promising classification technique

After evaluating various machine learning methods, we re-applied classification to the Control vs. Pathological data group, utilizing the approach identified as the most promising according to statistical analysis and by best accuracy. There was consensus among both analyses that using unfiltered data without feature extraction was the best approach. However, regarding the technique for feature selection, the statistical analysis pointed to the use of the RFE technique. In contrast, the analysis for the best accuracy with the entire signal indicated using the LassoCV (LA) technique. Both techniques were tested again, and the selection using LA combined with the LDA classifier proved to be the most promising.

This Control vs. Pathological data grouping was selected in this final analysis for its broad scope, effectively distinguishing between healthy individuals and those with any degree of glaucoma. While this approach does not discern the specific degree of glaucoma, it aligns with our primary objective of developing an effective screening tool, where detailed gradation is initially less critical.

Fig. 4 displays a violin plot illustrating the distribution of the 20 highest-ranked features from the “Control vs. Pathological” group, as identified by the LassoCV selector. The values of these features have been normalized using the Z-score method. The features correspond to specific positions on the pupillary signal, identified as follows: (a) p335, (b) p1448, (c) p1270, (d) p708, (e) p31, (f) p1369, (g) p1102, (h) p172, (i) p1069, (j) p348, (k) p1437, (l) p1080, (m) p176, (n) p1452, (o) p710, (p) p28, (q) p247, (r) p1393, (s) p496, (t) p1403. A violin plot combines the elements of a box plot with a density estimate, providing an overview of the data density throughout the entire range of each feature. This visualization enables the observation of data concentration and variability for each feature, categorized by

Table 6
Statistical feature selection technique analysis highlights the most effective approach for each grouping.

Feature selection								
Grouping	**Filter**	**Extraction**	**Filtering**	**Best feature selection**	**Summarization (Freq.)**			
					NS	**SB**	**RFE**	**LA**
Control vs Pathological	Under	No	NF	LA	0	0	0	3
			CF	LA				
			LP	LA				
		Yes	NF	RFE	0	0	2	1
			CF	RFE				
			LP	LA				
	Over	No	NF	RFE	0	0	3	0
			CF	RFE				
			LP	RFE				
		Yes	NF	RFE	0	1	1	1
			CF	SB				
			LP	RFE				
Control vs Early	Under	No	NF	RFE	0	0	2	1
			CF	LA				
			LP	RFE				
		Yes	NF	RFE	0	0	3	0
			CF	RFE				
			LP	RFE				
	Over	No	NF	RFE	0	0	3	0
			CF	RFE				
			LP	RFE				
		Yes	NF	SB	0	1	1	1
			CF	RFE				
			LP	SB				
Control vs Moderate	Under	No	NF	LA	0	0	1	2
			CF	RFE				
			LP	RFE				
		Yes	NF	RFE	0	0	1	2
			CF	LA				
			LP	LA				
	Over	No	NF	NS	1	0	0	2
			CF	LA				
			LP	LA				
		Yes	NF	LA	0	0	2	1
			CF	RFE				
			LP	RFE				
Control vs Severe	Under	No	NF	RFE	0	0	3	0
			CF	RFE				
			LP	RFE				
		Yes	NF	RFE	1	0	2	0
			CF	RFE				
			LP	NS				
	Over	No	NF	RFE	0	0	1	2
			CF	LA				
			LP	LA				
		Yes	NF	NS	1	0	1	1
			CF	LA				
			LP	RFE				
Control vs Early vs Moderate vs Severe	Under	No	NF	RFE	0	1	2	0
			CF	RFE				
			LP	SB				
		Yes	NF	RFE	0	2	1	0
			CF	SB				
			LP	SB				
	Over	No	NF	RFE	0	0	0	3
			CF	RFE				
			LP	RFE				
		Yes	NF	NS	3	0	0	0
			CF	NS				
			LP	NS				
			NF	RFE				
			CF	RFE				

(*continued on next page*)

Table 6 (*continued*).

Feature selection								
Grouping	**Filter**	**Extraction**	**Filtering**	**Best feature selection**	**Summarization (Freq.)**			
					NS	**SB**	**RFE**	**LA**
Early vs Moderate vs Severe	Under	No	LP	RFE	0	0	3	0
		Yes	NF	RFE	0	1	2	0
			CF	RFE				
			LP	SB				
	Over	No	NF	RFE	1	0	2	0
			CF	RFE				
			LP	NS				
		Yes	NF	RFE	2	0	1	0
			CF	NS				
			LP	NS				
Total					**9**	**6**	**37**	**20**

Under (Undersampling), Over (Oversampling), Yes Extraction (Yes), No Extraction (No), No Filtering (NF), Custom Filtering (CF), Low Pass Filtering (LP), Lasso Feature Selection (LA), Recursive Feature Elimination (RFE), SelectKBest Feature Selection (SB), No Feature Selection (NS), Frequency of best strategy (Freq.).

Table 7
Best classification accuracies achieved for each grouping by interest and approach utilizing a specific signal segment associated with a stimulation color.

Best accuracy — segmented signal						
Grouping	Approach					Best accuracy
	Filtering	Extraction	Segment	Balancing	Classifier	
Control vs Pathological	LP	Yes	White	Over	KNN	0.6488
Control vs Early	LP	No	Blue	Over	ET	0.6576
Control vs Moderate	NF	No	Green	Over	ET	0.9493
Control vs Severe	NF	No	Blue	Over	ET	0.9704
Control vs Early vs Moderate vs Severe	NF	No	Red	Over	GBM	0.7631
Early vs Moderate vs Severe	NF	No	Blue	Over	ET	0.9268

No Filtering (NF), Low Pass Filtering (LP), No Extraction (No), Yes Extraction (Yes), Over (Oversampling), k-Nearest Neighbors (KNN), Gradient Boosting Machine (GBM), Extra Tree (ET).

