Glaucoma Detection from Raw SD-OCT Volumes: a Novel Approach Focused on Spatial Dependencies

Gabriel García[1], Adrián Colomer[1] and Valery Naranjo[1]

[1]Instituto de Investigación e Innovación en Bioingeniería (I3B), Universitat Politècnica de València (UPV), 46022 Valencia, Spain

ARTICLE HISTORY
Compiled October 22, 2020

ABSTRACT

Background and objective: Glaucoma is the leading cause of blindness worldwide. Many studies based on fundus image and optical coherence tomography (OCT) imaging have been developed in the literature to help ophthalmologists through artificial-intelligence techniques. Currently, 3D spectral-domain optical coherence tomography (SD-OCT) samples have become more important since they could enclose promising information for glaucoma detection.

To analyse the hidden knowledge of the 3D scans for glaucoma detection, we have proposed, for the first time, a deep-learning methodology based on leveraging the spatial dependencies of the features extracted from the B-scans.

Methods: The experiments were performed on a database composed of 176 healthy and 144 glaucomatous SD-OCT volumes centred on the optic nerve head (ONH). The proposed methodology consists of two well-differentiated training stages: a slide-level feature extractor and a volume-based predictive model. The slide-level discriminator is characterised by two new, residual and attention, convolutional modules which are combined via skip-connections with other fine-tuned architectures. Regarding the second stage, we first carried out a data-volume conditioning before extracting the features from the slides of the SD-OCT volumes. Then, Long Short-Term Memory (LSTM) networks were used to combine the recurrent dependencies embedded in the latent space to provide a holistic feature vector, which was generated by the proposed sequential-weighting module (SWM).

Results: The feature extractor reports AUC values higher than 0.93 both in the primary and external test sets. Otherwise, the proposed end-to-end system based on a combination of CNN and LSTM networks achieves an AUC of 0.8847 in the prediction stage, which outperforms other state-of-the-art approaches intended for glaucoma detection. Additionally, Class Activation Maps (CAMs) were computed to highlight the most interesting regions per B-scan when discerning between healthy and glaucomatous eyes from raw SD-OCT volumes.

Conclusions: The proposed model is able to extract the features from the B-scans of the volumes and combine the information of the latent space to perform a volume-level glaucoma prediction. Our model, which combines residual and attention blocks with a sequential weighting module to refine the LSTM outputs, surpass the results achieved from current state-of-the-art methods focused on 3D deep-learning architectures.

KEYWORDS
Glaucoma detection, SD-OCT volumes, Convolutional attention blocks, Residual connections, LSTM networks, Sequential-weighting module.

CONTACT G.G. Author. Email: jogarpa7@i3b.upv.es

1. Introduction

Glaucoma is a group of progressive optic neuropathies that affects the optic nerve causing several visual field defects and structural changes [1]. Nowadays, this chronic disease is the leading cause of blindness worldwide [2], with a number of estimated cases of 111.8 million in 2040, according to [3]. Early diagnosis of glaucoma is essential for timely treatment in order to avoid the irreversible vision loss [2]. Currently, there is no single accurate test to certify the glaucoma diagnosis, so the procedure includes a lot of hardworking tests such as pachymetry (to measure the thickness of the cornea), tonometry (to assess the intraocular pressure), visual field tests and a subjective examination and interpretation of optical features from different experts who often disagree [4]. In this context, techniques based on image analysis like fundus image and optical coherence tomography (OCT) have become very important for the diagnosis and management of this degenerative disease. In particular, OCT imaging modality [5] is a non-contact and non-invasive technique able to quantify several retinal structures through generating high-resolution 2D and 3D images of the retina. Ophthalmologists usually make use of these 2D-OCT images centred on the optic disc to analyse structural changes in the retinal nerve fibre layer (RNFL) and in the ganglion cell inner plexiform layer (GCIPL). Both structures are reported as useful biomarkers of glaucoma for the disease progression [6]. Otherwise, fundus image analysis is postulated as a great cost-effectiveness technique which has reported promising results in the detection of several eye-focused diseases [7–9]. However, although fundus image-based studies are cheaper than OCT, this modality is the quintessential imaging technique for glaucomatous damage evaluation [10]. This is because fundus photography is colour-dependent on the training data set and its interpretation remains subjective [11, 12], whereas OCT modality can provide reproducible and objective measurements of optic nerve head (ONH) and RNFL thickness [13]. Besides, glaucoma disease is evident in the deterioration of the cell layer around the optic disc, which is very hard to distinguish in the 2D projection of the fundus images. Therefore, since OCT imaging modality allows focusing on the depth axis to identify structural retinal changes, glaucoma disease can be easier detected via OCT, instead of fundus image. Furthermore, OCT system can provide high-resolution three-dimensional images of the macula and ONH in the spectral domain (SD), which emerges as a powerful tool for detecting glaucoma [10]. However, due to around 30 million of OCT scans are acquired each year, experts rarely scroll through the entire cube because it supposes a workload difficult to face [14]. For this reason, in this paper, we propose a promising volume-based predictive model to evidence the added value that SD-OCT volumes can provide for glaucoma diagnosis.

1.1. Related work

1.1.1. 2D-OCT approximation for glaucoma detection

Many state-of-the-art studies, focusing on OCT techniques, have been proposed to address the automatic detection of glaucoma with the aim of reducing the workload and the rate of discordance between experts.

Hand-driven learning on 2D-OCT projection. Most of glaucoma diagnosis-based studies made use of 2D-OCT scans centred on the optic disc, a.k.a circumpapillary images, due to their known potential when diagnosing [15]. To the best of the authors' knowledge, all the circumpapillary-based studies intended to glaucoma detection were performed by applying hand-driven learning methods, such as [16,17], which required hand-crafted encoding phases before accomplishing the classification stage, e.g. segmentation of regions of interest and hand-crafted feature extraction.

Deep learning on 2D-OCT projection. Another way to address the glaucoma identification from circumpapillary images would be via deep learning, which would allow operating directly on the 2D-OCT scans without defining previous biomarkers, as we did in our previous study [18]. However, all the studies found in the literature (which apply deep-learning techniques from 2D scans) were based on fundus images [9,19] or RNFL probability maps [20,21] combining fundus images and OCT B-scans, but no previous studies were addressed just from circumpapillary images. This fact could be explained taking into account that researchers focused their efforts on identifying useful patterns (e.g. RNFL and GCIPL) capable of providing a tangible interpretation for the clinicians. It is the reason because many other studies were carried out for the sole purpose of segmenting the retinal layers of interest [22,23].

1.1.2. 3D-OCT approximation for glaucoma detection

Going deeper into the glaucoma detection, the real challenge today lies in the analysis of the unknown potential enclosed in the 3D-OCT scans, since specialists postulate that SD-OCT volumes hide a key knowledge that is not currently being traced due to their large associated workload. Therefore, we propose here a clinical decision support system based only on the analysis of ONH-centred cubes to claim the importance of the 3D cross-sectional information about the glaucoma diagnosis.