Table 8
Best accuracies achieved by classifiers grouped by interest and approach using full signal.

Best accuracy — full signal						
Grouping	Approach					Best accuracy
	Filtering	Extraction	Selection	Balancing	Classifier	
Control vs Pathological	NF	No	LA	Under	LDA	0.7390
Control vs Early	NF	No	LA	Under	NB	0.7297
Control vs Moderate	NF	No	NS	Over	SVM	0.9810
Control vs Severe	NF	No	RFE	Over	RF	0.98733
Control vs Early vs Moderate vs Severe	NF	No	NS	Over	ET	0.7921
Early vs Moderate vs Severe	NF	No	NS	Over	ET	0.9221

No Filtering (NF), No Extraction (No), No Feature Selection (NS), Recursive Feature Elimination (RFE), Lasso Feature Selection (LA), Under (Undersampling), Over (Oversampling), Linear Discriminant Analysis (LDA), Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Extra Tree (ET).

class. It is noticeable that the features are not entirely symmetric; these differences between classes can be instrumental in data classification.

Additionally, Fig. 5 presents a heatmap illustrating the correlation among the 20 highest-ranked features, comparing the use of both the LassoCV selector (A) and RFE (B). In this heatmap, darker quadrants are preferable, as they indicate reduced interdependence between the features. Comparing heatmap A (LassoCV) with heatmap B (RFE) shows that heatmap A is slightly darker than B, which helps explain the better accuracy achieved using this feature selector in the analysis by best accuracy.

The LassoCV feature selector identified the following key features from the pupillary signal data: (a) p335, (b) p1448, (c) p1270, (d) p708, (e) p31, (f) p1369, (g) p1102, (h) p172, (i) p1069, (j) p348, (k) p1437, (l) p1080, (m) p176, (n) p1452, (o) p710, (p) p28, (q) p247, (r) p1393, (s) p496, (t) p1403. In contrast, the Recursive Feature Elimination (RFE) selector pinpointed these features as most significant: (a) p98, (b) p178, (c) p343, (d) p349, (e) p350, (f) p367, (g) p403, (h) p498, (i) p499, (j) p713, (k) p854, (l) p1073, (m) p1105, (n) p1272, (o) p1469, (p) p1475, (q) p1489, (r) p1490, (s) p1491, (t) p1492. Each feature set emphasizes the variations in selection criteria and results between the two methodologies. Selecting less correlated features in classification models is advantageous because it reduces redundancies and ensures that each feature contributes unique information, enhancing the model's generalization to new data by avoiding excessive dependencies on similar features. Moreover, it simplifies the model, speeds up training, and reduces the risk of overfitting.

Fig. 6 presents the accuracy levels achieved by various classifiers in the optimal approach for the Control vs. Pathological data group. From this, it is possible to see that the classifier LDA achieved a little high accuracy, reaching 73.9%. Furthermore, Fig. 7 provides visualizations of a confusion matrix that illustrates the accuracy and misclassifications of the LDA and an area under the ROC curve, demonstrating its ability to discriminate between the control and pathological group data.

After gathering the outcomes from the machine learning techniques, we will analyze and discuss these results to assess the effectiveness and implications of the applied methods.

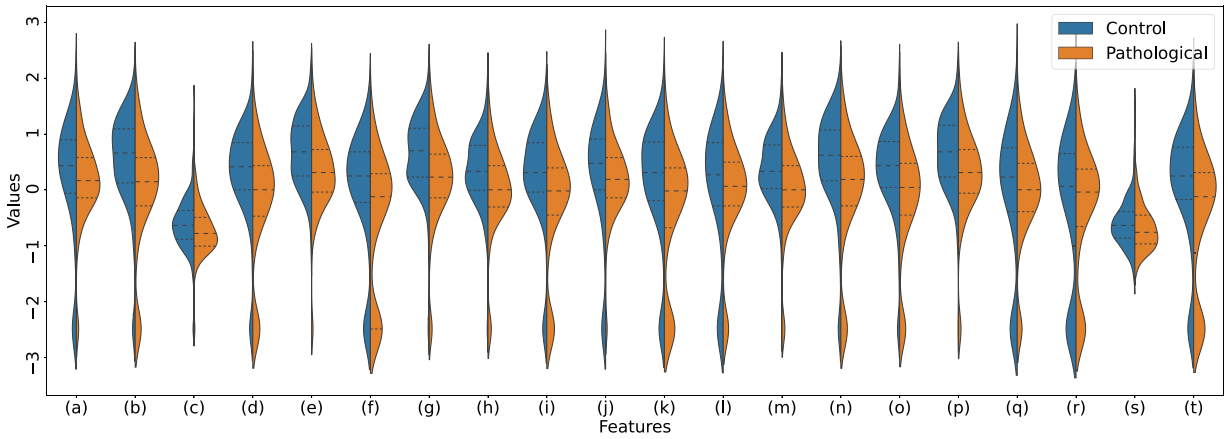


Fig. 4. Violin plot illustrating the normalized distribution (by Z-score) of the values of the 20 top features from the LassoCV feature selector in each class: (a) p335, (b) p1448, (c) p1270, (d) p708, (e) p31, (f) p1369, (g) p1102, (h) p172, (i) p1069, (j) p348, (k) p1437, (l) p1080, (m) p176, (n) p1452, (o) p710, (p) p28, (q) p247, (r) p1393, (s) p496, (t) p1403.

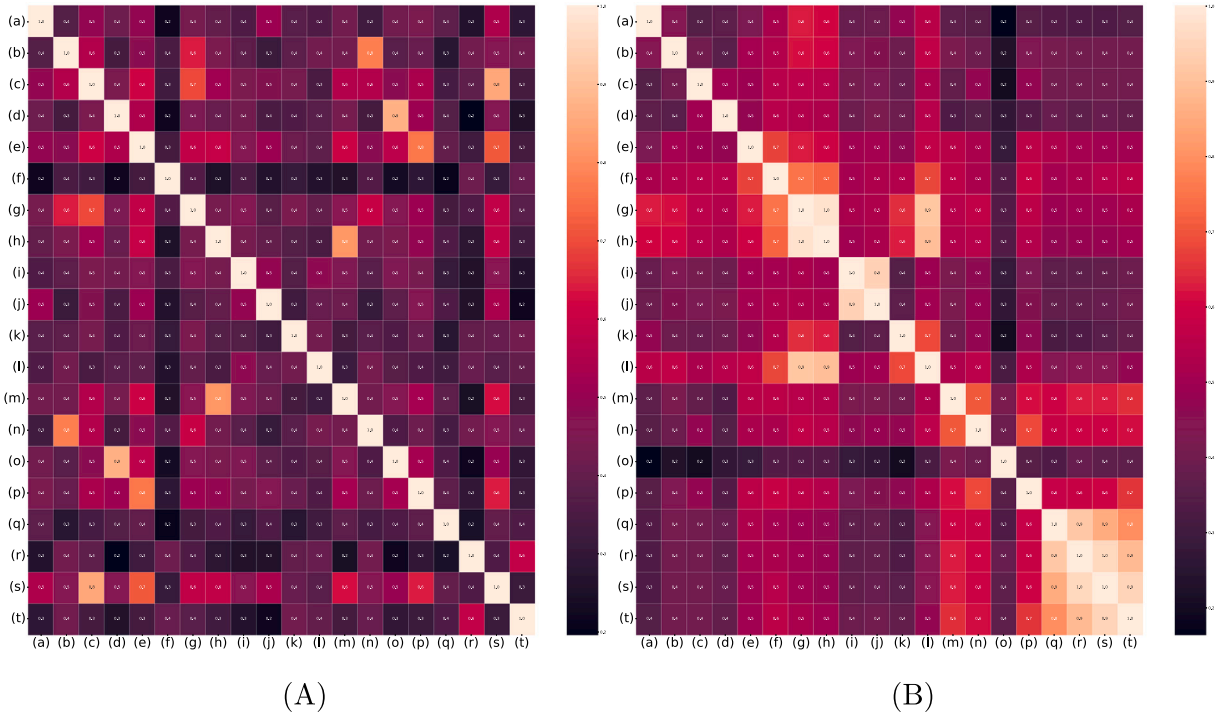


Fig. 5. Heatmap comparison of top 20 features selected by two methods. Panel (A) illustrates features selected by the LassoCV feature selector, including p335, p1448, and p1270, among others. Panel (B) shows features the RFE selector identifies, such as p98, p178, p343, etc. Each feature is represented in a heatmap format, indicating its relative importance in the dataset according to each selection technique.

4. Discussion and conclusions

This study explores machine-learning techniques for classifying glaucoma based on pupillary light reflex signals. Different approaches, such as filtering, sample balancing, and feature extraction and selection techniques were investigated.

4.1. Regarding filtering:

The statistical analysis in Table 3 indicated that refraining from signal filtering generally yielded superior results, irrespective of the other techniques implemented.

This finding was corroborated by the best accuracy analysis of the whole signal in Table 8, highlighting that non-filtered signals maintained their integrity and were more effective for classification.

4.2. Feature extraction:

The statistical analysis in Table 4 yielded mixed results, varying effectiveness based on the balancing and feature selection methods.

Nonetheless, the analysis of best accuracy in Table 8 suggested a preference for utilizing the full range of signal diameters, hinting that extensive feature extraction might not be necessary.

4.3. On sample balancing:

The oversampling technique significantly enhances classifier performance when there is a substantial imbalance.

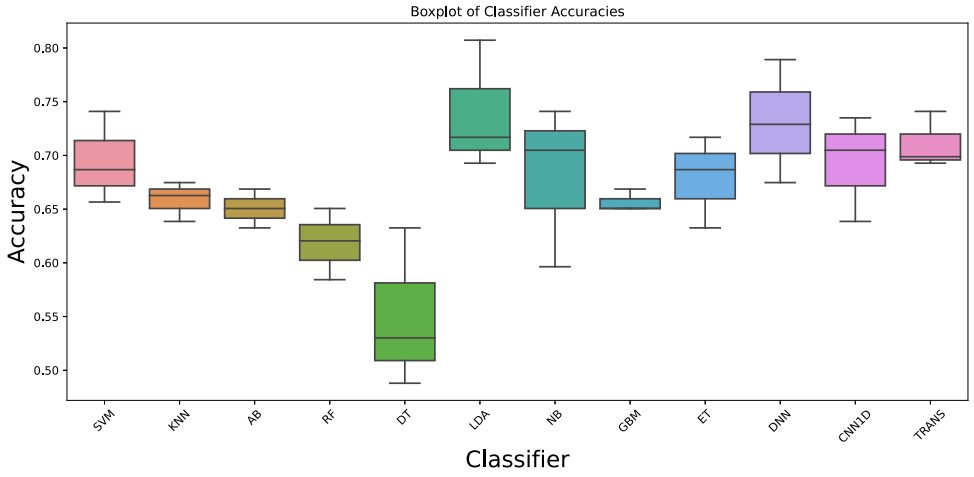


Fig. 6. Comparative accuracies of classifiers for control vs. pathological group illustrated through a box plot of the most effective approach.