Hand-driven learning on 3D-OCT approach. Similarly to the 2D approximation, some studies in the literature applied hand-crafted algorithms on 3D scans to face the glaucoma discrimination [24–26]. In particular, both [24] and [25] manually extracted features related to the RNFL and the optic nerve throughout the cube. The authors proposed a similar methodology, but they tested the models on different databases. In [24], the best AUC reported was 0.877 using a random forest classifier from a database composed of 46 healthy and 57 glaucomatous patients, whereas in [25], the same researchers provided an AUC of 0.818 by applying bagging methods on a database of 48 and 62 healthy and glaucomatous patients, respectively. Another creative approach was proposed in [26], where the authors made use of a superpixel segmentation technique before addressing the feature extraction stage. They combined the features extracted from the superpixel maps with other common RNFL measurements to feed an adaptive boosting classifier. The researchers obtained an AUC of 0.855 from a database of 44 healthy and 89 glaucomatous eyes.

Deep learning on 3D-OCT approach. The use of deep-learning methods to address the glaucoma detection via SD-OCT volumes has been increased in recent times. In fact, most studies have been published during the last two years, which claims the current interest of OCT volumes for glaucoma diagnosis [27–29]. A research group from Hong Kong deserves a special mention because most of the contributions in this field come from their work. In particular, they carried out two closely similar studies, [28] and [29] to detect glaucoma by means of 3D-Convolutional Neural Networks (3D-CNNs). The main differences between them lied in the database and inclusion/exclusion criteria, as the authors concluded in [28]. Noury et al. [28] made use of a private database composed of 316 glaucomatous and 247 healthy eyes from people of different ethnicity. They developed an end-to-end classification model based on the network proposed in [30]. The researchers achieved an AUC of 0.8883 in the primary test set and this value was lower when testing external data sets. Otherwise, the authors in [29] applied similar techniques on a homogeneous database only composed of Chinese Asian people. Particularly, 2926 glaucomatous and 1961 healthy eyes. The work demonstrated good performance with an AUC of 0.969, a sensitivity of 0.89, a specificity of 0.96 and an accuracy of 0.91 when testing the primary data set. However, the results fell when the researchers assessed their network with an external database from Stanford, reaching 0.893, 0.78, 0.79 and 0.80 of AUC, sensitivity, specificity and accuracy, respectively. More recent works from the same authors [31, 32] performed a multi-output architecture by including other well-known measures (for glaucoma diagnosis) such as Visual Field Index (VFI), Mean Deviation (MD) and Pattern Standard Deviation (PSD). Specifically, a neural branch of the network was responsible for the classification between normal and glaucomatous cases, whereas the other branch was intended to regression tasks for predicting VFI, MD and PSD values. In this way, the model was fed with information from VFI, MD and PSD metrics during the backward propagation step in order to update the weights in each epoch taking into account interesting parameters associated with glaucoma disease. However, these two last studies are not comparable with our work because additional information was used besides the raw OCT volumes, unlike the works [28, 29] accomplished by the same research group. Another interesting study was carried out by IBM team in [27], where the authors made a comparison between hand-driven and data-learning approaches. They proposed a 3D-CNN architecture trained from scratch and they achieved an AUC of 0.94 in the prediction of the test set. However, it should be noted that, in this case, the experiments were performed on a significant unbalanced database, whose test set was composed of 17 healthy and 93 glaucomatous patients.

In this context, other works could be mentioned because they also applied deep-learning techniques on SD-OCT volumes, but with other purposes. For example, in [33] the researchers from Hong Kong developed a deep-learning algorithm for discriminating ungradable OCT optic disc scans. Otherwise, the authors in [14] implemented deep-learning techniques to detect specific Age-Related Macular Degeneration (AMD) patterns in the B-scans of the three-dimensional cubes. Also, De Fauw et al. in [34] applied artificial-intelligence algorithms on OCT volumes to diagnosis several retinal injuries via tissue segmentation.

1.2. Contribution of this work

This paper documents several key contributions concerning the glaucoma detection from SD-OCT volumes. Unlike the previous studies that addressed the problem using 3D CNNs, we reveal a new approach characterised by extracting features from the B-scans by an innovative 2D-CNN, and preserving the feature dependencies embedded in the latent space making use of LSTM networks [35] along with an additional proposed module. The combination of CNNs and LSTM networks has been successfully performed in recent studies to identify pathological biomarkers associated to AMD and diabetic macular edema (DME) [14], as well as to predict the progression of the ophthalmic diseases from different slit-lamp images [36]. However, to the best of the author's knowledge, we are the first that suggest the use of CNN-LSTM to address the glaucoma detection, by assuming each spatial slide of the volume as a temporary instance. As a novelty, in order to attain the feature-extraction stage, we propose a new slide-level discriminator based on a pre-trained 2D-CNN model able to discern between healthy and glaucomatous cases just from raw circumpapillary OCT images. The proposed 2D-CNN feature extractor is composed of a novel combination of pre-trained convolutional blocks in parallel with residual modules trained from scratch. Additionally, an attention block was also included via skip-connection to focus on local related-glaucoma areas during the training phase. Moreover, we propose an innovative way of codifying the LSTM outputs implementing a sequential-weighting module (SWM) before addressing the final classification stage. The flowchart of the designed end-to-end system is exposed in Figure 1, where we represent how the pre-trained circumpapillary base model extracts the features from the SD-OCT slides and how the three-dimensional information is analysed making use of LSTM networks to finally predict the class of each specific ONH-centred cube.

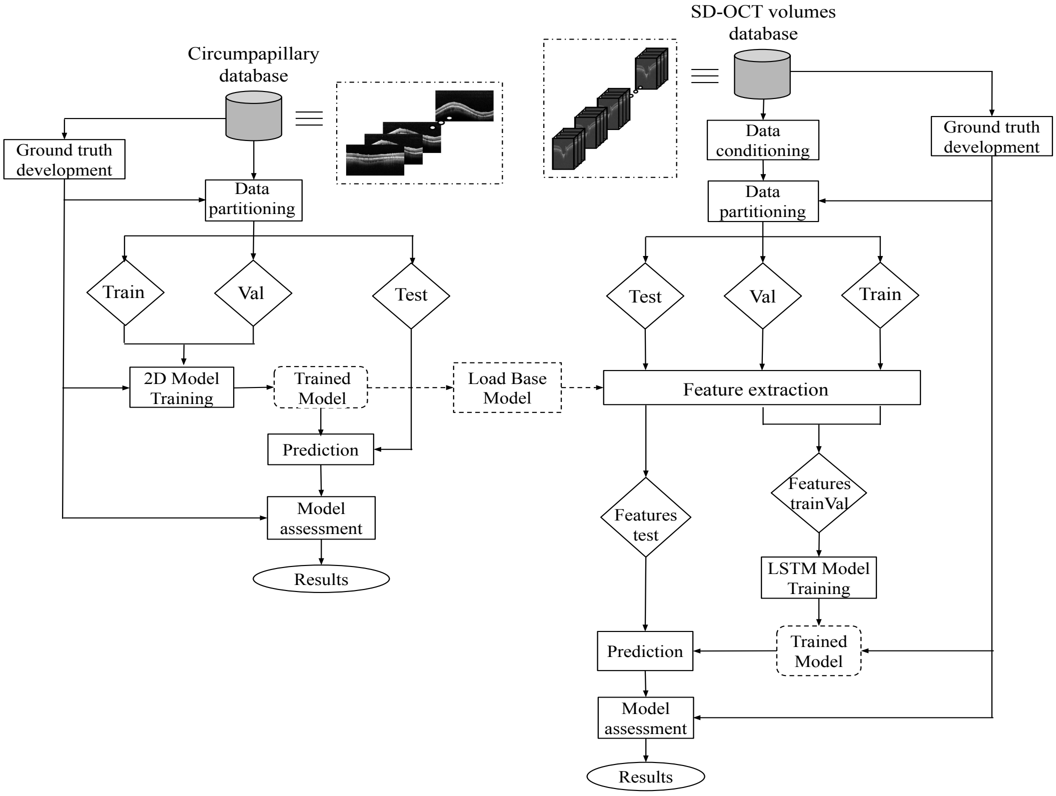


Figure 1. Flowchart of the proposed classification approach based on combining CNN and LSTM networks to distinguish between healthy and glaucomatous eyes from SD-OCT volumes.

In the recent study [27], the authors claimed that they used 3D convolutions to be able to accomplish the 3D Class Activation Maps (CAMs) because otherwise the resulting CAM would be 2D and the depth information would be lost. Against the statement of [27], our LSTM-based model is capable of leveraging the spatial dependencies extracted from the SD-OCT slides to compute the 2D-CAMs sequentially. Thereby, we enable an interpretation of SD-OCT volumes based not only on identifying the regions of interest (ROIs) of each slide, but also the most relevant B-scans of the volume for glaucoma classification. At this point, it is important to note that we also replicate several architectures proposed in the literature to make a direct comparison between different methods. In particular, we test in our database the models of the state-of-the-art studies intended to glaucoma detection just from SD-OCT volumes, i.e. the work that reported the best results by Hong Kong and Stanford association [29], and the work carried out by IBM research group [27].

2. Material

Three different and independent databases were employed to accomplish this study, as indicated in Table 1. Two of them are related to circumpapillary OCT images and they were used to train and validate the proposed slide-level feature extractor. The third database is composed of the SD-OCT volumes from which we built the predictive models for glaucoma detection. Both the circumpapillary and SD-OCT volumes databases are centred around the optic nerve head (ONH) of the retina to extract the B-scans. Note that, although circumpapillay and B-scans of the volumes are extracted following different patterns, the structure involved in both kind of scans is the same, i.e. the layer of fibres of the retina. For that reason, we used the models trained with the circumpapillary images as a feature extractor of the B-scans of the volumes.

The first data set (*circ-DB-1*), intended to train the slide-level discriminator, consists of 249 cross-sectional images around the ONH of the retina (B-scans). Specifically, 156 healthy and 93 glaucomatous samples from 174 patients were labelled by an expert to create the ground truth for training and evaluating the models. The second circumpapillary database (*circ-DB-2*), which comes from another hospital, was used to perform an external validation of the proposed feature extractor. Particularly, *circ-DB-2* is composed of 336 OCT images (143 glaucomatous and 193 healthy cases from 199 patients) which were annotated by another ophthalmologist. It should be noted that a *Heidelberg Spectrallis OCT*-called system was used to extract the B-scans (496 × 768 pixels) from both databases with an axial resolution of 4-5 μm. This equipment employs a super-luminescence diode with an infrared beam of average wavelength of 870 nm and a bandwidth of 25 nm. Patients with primary open-angle glaucoma (POAG) were included in the study, whereas subjects with other eye diseases, e.g. cataract, closed-angle glaucoma, and pseudo exfoliation syndrome were excluded. More information related to the age and gender of the patients is detailed in Table 2. It should be noted that all the subjects that compose the different databases are Caucasian.

The third database (vol-DB-3) contains the spectral-domain OCT samples that we used to develop our volume-based classification model. It consists of 320 OCT scans centred on the ONH which were captured on a *Topcon* 2000 OCT machine. This equipment allows measuring up to 45^o and a depth of 2.3 mm with a resolution of less than 6 μm using a super luminescent diode of 840 nm. Specifically, vol-DB-3 is composed of 176 healthy and 144 glaucomatous three-dimensional scans of $885 \times 512 \times 128$ voxels per volume, as detailed in Table 1. An expert ophthalmologist performed a volume-level annotation of the database containing cases of 200 patients with an age comprised between 18 and 80 years (see Table 2). Concerning the inclusion and exclusion criteria for diagnosis, the healthy group included samples with best-corrected visual acuity 20/40 or better, normal intra-ocular pressure (IOP) and normal-appearing optic nerves. Otherwise, scans with refractive error of >5D of the sphere or 2.5D of cylinder, history retinal disease, intra-ocular pressure >21 mmHg or unusable OCT were excluded from the study. For the glaucoma group, the inclusion criteria lied in any glaucomatous visual field defect, whereas the exclusion rules comprised refractive error of >5D of the sphere or 2.5D of cylinder, optic nerve-related diseases and unusable OCT samples.

Note that the protocol used for glaucoma labelling was carried out via European guideline for Glaucoma diagnosis. A thorough examination includes intra-ocular pressure analysis (using Goldmann applanation tonometry), study of the central corneal thickness, assessment of the anterior chamber angle (Gonioscopy), optic nerve head assessment (via slit lamp examination), Standard Automated Perimetry (using Octopus system) and measurement of the thickness of retinal nerve fibre layer and ganglion cell layer (with OCT+HRT equipment). Based on the above-mentioned examinations, a diagnosis was made into Healthy or Glaucoma for each sample of all databases. It is also important to highlight that any of the following criteria, if repeatable, was considered sufficient evidence of glaucomatous visual field defect: glaucoma hemifield test outside normal limits, pattern standard deviation with p-value < 0.05 or a cluster of three points or more in the pattern deviation in a single hemifield (superior or inferior) with p-values < 0.05, one of which must have a p-value < 0.01.

Table 1. Material corresponding to the databases used in this study.

Database	Label	Patients	Samples	Dimensions
circ-DB-1	Healthy	107 (61.49%)	156 (62.65%)	496 × 768
	Glaucoma	67 (38.51%)	93 (37.35%)	
circ-DB-2	Healthy	99 (49.75%)	193 (57.44%)	496 × 768
	Glaucoma	100 (50.25%)	143 (42.56%)	
vol-DB-3	Healthy	100 (50.00%)	176 (55.00%)	885 × 512 × 128
	Glaucoma	100 (50.00%)	144 (45.00%)	

[1] Database employed to train the feature extractor model.
[2] Database used in the external validation of the feature extractor.
[3] Database intended to create the volume-based predictive model.

Table 2. Demographic data related to the age and gender of the patients from each database used in this study.