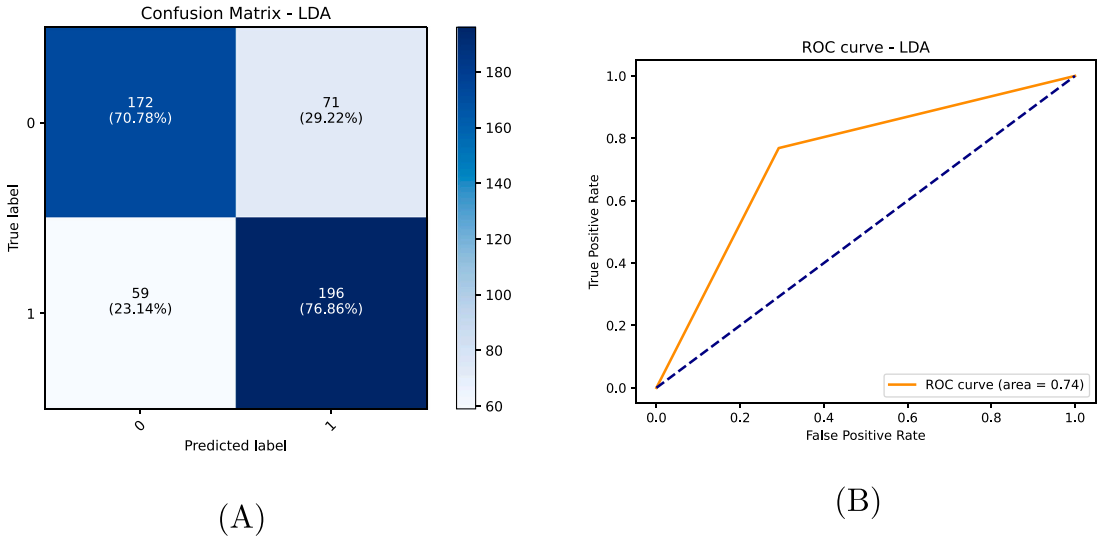


Fig. 7. Confusion Matrix (A) and ROC Curve (B) displaying the performance of the LDA in classifying the Control vs. Pathological Group using the optimal approach.

4.4. Feature selection:

The analyses reveal that the effectiveness of different feature selection techniques varied, with no single method showing complete dominance over the others. However, statistical analysis shows a slight preference for Recursive Feature Elimination. Conversely, LassoCV is the most effective strategy according to analysis by best accuracy.

Both analyses unanimously identified the SelectKBest method as the least recommended. This likely occurs because SelectKBest is a relatively simple technique that selects features based on individual statistical tests. This method evaluates each feature independently, without considering the interactions between them. On the other hand, RFE and LassoCV account for these interactions, resulting in more effective feature selection.

4.5. Cropping signal:

The analysis for the best accuracy using only the signal segment corresponding to a specific stimulus color did not produce definitive conclusions about which part of the signal is most critical, as shown in Table 7. Experiments have shown that each stimulus color proved to be more effective in different types of groupings.

The observed accuracies were slightly lower compared to the use of the complete signal in almost all groupings. This suggests that using the complete signal may be more advantageous with the techniques used, although the difference in performance was slight. The strategies of not filtering the signal, avoiding feature extraction, and applying oversampling balance remained the most effective. Traditional classifier methods prevailed in performance, particularly emphasizing the Extra Tree classifier.

4.6. Classifier analysis:

Traditional classifiers such as LDA, SVM, RD, and ET demonstrated the best performance in the classification task covering all six studied groupings, as detailed in Table 8. They maintained their effectiveness without being surpassed by artificial neural network algorithms. It is possible that neural network-based algorithms require more data to achieve superior effectiveness.

The findings suggest that preserving the full spectrum of signal data without filtering or feature extraction could be more advantageous for classification. This is particularly true for our research database, which likely has an acceptable noise level. We also observed the significant impact of oversampling with SMOTE on enhancing the classifiers' decision boundary delineation.

It was also observed, as expected, that classifying glaucoma in its severe stages is relatively straightforward. However, identifying glaucoma in its initial stages presents significant challenges.

4.7. Practical recommendations:

After analyzing the machine learning techniques employed, our recommendation is as follows:

When the noise levels are tolerable, unfiltered datasets can yield better results.

Using all the diameters of the pupillary signal has proven to be more effective for classification than extracting features commonly described in the literature.

In cases of data imbalance, implement the oversampling technique with SMOTE for balancing.

4.8. Limitations of the study:

Our database has a relatively small proportion of patients with moderate and severe glaucoma. This limitation is mitigated by adequately representing early glaucoma cases, which is crucial as pupillometry is particularly useful in early-stage glaucoma screening.

The feature extraction approach may yield enhanced results by discovering and including new features from the pupillary signal, indicating potential for further research in this area.

Exploring additional feature selection techniques and evaluating alternative data balancing strategies could further enhance the assessment.