	Age		Gender	
	Range	$\mu \pm \sigma$	**Male**	**Female**
circ-DB-1	[15-89]	55.95±18.77	74 (42.54%)	100 (57.47%)
circ-DB-2	[24-93]	60.82±12.32	86 (43.22%)	113 (56.78%)
vold-DB-3	[18-80]	50.13±15.54	89 (44.50%)	111 (55.50%)

3. Methodology

3.1. Slide-level feature extractor design

The objective in this stage is to build a 2D-CNN architecture able to extract discriminatory features from the slides of the SD-OCT volumes. So, in our previous work [18], we carried out a validation of different architectures making use of the raw circumpapillary OCT samples. Specifically, the most common state-of-the-art architectures, as well as other CNNs trained from scratch, were considered. As detailed in [18], we proposed shallow networks from scratch due to the small amount of data, and we studied the use of data augmentation techniques to alleviate that problem. Additionally, we fine-tuned some of the most popular architectures of the literature, such as VGG16, VGG19, ResNet50, InceptionV3 and Xception [37], to take advantage of the wide knowledge acquired by these networks when they were trained on the *ImageNet* data set. Thus, we loaded the weights ω pre-trained with around 14 million of natural images to initialise the coefficients of the networks and then, we performed a *deep fine-tuning* strategy [38] to freeze the coefficients of the three first convolutional blocks and retrain the lasts, making use of the specific samples. Note that, we replicated ×3 the channels of the grey-scale images to adapt the input dimensionality to fine-tune the CNNs. Also, we applied a ×0.5 down-sampling to face the GPU memory constraints.

According to [18], the VGG family of networks reported the best glaucoma detection performance from raw circumpapillary OCT images. Therefore, to accomplish this study, we made use of these family of architectures as a starting point to develop the new feature extractor. In particular, we kept the same *deep fine-tuning* strategy previously implemented in [18] concerning the VGG16 architecture. However, in this paper, we propose two innovative modules to improve the models' performance through skip-residual connections.

The first module (M_{res}) consists of a combination of the fine-tuned VGG16 architecture with a residual structure applied in parallel to the unfrozen blocks, followed by a 1×1 convolution layer, as we show in Figure 2. Thereby, we connected fine-tuning techniques with other convolutional layers to propagate the information from initial layers to the lasts, using residual connections in a novel way. This makes possible to mitigate the problem of vanishing gradients by allowing the shortcut to flow through the gradient of a deeper architecture. Unlike the traditional shortcuts defined in [39], where a specific input fed the network at two different points of it, the proposed system introduces a convolutional shortcut inspired by the basic structure of the ResNet-50 architecture. Such structure aims to optimise the dimensionality of the filters by alternating convolution layers of 1×1 and 3×3 kernel sizes, which are represented in green and blue boxes, respectively, in Figure 2. Also, an initial batch normalisation layer (in brown) and a final max-pooling layer (in red) were implemented as a part of the residual block (see Figure 2). Note that kernel and stride sizes of 4×4 were specified for the max-pooling layer guaranteeing the consistency of the filter dimensions to concatenate the residual features with the output from the Block_5 VGG16, as observed in Figure 2. Finally, the M_{res} module contains a 1×1 convolution layer which generates a volume of features $G = \{g_1, g_2, ..., g_k, ..., g_C\}$, where $C = 512$ is the number of filters in the volume and g_k is the *k-th* feature map with dimensions $H \times W = 7 \times 12$.

The second module (M_{att}) of the proposed network includes an attention block characterised by a succession of 1×1 convolutional layers intended to refine the features in the spatial dimension. Specifically, the proposed module is a kind of bottleneck architecture composed of a batch normalisation layer followed by two successive ReLu-activated convolutions, in which the size of the filters is decreased progressively, as specified in Figure 2. Also, a 1×1 convolution layer with a unique filter passing through a sigmoid function (purple) was used to recalibrate the inputs. At the end of the bottleneck, it includes another 1×1 convolution layer which increases the number of filters to make possible the concatenation between the inputs and the outputs of the attention block. In this way, a basic skip-connection was implemented to flow larger gradients to previous layers by learning an identity function as a shortcut, as observed in Figure 2. Finally, the second module M_{att} is provided with another 1×1 convolution layer used to obtain a feature volume map $F = \{f_1, f_2, ..., f_k, ..., f_C\}$.

Regarding the top model, a spatial squeeze was performed by a Global Average Pooling (GAP) layer, which provides a vector $x \in \mathbb{R}^{1 \times 1 \times C}$ according to the Equation (1). Finally, we defined a softmax-activated dense layer with two neurons corresponding to the two classes (healthy and glaucoma) in which the OCT images had to be classified.

$$x_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_k(i,j). \tag{1}$$

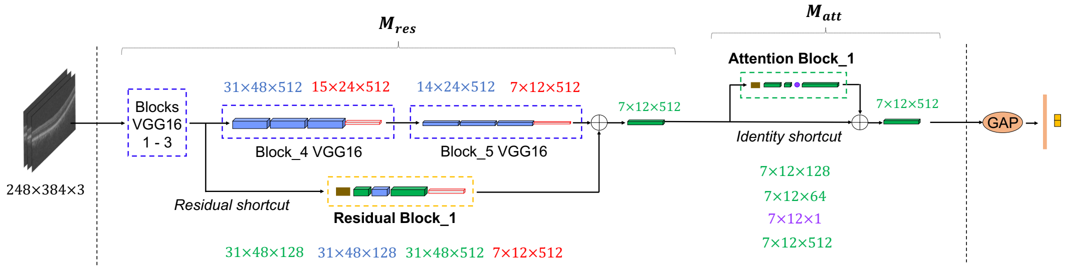
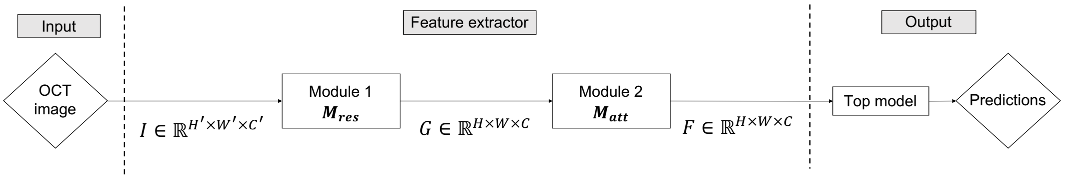


Figure 2. Developed architecture to distinguish between healthy and glaucomatous eyes from raw circumpapillary OCT images. Note that blue, red, brown and green colours denote 3 × 3 convolution, max-pooling, batch normalisation and 1 × 1 convolution layers, respectively.

As summarised in Figure 3, given an input image $I \in \mathbb{R}^{H' \times W' \times C'}$, being $H' \times W' \times C' = 248 \times 384 \times 3$ the dimensions of I, the first module M_{res} of the feature extractor generates a volume map $G \in \mathbb{R}^{H \times W \times C}$, $M_{res} : I \rightarrow G$. From here, G becomes the input to the second module M_{att}, which provides a refined output $F \in \mathbb{R}^{H \times W \times C}$, $M_{att} : G \rightarrow F$ that corresponds to the feature volume embedded in the latent space, which will be used to extract information from each B-scan of the SD-OCT volumes.



Input
Feature extractor
Output
OCT image
Module 1 M_{res}
Module 2 M_{att}
Top model
Predictions
$I \in \mathbb{R}^{H' \times W' \times C'}$
$G \in \mathbb{R}^{H \times W \times C}$
$F \in \mathbb{R}^{H \times W \times C}$

Figure 3. Flowchart of the proposed circumpapillary architecture highlighting the connections between the different modules of the feature extractor.

3.2. Volume-based predictive model development

3.2.1. Data-volume conditioning

As reported by Maetschke et al. [27], it was necessary to prepare the SD-OCT volumes database to face the constraints of the GPU memory caused by a large amount of data. In this paper, we propose a conditioning step of the slides based on extracting the useful information from each image, instead of applying a mere down-sampling from the volumes like [27,28]. First of all, we discarded the 32 initial and the 32 final slides from the total of 128 because the rich information seems to be located around the ONH, i.e. around the central slides of the cube, according to to the recent study [32].

In addition, we also developed a series of algorithms to remove useless pixels from each slide ensuring the same dimensions for all the slides. In this way, given $S = \{s_1, s_2, s_3, ..., s_P\}$ and $V = \{v_1, v_2, v_3, ..., v_Q\}$, where S and V are sets of slides and volumes composed of P and Q instances, respectively, the algorithm is able to reduce the dimensions $M \times N$ of each slide $s_i \in v_j$, with $i = \{1, 2, 3, ..., P\}$ and $j = \{1, 2, 3, ..., Q\}$, to dimensions $m \times N$. In order to calculate m, it was first necessary to extract the dimensions of each specific bounding box $B_i \subset s_i$, which corresponds to the region of the slide s_i that maximises the target retina area. Specifically, B_i was obtained by applying the *ROIret* function which consists of a succession of morphological operations followed by the Otsu's binarisation method, according to Algorithm 1.

Algorithm 1: *ROIret* function to extract the region B_i from each slide $s_i \in v_j$.

```
Data: Specific slide $s_i \in v_j$.
Functions: $Otsu$, to find the optimal threshold.
$Rectangle$, to extract the region from a mask.
Result: Bounding box $B_i \subset s_i \in v_j$.

Initialisation:
$E_O \leftarrow 1$;            %Disk structuring element for opening
$E_D \leftarrow [10, 20]$;     %Rectangular structuring element for dilation
$E_C \leftarrow 10$;           %Disk structuring element for closing

Bounding Box extraction:
$I_{open} \leftarrow (s_i \ominus E_O) \oplus E_O$
$I_{dil} \leftarrow I_{open} \oplus E_D$
$I_{close} \leftarrow (I_{dil} \oplus E_C) \ominus E_C$
$th \leftarrow Otsu(I_{close})$
$Mask \leftarrow I_{close} \geq th$
$B_i \leftarrow Rectangle(Mask)$
```

Once all $B_{i,j}$ were achieved, m was defined by the dimensions of the largest $B_{i,j}$, according to Algorithm 2. Then, we extracted a new set of slides $I = \{I_1, I_2, ..., I_i, ...I_P\}$, where $I_i \subset s_i \in v_j$ corresponds to the region $m \times N$ centred on the computed B_i, as detailed in Algorithm 2. After the conditioning phase, each OCT volume was defined by $P = 64$ B-scans of dimensions $m \times N = 550 \times 512$. However, to adapt the input dimensionality to the trained circumpapillary feature extractor, we resized each new slide $I_i \in v_i$ to 248×384 pixels, as illustrated in Figure 4.

Algorithm 2: Data-volume conditioning to remove useless pixels from the slides of the SD-OCT volumes.

```
Data: Slides $S \in V$ with dimensions $M \times N$.
Function: $centroid$, to extract the centroid of the $B \subset s_i \in v_j$.
Result: New slides $I \subset S \in V$ with dimensions $m \times N$.

Data conditioning:
for $j \leftarrow 1$ to Q do
    for $i \leftarrow 1$ to P do
        $B_{i,j} \leftarrow ROIret$ from $s_i \in v_j$ ;
        $x, y \leftarrow centroid$ from $B \subset s_i \in v_j$ ;
        $c_{i,j} \leftarrow x$ ;
        $d_{i,j} \leftarrow |B(1,1) - B(end,1)|$ ;
$m \leftarrow MAX(d)$ ;
for $j \leftarrow 1$ to Q do
    for $i \leftarrow 1$ to P do
        $I_{i,j} \leftarrow s_i(c_{i,j} - \frac{m}{2}$ to $c_{i,j} + \frac{m}{2}, 1$ to $N) \in v_j$ ;
```

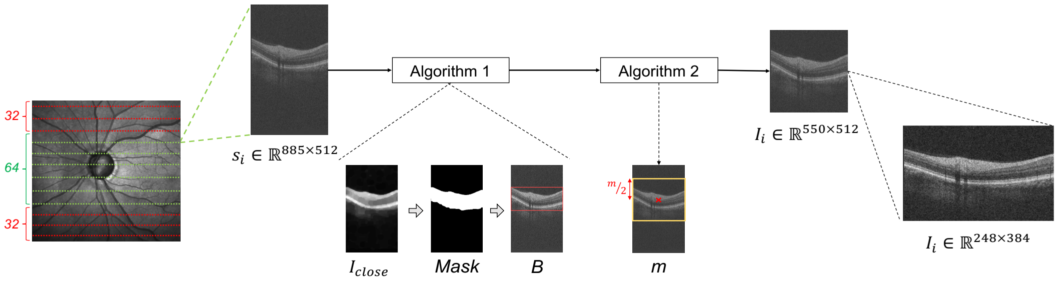

32
64
32
$s_i \in \mathbb{R}^{885\times512}$
Algorithm 1
Algorithm 2
$I_i \in \mathbb{R}^{550\times512}$
$I_i \in \mathbb{R}^{248\times384}$
$m/2$
I_{close}
Mask
B
m

Figure 4. Complete process to adapt the B-scans of the volumes to the slide-level discriminator dimensions. This new set of volumes of dimensions $P \times m \times N$ will constitute the input to the feature extraction stage.

3.2.2. LSTM network construction

In this stage, we propose the use of LSTM networks to feed the model with the spatial dependencies of the features extracted in the latent space from each new slide $I_i \in v_j$. Specifically, LSTM is a kind of Recurrent Neural Network (RNN) for sequence modelling, widely used in handwriting recognition [40], speech recognition [41] or video classification [42] tasks, among others. Unlike traditional RNNs, LSTM networks contain a memory cell c_t able to accumulate the state information to avoid the long-term dependency problem. A common LSTM unit is composed of a series of *gates* that control the flow of information around the cell. An *input gate* i_t regulates the new information that enters the cell to be accumulated. The activation of a *forget gate* f_t determines whether the past cell status c_{t-1} is forgotten or not. Finally, an *output gate* o_t controls the propagation of the latest cell output c_t to the final state h_t, being t each temporary instance.