In terms of classification, while we tested key classifiers, exploring other network architectures and fine-tuning hyperparameters could lead to even better results.

4.9. Limitations of PLR evaluation:

Various studies on PLR-based diagnosis indicate its susceptibility to factors like substances, other pathologies, and physiological conditions such as sleep deprivation and stress. Isolating these variables in glaucoma diagnosis remains a significant challenge.

5. Future work

After achieving satisfactory accuracy rates in glaucoma classification using a protocol with relatively long periods of dark adaptation (10 min) and recording time (4 min), the subsequent step is to explore the possibility of reducing these durations without compromising high precision. This optimization would make the pupillary reflex examination more efficient and practical for medical applications.

Currently, pupillometry relies on pupillometers for recording, essential for controlled stimulus and recording conditions in a light-shielded environment. A significant advancement would be developing an application that enables smartphone camera recordings, eliminating the need for specialized equipment. Ideally, users in a dimly-lit environment would record their pupillary reflex at a controlled distance, like arm's length, aligning the phone at eye level. The app could then analyze the recorded data to provide immediate screening results, enhancing the procedure's accessibility and convenience.

Another possible extension of this work would be the application of multiview learning [54,55], which synthesizes various data views to achieve more comprehensive data descriptions. Combining pupillometry data with other clinical information, such as imaging exams, laboratory results, and medical histories.

Funding

This research received support from Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG), Brazil.

CRediT authorship contribution statement

Hedenir Monteiro Pinheiro: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Eduardo Nery Rossi Camilo:** Validation, Resources, Formal analysis, Data curation. **Augusto Paranhos Junior:** Validation, Resources. **Afonso Ueslei Fonseca:** Visualization, Software. **Gustavo Teodoro Laureano:** Writing – review & editing, Formal analysis. **Ronaldo Martins da Costa:** Supervision, Project administration, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hedenir Monteiro Pinheiro reports financial support was provided by Fundação de Amparo á Pesquisa do Estado de Goiás (FAPEG). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgment

LaMCAD/UFG supported this research by providing computational infrastructure for data processing used in this study.

Appendix A. Customized algorithm for signal filtering

This appendix section presents a pseudocode for a specially designed algorithm used in pupil signal filtering.

Algorithm 1: Signal filtering.

```
Result: Returns the filtered signal
max_variation ← 3;
repetitions ← 100;
filtered_signal ← signal_original;
if discontinuity(filtered_signal, FACTOR_OF_DISCONTINUITY) is True then
    return filtered_signal;
else
    for i ← 0 to repetitions do
        foreach element in filtered_signal do
            if element is not the last and element > -1 then
                if difference(element, next element) > max_variation then
                    next element ← -1;
                    foreach subwindow in filtered_signal from next element do
                        if difference(element, subwindow) > max_variation then
                            next next element ← -1;
                        else
                            break;
                        end
                    end
                end
            end
        end
        if discontinuity(filtered_signal, FACTOR_OF_DISCONTINUITY) is True then
            max_variation ← max_variation + 1;
            filtered_signal ← signal_original;
        else
            break;
        end
    end
    return filtered_signal;
end
```

Appendix B. The architecture of the neural networks

This section presents diagrams illustrating the architecture of three neural networks utilized as classifiers in this study: a Fully Connected Network (FCN), a One-Dimensional Convolutional Neural Network (1D-CNN), and a Transformer Neural Network (see Figs. B.8–B.10).

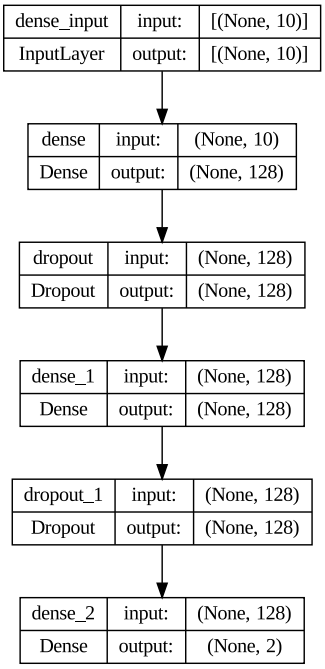


Fig. B.8. Fully Connected Neural Network (FCN) Classifier architecture.

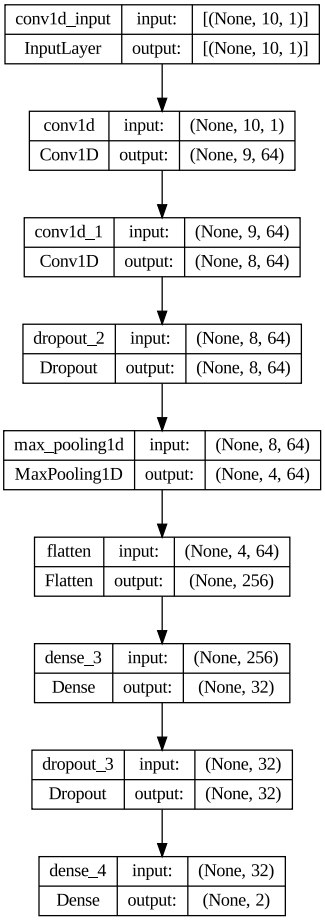


Fig. B.9. One-Dimensional Convolutional Neural Network (1D-CNN) classifier architecture.