In this paper, we follow the CNN-LSTM strategy carried out in [42] to consider sequences of CNN activations. Unlike the aforementioned video-based study, the proposed approach is intended to OCT-volume classification, so we consider each slide as a frame, i.e. each spatial dependency as a temporary instance. Therefore, once slide-level discriminator and data-volume conditioning stages were performed in the previous sections, we used the pre-trained base model to carry out the feature extraction from each conditioned B-scan of the SD-OCT volumes. As illustrated in Figure 3, given an input image $I_i \in v_j$, a feature volume f_i in the latent space was obtained after the feature extraction phase. Then, 1D array a_i was generated from each feature volume f_i by flattening their dimensions $7 \times 12 \times 512$, which correspond to the output of the last 1×1 convolution layer of the slide-level discriminator (see Figure 5). In this way, the inputs to the LSTM network consists of an array $A = \{a_1, a_2, ..., a_i, ..., a_P\}$, being $P = 64$ the number of slides per volume (see Figure 5). Otherwise, the output of each LSTM memory cell h_i corresponds to the concatenation of all outputs h_{i_u} obtained from each LSTM unit u, so that $h_i = [h_{i_1}, h_{i_2}, ..., h_{i_u}, ..., h_{i_U}]$, where U is the number of specified LSTM units, as deduced from Figure 5. Note that each h_i constitutes the input (together with the a_{i+1}) to the next LSTM memory cell c_t, which is graphically represented by the discontinue lines in Figure 5. Traditional LSTM networks can return, per volume v_j, either a set of all spatial dependencies $H_j = \{h_1, h_2, ..., h_i, ..., h_P\}$ or just the last one h_P, which contains information from all the previous B-scans. As a novelty, we developed in this case a sequential-weighting module (SWM) that allows taking into account all LSTM outputs by weighting them, in a sequential way, to provide a holistic feature vector O_j before the top model, according to Algorithm 3. Firstly, a flatten operation was applied to concatenate the LSTM outputs $H_j = \{h_1, h_2, ..., h_i, ..., h_P\}$ into an array R_j of length $T = P * U$, as illustrated in Figure 5. Then, we included a weighting layer $W = \{\frac{1}{T}, \frac{2}{T}, ..., \frac{k}{T}, ..., 1\}$, with $k = \{1, 2, 3, ..., T\}$, to generate a vector $L_j = R_j \circ W$, so that L_j became a weighting output from the initial LSTM output vector H_j. Additionally, as observed in Figure 5, a skip-connection module was defined to perform a spatial squeeze of F_j by means of a 3D global average-pooling layer (3DGAP), which generates an array Z_j that allows combining the set of features F_j embedded in the latent space with the weighted LSTM outputs L_j. Finally, a holistic feature vector O_j was obtained per volume by concatenating Z_j and L_j outputs (see Figure 5). Regarding the top model structure, a classifier based on a Multi-Layer Perceptron (MLP) was implemented to achieve the probability p_j corresponding to the class predicted from each specific volume v_j, as specified in Algorithm 3. It should be remarked that aspects related to the hyper-parameters of the architecture and the top model are detailed in Section 4.2.2.

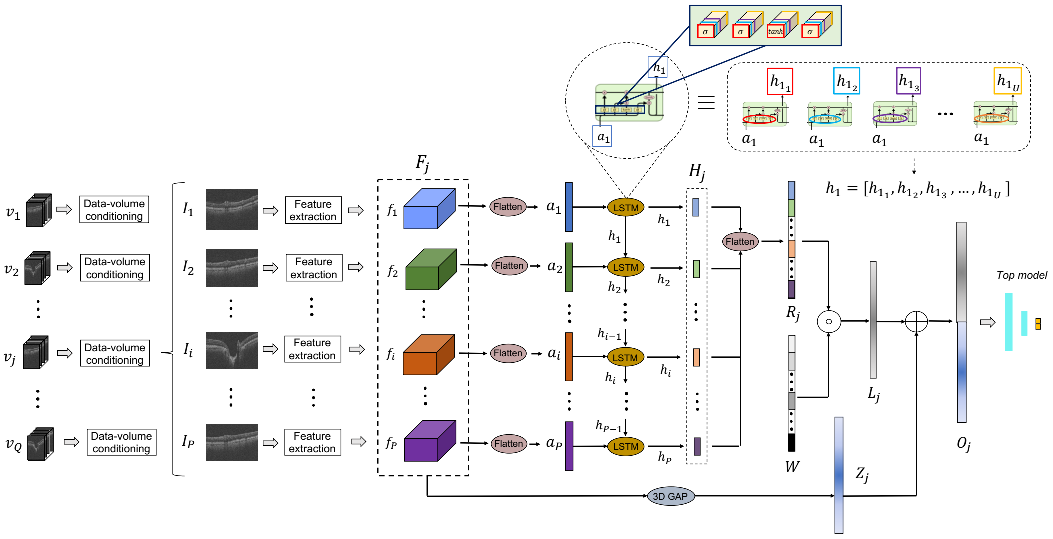


Figure 5. Illustration of the proposed end-to-end CNN-LSTM approach to address the glaucoma detection just from raw SD-OCT volumes centred on the ONH.

4. Results

In this stage, we describe separately the experiments carried out to develop the 2D CNN-based feature extractor, and those performed to achieve the volume-based predictive model. For both stages, we detail three well-differentiated sections: *data partitioning*, *validation phase* and *prediction stage*. Additionally, in the case of the slide-level discriminator, we also report the results from an *external validation* to demonstrate that the proposed feature extractor can generalise to other databases and, therefore, it can be used to extract the features from the slides of the ONH-centred SD-OCT volumes. Otherwise, with respect to the volume-based predictive model, we show the Class Activation Maps (CAMs) generated by the best approach to identify the regions in which the proposed model was paying attention for classifying each OCT volume, as healthy or glaucomatous, during the prediction stage.

Algorithm 3: Proposed methodology to predict glaucoma per volume.

```
Data: $P$ slides per $Q$ raw OCT volumes centred on ONH.
Functions: $DC \equiv$ Data Conditioning stage.
$FE \equiv$ Feature Extraction phase
$MLP \equiv$ Multi-Layer Perceptron classifier
Result: Predictions $p_j$ from each SD-OCT volume $v_j$.

End-to-end LSTM approach:
for $j \leftarrow 1$ to Q do
    for $i \leftarrow 1$ to P do
        $I_i \leftarrow DC(s_i)$, where $s_i \in v_j$ ;
        $f_i \leftarrow FE(I_i)$, where $I_i \subset s_i \in v_j$ ;
        $a_i \leftarrow Flatten(f_i)$ ;
        $h_i \leftarrow LSTM(a_i)$ ;
    % Sequential-Weighting Module (SWM):
    $R_j \leftarrow Flatten(H_j)$ ;
    $L_j \leftarrow R_j \circ W$ ;
    $Z_j \leftarrow 3DGAP(F_j)$;
    $O_j \leftarrow Concatenate(L_j, Z_j)$ ;
    % Top model:
    $p_j \leftarrow MLP(O_j)$ ;
```

4.1. Experiments for the slide-level feature extractor

4.1.1. Data partitioning

To provide robust and reliable results about the feature extractor, we performed a patient-level data partitioning of the *circ-DB-1* database. In particular, 52 circumpapillary images (20 glaucomatous and 32 healthy) were grouped in an independent set to test the model. With the rest of the data (training set), we performed an internal 5-fold cross-validation technique to optimise the hyper-parameters of the neural networks. Specifically, in each iteration, $\frac{4}{5}$ of the training set (58 glaucomatous and 99 healthy eyes) were employed to train a specific model and $\frac{1}{5}$ (15 with glaucoma and 25 normal images) to validate it in order to monitor and prevent overfitting. Once the five iterations were attained, we used the entire training data set (197 circumpapillary samples) to train the final model with the architecture and parameters that reported the best performance during the internal cross-validation (ICV) stage. The final model was validated with the test set and evaluated with the external circ-DB-2 database, which is composed of 143 glaucomatous and 193 healthy circumpapillary OCT images.

4.1.2. Validation phase

At this point, it should be remembered that, in our previous work [18], we carried out a rigorous comparison between different neural networks for glaucoma detection just from raw circumpapillary OCT images. For this reason, we detail in this section a new comparison between the VGG family of networks, which reported the best performance in [18], and the proposed *Residual Attention Glaucoma Network* (RAGNet).

Ablation experiments. To accomplish the empirical exploration of the optimal hyper-parameter combination for the feature extractor, we made a sweep of different values that we went refining during the training phase. In Table 3, we show a summary of the main parameters considered for this study, as well as the range of values used during the exploration. Otherwise, in Table 4, we expose the final hyper-parameters and components selected to carry out each approach under study. The abbreviations set out in the Table 3 correspond as follows: SGD - Stochastic Gradient Descent; MSE - Mean Squared Error; WBCE - Weighted Binary Cross Entropy; MLP - Multi-Layer Perceptron; GMP - Global Max Pooling; GAP - Global Average Pooling.

Table 3. Summary of the main hyper-parameters and CNN configurations considered during the training experiments to build the slide-level discriminator.

Model hyper-parameters	**Range**	**Top model**	**Range**
Learning rate	$[5e^{-1}, 1e^{-5}]$	*Initial dropout*	[0, 0.5]
Optimizer	SGD Adadelta Adam	*Structure*	Flatten + MLP GMP + Dense GAP + Dense
Loss function	MSE WBCE Hinge	*Final Activation*	Softmax Sigmoid
Batch size	8, 16, 32, 64	**Fine-tuning**	**Range**
Number of epochs	[50, 500]	*Unfrozen blocks*	1, 2, 3, 4, 5

Table 4. Hyper-parameters and components selected after the empirical exploration carried out during the training of each feature-extractor approach.

		VGG16	**VGG19**	**RAGnet (with VGG16)**	**RAGnet (with VGG19)**
Model hyperparameters	*Learning rate*	0.001	0.001	0.0005	0.0005
	Optimizer	Adadelta	Adadelta	SGD	SGD
	Loss function	WBCE	WBCE	WBCE	WBCE
	Batch size	16	16	16	16
	Number of epochs	125	125	120	120
Top Model	*Initial dropout*	0.4	0.4	0.4	0.4
	Structure	GAP	GAP	GAP	GAP
	Final activation	Softmax	Softmax	Softmax	Softmax
Fine-tuning	*Unfrozen blocks*	3	3	3	3

Based on the experiments attained in [18], we performed a deep fine-tuning strategy of the VGG16 and VGG19 architectures, since they reported the best results. For both CNNs, data augmentation techniques [43] were implemented to face the overfitting problem by increasing the number of images of the database with synthetic samples. A factor ratio of 0.2 was applied to perform random geometric and dense elastic transformations from the original images, according to [18]. As observed in Table 4, we unfroze the two last convolutional blocks of the VGGs to retrain the weights with the specific information contained in the circumpapillary OCT images. Additionally, a weighted binary cross-entropy (WBCE) was used as a loss function by employing an optimal balanced factor $\alpha = [1.35, 0.79]$, following the Equations (2) and (3), to alleviate the unbalanced problem between glaucoma and healthy classes, respectively (see Table 4). The models reached the best performance when they were trained during 125 epochs trying to minimise the WBCE loss function, using Adadelta optimiser with a learning rate of 0.001 and a batch size of 16. It is noticeable that we added an initial dropout layer with a coefficient of 0.4 before the top model, as exposed in Table 4.

$$WBCE = -\alpha_2 \; y_i \; log(\hat{y}_i) - \alpha_1(1 - y_i)log(1 - \hat{y}_i) \;, \quad being \tag{2}$$

$$\alpha_c = \frac{Ns}{Nc \sum y_i^c} \;, \qquad with \; c \in [1, 2] \tag{3}$$

where $\hat{y}$ and y represent the outputs and the ground truth, respectively. Ns denotes the total number of samples and $Nc = 2$ corresponds to the number of classes. Note that $c = 1$ and $c = 2$ are associated with glaucoma and healthy classes, in that order.

Regarding the proposed model (RAGNet), we performed the same deep fine-tuning strategy as before, i.e. we only retrained the two last blocks of the VGG networks, according to Table 4. Besides, the same data augmentation processing and weighted loss function were specified to make an objective comparison. The major innovation of the proposed method lies in the inclusion of the residual M_{res} and attention M_{att} modules, whose filters and dimensions parameters were exposed in Figure 2. The combination of the best hyper-parameters was carried out following the same empirical exploration as before (see Table 3). In this case, we trained the models (using VGG16 and VGG19 architectures as a base model) during 120 epochs, instead of 125 like in the VGGs case. WBCE (Equation 2) was selected as the loss function to be minimised, using SGD optimiser with a learning rate of 0.0005 and a batch size of 16 (see Table 4). As before, we also included a dropout layer of 0.4 to address the classification stage.

Concerning the top model, we made use of the same structure for all approaches, which is described in Table 4. Particularly, we included a Global Average Pooling (GAP) layer to obtain a spatial squeeze before the softmax-activated dense layer, which is composed of two neurons to predict healthy or glaucoma for each sample.

Quantitative results. Different figures of merit, such as sensitivity (SN), specificity (SP), positive predictive value (PPV), negative predictive value (NPV), F-score (FS) and area under the ROC curve (AUC) are reported in Table 5 to provide objective results. Note that AUC metric was calculated by means of a polynomial approximation using a gradient descent method, according to [44]. As a novelty, we also calculated quantitative metrics related to the learning curves achieved during the training of the models (see Table 6). In particular, we consider the validation accuracy (ACC) and loss (LOSS) values and propose two additional measures to quantify the overfitting (OVFT) and the quality (QLTY) of the validation learning curves. According to Equation (4), OVFT indicator aims to provide measurable information related to the generalisation ability of the models, so that the closer OVFT is to 0 the better. OVFT < 0 indicates underfitting and OVFT > 0 supposes overfitting, proportionally.

$$OVFT = \frac{1}{\epsilon - p} \sum_{e=p}^{\epsilon} (VL_e - TL_e), \quad with \quad OVFT \in [-\infty, \infty] \tag{4}$$

VL_e and TL_e correspond to the validation and training loss curves, respectively, at the epoch e. Additionally, ϵ denotes the total number of epochs and p represents the position (epoch) in which the validation loss curve VL reaches the global minimum.