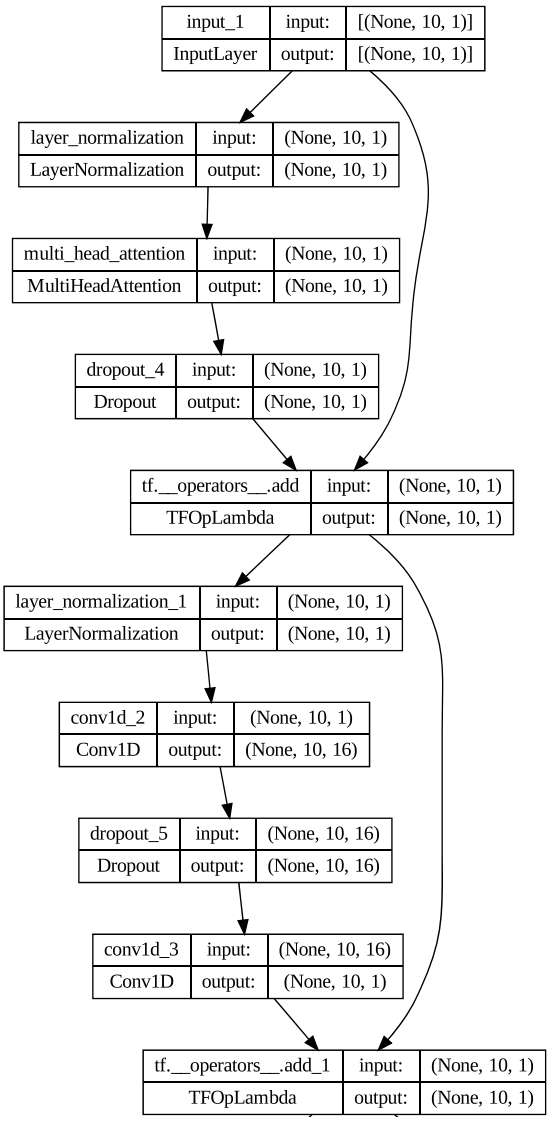


Fig. B.10. First head of the Transformer Neural Network (TRANS) classifier architecture.

References

[1] Gracitelli Carolina PB, Duque-Chica Gloria Liliana, Roizenblatt Marina, de Araújo Moura Ana Laura, Nagy Balazs V, de Melo Geraldine Ragot, Borba Paula Delegrego, Teixeira Sérgio H, Tufik Sergio, Ventura Dora Fix, et al. Intrinsically photosensitive retinal ganglion cell activity is associated with decreased sleep quality in patients with glaucoma. Ophthalmology 2015;122(6):1139–48.

[2] Weinreb Robert N, Khaw Peng Tee. Primary open-angle glaucoma. The Lancet 2004;363(9422):1711–20.

[3] Weinreb Robert N, Aung Tin, Medeiros Felipe A. The pathophysiology and treatment of glaucoma: a review. Jama 2014;311(18):1901–11.

[4] Tatham Andrew J, Meira-Freitas Daniel, Weinreb Robert N, Zangwill Linda M, Medeiros Felipe A. Detecting glaucoma using automated pupillography. Ophthalmology 2014;121(6):1185–93.

[5] Eddy David M, Sanders Lauri E, Eddy Judy F. The value of screening for glaucoma with tonometry. Survey Ophthal. 1983;28(3):194–205.

[6] Wu Zhichao, Medeiros Felipe A. Recent developments in visual field testing for glaucoma. Curr. Opin. Ophthalmol. 2018;29(2):141–6.

[7] Yaqub Moustafa. Visual fields interpretation in glaucoma: a focus on static automated perimetry. Commun. Eye Health 2012;25(79–80):1.

[8] Hoyt William F, Frisen Lars, Newman Nancy M. Fundoscopy of nerve fiber layer defects in glaucoma. Investigat. Ophthalmol. Visual Sci. 1973;12(11):814–29.

[9] Bussel Igor I, Wollstein Gadi, Schuman Joel S. OCT for glaucoma diagnosis, screening and detection of glaucoma progression. Br. J. Ophthalmol. 2014;98(Suppl 2):ii15–9.

[10] Chang Dolly S, Arora Karun S, Boland Michael V, Supakontanasan Wasu, Friedman David S. Development and validation of an associative model for the detection of glaucoma using pupillography. Am. J. Ophthalmol. 2013;156(6):1285–96.

[11] Najjar Raymond P, Sharma Sourabh, Atalay Eray, Rukmini Annadata V, Sun Christopher, Lock Jing Zhan, Baskaran Mani, Perera Shamira A, Husain Rahat, Lamoureux Ecosse, et al. Pupillary responses to full-field chromatic stimuli are reduced in patients with early-stage primary open-angle glaucoma. Ophthalmology 2018;125(9):1362–71.

[12] Duque-Chica Gloria L, Gracitelli Carolina PB, Moura Ana LA, Nagy Balázs V, Vidal Kallene S, Paranhos Augusto, Ventura Dora F. Inner and outer retinal contributions to pupillary light response: correlation to functional and morphologic parameters in glaucoma. J. Glaucoma 2018;27(8):723–32.

[13] Sarezky Daniel, Krupin Theodore, Cohen Aaron, Stewart Charles Wm, Volpe Nicholas J, Tanna Angelo P. Correlation between intereye difference in visual field mean deviation values and relative afferent pupillary response as measured by an automated pupillometer in subjects with glaucoma. J. Glaucoma 2014;23(7):419–23.

[14] Sarezky Daniel, Volpe Nicholas J, Park Meghan S, Tanna Angelo P. Correlation between inter-eye difference in average retinal nerve fiber layer thickness and afferent pupillary response as measured by an automated pupillometer in glaucoma. J. Glaucoma 2016;25(3):312–6.

[15] Park Hae-Young Lopilly, Jung Suk Hoon, Park Sung-Hwan, Park Chan Kee. Detecting autonomic dysfunction in patients with glaucoma using dynamic pupillometry. Medicine 2019;98(11).

[16] Martucci Alessio, Cesareo Massimo, Napoli Domenico, Sorge Roberto Pietro, Ricci Federico, Mancino Raffaele, Nucci Carlo. Evaluation of pupillary response to light in patients with glaucoma: a study using computerized pupillometry. Int. Ophthalmol. 2014;34(6):1241–7.

[17] Tatham Andrew J, Meira-Freitas Daniel, Weinreb Robert N, Marvasti Amir H, Zangwill Linda M, Medeiros Felipe A. Estimation of retinal ganglion cell loss in glaucomatous eyes with a relative afferent pupillary defect. Invest Ophthalmol Vis Sci 2014;55(1):513–22.

[18] Lawlor Mitchell, Quartilho Ana, Bunce Catey, Nathwani Neil, Dowse Emily, Kamal Debbie, Gazzard Gus. Patients with normal tension glaucoma have relative sparing of the relative afferent pupillary defect compared to those with open angle glaucoma and elevated intraocular pressure. Invest Ophthalmol Vis Sci 2017;58(12):5237–41.

[19] Charalel Resmi A, Lin Hugh S, Singh Kuldev. Glaucoma screening using relative afferent pupillary defect. J. Glaucoma 2014;23(3):169–73.

[20] Pillai Manju R, Sinha Sapna, Aggarwal Pradeep, Ravindran Ravilla D, Privitera Claudio M. Quantification of RAPD by an automated pupillometer in asymmetric glaucoma and its correlation with manual pupillary assessment. Indian J. Ophthalmol. 2019;67(2):227.

[21] Kankipati Laxmikanth, Girkin Christopher A, Gamlin Paul D. The post-illumination pupil response is reduced in glaucoma patients. Investigat. Ophthalmol. Visual Sci. 2011;52(5):2287–92.

[22] Pradhan Zia S, Rao Harsha L, Puttaiah Narendra K, Kadambi Sujatha V, Dasari Srilakshmi, Reddy Hemanth B, Palakurthy Meena, Riyazuddin Mohammed, Rao Dhanaraj AS. Predicting the magnitude of functional and structural damage in glaucoma from monocular pupillary light responses using automated pupillography. J. Glaucoma 2017;26(5):409–14.

[23] Rukmini Annadata V, Milea Dan, Baskaran Mani, How Alicia C, Perera Shamira A, Aung Tin, Gooley Joshua J. Pupillary responses to high-irradiance blue light correlate with glaucoma severity. Ophthalmology 2015;122(9):1777–85.

[24] Carle Corinne F, James Andrew C, Kolic Maria, Essex Rohan W, Maddess Ted. Luminance and colour variant pupil perimetry in glaucoma. Clinical Exper. Ophthalmol. 2014;42(9):815–24.

[25] Lee Jinho, Kim Young Kook, Ha Ahnul, Kim Yong Woo, Baek Sung Uk, Kim Jin-Soo, Lee Haeng Jin, Jeoung Jin Wook, Kim Seong-Joon, Park Ki Ho, et al. Temporal raphe sign for discrimination of glaucoma from optic neuropathy in eyes with macular ganglion cell–inner plexiform layer thinning. Ophthalmology 2019;126(8):1131–9.

[26] Arévalo-López Carla, Gleitze Silvia, Madariaga Samuel, Plaza-Rosales Iván. Pupillary response to chromatic light stimuli as a possible biomarker at the early stage of glaucoma: a review. Int. Ophthalmol. 2023;43(1):343–56.

[27] Bayraktar Serdar, Hondur Gözde, Şekeroğlu Mehmet Ali, Şen Emine. Evaluation of static and dynamic pupillary functions in early stage primary open angle glaucoma. J. Glaucoma 2023.

[28] Adhikari Prakash, Zele Andrew J, Thomas Ravi, Feigl Beatrix. Quadrant field pupillometry detects melanopsin dysfunction in glaucoma suspects and early glaucoma. Sci Rep 2016;6:33373.

[29] Pattan Hadiya Farhath, Liu Xiao, Williams Mark, King Brett, Port Nicholas, Tankam Patrice. Assessing the pupillary response in healthy and primary open-angle glaucoma. Invest Ophthalmol Vis Sci 2023;64(8):2034.

[30] Najjar Raymond P, Rukmini AV, Finkelstein Maxwell T, Nusinovici Simon, Mani Baskaran, Nongpiur Monisha Esther, Perera Shamira, Husain Rahat, Aung Tin, Milea Dan. Handheld chromatic pupillometry can accurately and rapidly reveal functional loss in glaucoma. Br. J. Ophthalmol. 2023;107(5):663–70.

[31] Wu Lianyi, Liu Yiming, Shi Yelin, Sheng Bin, Li Ping, Bi Lei, Kim Jinman. Detect glaucoma with image segmentation and transfer learning. In: Proceedings of the 32nd international conference on computer animation and social agents. 2019, p. 37–40.

[32] Gaddipati Divya Jyothi, Sivaswamy Jayanthi. Glaucoma assessment from fundus images with fundus to OCT feature space mapping. ACM Trans. Comput. Healthc. (HEALTH) 2021;3(1):1–15.

[33] Talaat Mennato-Allah, Raed Nataly, Medhat Aya, Ashraf Romisaa, Essam Mohammad, ElKashlan Rana Y, Abdel-Hamid Lamiaa. Glaucoma detection from retinal images using generic features: Analysis & results. In: Proceedings of the 2019 2nd international conference on watermarking and image processing. 2019, p. 10–5.

[34] An Guangzhou, Omodaka Kazuko, Tsuda Satoru, Shiga Yukihiro, Takada Naoko, Kikawa Tsutomu, Nakazawa Toru, Yokota Hideo, Akiba Masahiro. Comparison of machine-learning classification models for glaucoma management. J. Healthc. Eng. 2018;2018.

[35] Quan Yadan, Duan Huiyu, Zhan Zongyi, Shen Yuening, Lin Rui, Liu Tingting, Zhang Ting, Wu Jihong, Huang Jing, Zhai Guangtao, et al. Binocular head-mounted chromatic pupillometry can detect structural and functional loss in glaucoma. Front Neurosci 2023;17.

[36] Pinheiro Hedenir Monteiro, da Costa Ronaldo Martins. Pupillary light reflex as a diagnostic aid from computational viewpoint: a systematic literature review. J Biomed Inform 2021;117:103757.

[37] Rukmini AV, Milea Dan, Gooley Joshua J. Chromatic pupillometry methods for assessing photoreceptor health in retinal and optic nerve diseases. Front. Neurol. 2019;10:76.

[38] Crippa Sylvain V, Pedrosa Domellöf Fatima, Kawasaki Aki. Chromatic pupillometry in children. Front. Neurol. 2018;9:669.

[39] Park Jason C, Moura Ana L, Raza Ali S, Rhee David W, Kardon Randy H, Hood Donald C. Toward a clinical protocol for assessing rod, cone, and melanopsin contributions to the human pupil response. Investigat. Ophthalmol. Visual Sci. 2011;52(9):6624–35.

[40] Gracitelli Carolina PB, Duque-Chica Gloria L, Moura Ana Laura, Nagy Balazs V, de Melo Geraldine R, Roizenblatt Marina, Borba Paula D, Teixeira Sérgio H, Ventura Dora F, Paranhos Augusto. A positive association between intrinsically photosensitive retinal ganglion cells and retinal nerve fiber layer thinning in glaucoma. Investigat. Ophthal. Visual Sci. 2014;55(12):7997–8005.

[41] Pinheiro Hedenir Monteiro, da Costa Ronaldo Martins, Camilo Eduardo Nery Rossi, da Silva Soares Anderson, Salvini Rogerio, Laureano Gustavo Teodoro, Soares Fabrizzio Alphonsus, Hua Gang. A new approach to detect use of alcohol through iris videos using computer vision. In: International conference on image analysis and processing. Springer; 2015, p. 598–608.

[42] Silva Cleyton RG, Gonçalves Cristhiane, Camilo Eduardo NR, Santos Fabio B, Siqueira Joyce, Albuquerque Eduardo S, Soares Fabrízzio AAMN, Oliveira Leandro LG, Costa Ronaldo Martins da. Automated evaluation system for human pupillary behavior. In: International Medical Informatics Association (IMIA) and IOS Press. 2017, p. 589–93, Volume 245: MEDINFO 2017: Precision Healthcare through Informatics.

[43] Stockman George, Shapiro Linda G. Computer Vision. Prentice Hall PTR; 2001, p. 279–325.

[44] Wang Chien-Yao, Bochkovskiy Alexey, Liao Hong-Yuan Mark. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 7464–75.

[45] Zandi Babak, Lode Moritz, Herzog Alexander, Sakas Georgios, Khanh Tran Quoc. Pupilext: Flexible open-source platform for high-resolution pupillometry in vision research. Front. Neurosci. 2021;15:603.

[46] Khalid Samina, Khalil Tehmina, Nasreen Shamila. A survey of feature selection and feature extraction techniques in machine learning. In: 2014 science and information conference. IEEE; 2014, p. 372–8.

[47] Ngo Quoc Cuong, Bhowmik Susmit, Sarossy Marc, Kumar Dinesh Kant. Pupillary complexity for the screening of glaucoma. IEEE Access 2021;9:144871–9.

[48] He Haibo, Ma Yunqian. Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons; 2013, p. 47.

[49] Kumar Vipin, Minz Sonajharia. Feature selection: a literature review. SmartCR 2014;4(3):211–29.

[50] Kohavi Ron, et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, vol. 14.2. Montreal, Canada; 1995, p. 1137–45.

[51] Razali Nornadiah Mohd, Wah Yap Bee, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. J. Statist. Model. Analytics 2011;2(1):21–33.

[52] Efron Bradley, Hastie Trevor. Computer age statistical inference, student edition: algorithms, evidence, and data science, vol. 6, Cambridge University Press; 2021.

[53] Conover William Jay. Practical nonparametric statistics, vol. 350, john wiley & sons; 1999.

[54] Xu Cai, Guan Ziyu, Zhao Wei, Wu Hongchang, Niu Yunfei, Ling Beilei. Adversarial incomplete multi-view clustering.. In: IJCAI, Vol.7. 2019, p. 3933–9.

[55] Xu Cai, Zhao Wei, Zhao Jinglong, Guan Ziyu, Song Xiangyu, Li Jianxin. Uncertainty-aware multiview deep learning for internet of things applications. IEEE Trans Ind Inf 2022;19(2):1456–66.