Otherwise, QLTY metric provides information about how stable is the model and how much does it learn. Therefore, we measured the stability (STBL) of the model taking into account the variations between adjacent epochs of the validation loss curve. Specially, STBL was calculated according to Equation (5), where $D = \{d_1, d_2, ..., d_e, ..., d_{\epsilon-1}\}$ is an array containing the differences between adjacent epochs, so that $d_e = VL_e - VL_{e+1}$ denotes the difference between the values of the validation loss curve achieved for the epochs e and $e + 1$. The closer STBL is to 0, the more stable is the validation curve. Regarding the amount learned by the model (LRNG), it was calculated as the difference between the validation loss values at the initial and final epochs, according to Equation (6). In this case, negative values correspond to a significant overfitting, whereas higher values denote greater learning. Finally, QLTY metric was achieved following the Equation (7), where LRNG and STBL measures were combined to evaluate the quality of the model training. In this way, lower variations of validation loss between adjacent epochs and higher positive differences between initial and final epochs would entail a QLTY value closer to 1, which is associated with a more reliable model.

$$STBL = \sqrt{\frac{\sum_{e=1}^{\epsilon-1}(d_e - \bar{D})^2}{\epsilon - 1}}, \quad with \quad STBL \in [0, \infty] \tag{5}$$

$$LRNG = VL_1 - VL_{\epsilon-1}, \quad with \quad LRNG \in [-\infty, \infty] \tag{6}$$

$$QLTY = \begin{cases} \frac{1}{1+e^{-\frac{LRNG}{STBL}}} & \text{if} \ \ STBL > 0 \\ \frac{1}{1+e^{-\frac{LRNG}{10^{-\epsilon}}}} & \text{if} \ \ STBL = 0 \end{cases} \quad with \quad QLTY \in [0, 1] \tag{7}$$

Table 5. Quantitative results reached from the *circ-DB-1* database during the internal cross-validation stage with a confidence level of 95%. Metrics are reported in terms of average ± standard deviation followed by the confidence intervals (CIs).

	VGG16	**VGG19**	**RAGNet-VGG16**	**RAGNet-VGG19**
SN	0.81±0.11 (0.70-0.91)	0.77±0.17 (0.60-0.93)	**0.86±0.11 (0.75-0.97)**	0.73±0.12 (0.61-0.84)
SP	0.96±0.03 (0.93-0.99)	0.95±0.03 (0.92-0.98)	0.95±0.04 (0.91-0.99)	**0.97±0.02 (0.95-0.98)**
PPV	0.92±0.06 (0.87.0.98)	0.90±0.07 (0.84-0.97)	0.92±0.07 (0.86-0.99)	**0.93±0.04 (0.89-0.97)**
NPV	0.90±0.06 (0.84-0.95)	0.88±0.08 (0.80-0.96)	**0.93±0.05 (0.87-0.98)**	0.86±0.05 (0.81-0.91)
FS	0.86±0.08 (0.79-0.93)	0.82±0.11 (0.72-0.93)	**0.88±0.04 (0.85-0.92)**	0.81±0.08 (0.73-0.89)
AUC	0.88±0.05 (0.85-0.91)	0.86±0.08 (0.83-0.88)	**0.91±0.04 (0.87-0-94)**	0.85±0.06 (0.82-0.88)

4.1.3. Prediction stage

In this section, we expose the results for the prediction of the primary and external test sets. Specifically, the results achieved when evaluating the primary test set from the *circ-DB-1* database are reported in Table 7). Additionally, the results corresponding to the external validation are exposed in Table 8. The goal of this section is to demonstrate that the proposed slide-level discriminator could be valid to perform the feature extraction from the B-scans of the SD-OCT volumes. Therefore, we made use

Table 6. Performance of the learning curves obtained from the *circ-DB-1* database during the internal cross-validation stage with a confidence level of 95%. Metrics are reported in terms of average ± standard deviation followed by the confidence intervals.

	VGG16	**VGG19**	**RAGNet-VGG16**	**RAGNet-VGG19**
ACC	0.90±0.05 (0.86-0.95)	0.88±0.06 (0.82-0.94)	**0.92±0.02 (0.90-0.94)**	0.89±0.05 (0.83-0.92)
LOSS	0.27±0.11 (0.16-0.37)	0.34±0.18 (0.17-0.5)	**0.25±0.09 (0.16-0.35)**	0.38±0.18 (0.2-0.54)
OVFT	**0.11±0.10 (0.02-0.20)**	0.16±0.14 (0.02-0.29)	0.12±0.08 (0.04-0.20)	0.17±0.10 (0.08-0.26)
STBL	**0.01±0.01 (0.01-0.02)**	0.02±0.01 (0.02-0.03)	0.07±0.03 (0.05-0.10)	0.08±0.01 (0.07-0.08)
QLTY	**1±3e^{-13} (0.99-1)**	0.99±0.01 (0.98-1)	1±2e^{-3} (0.99-1)	0.89±0.21 (0.69-1)

of the circ-DB-2 database as an external test set to check how did the proposed feature extractor work with new OCT samples centred on the optic nerve head (ONH).

Table 7. Comparison between different strategies carried out during the prediction of the primary test set from *circ-DB-1* database.

	VGG16	**VGG19**	**RAGNet-VGG16**	**RAGNet-VGG19**
SN	0.8500	0.8500	0.8500	**0.9000**
SP	0.8750	0.8438	**0.9375**	0.8750
PPV	0.8095	0.7727	**0.8947**	0.8182
NPV	0.9032	0.9000	0.9091	**0.9333**
FS	0.8293	0.8095	**0.8718**	0.8571
AUC	0.8625	0.8469	**0.8938**	0.8875
ACC	0.8654	0.8462	**0.9038**	0.8846

Table 8. Comparison of the results achieved from the different methods when testing the external circ-DB-2 database.

	VGG16	**VGG19**	**RAGNet-VGG16**	**RAGNet-VGG19**
SN	0.8741	**0.8951**	**0.8951**	0.8741
SP	0.8446	0.8238	**0.8756**	0.8653
PPV	0.8065	0.7901	**0.8432**	0.8278
NPV	0.9006	0.9138	**0.9185**	0.9017
FS	0.8389	0.8393	**0.8678**	0.8503
AUC	0.8593	0.8595	**0.8789**	0.8697
ACC	0.8571	0.8542	**0.8839**	0.8690

4.2. Experiments for the volume-based predictive model

4.2.1. Data partitioning

As before, we also carried out a data partitioning stage of the *vol-DB-3* database to guarantee the rigour of the experiments performed with the SD-OCT volumes. Particularly, $\frac{1}{5}$ of the data (29 glaucomatous and 35 healthy) was used as a test set to evaluate the proposed model. The rest of the database was used to train the model through a 5-fold cross-validation technique. In each of the five iterations, $\frac{4}{5}$ of the training data (92 glaucomatous and 113 healthy eyes) were employed to develop a specific model, whereas $\frac{1}{5}$ (23 with glaucoma and 28 normal volumes) was used as a validation set to control overfitting and optimise the hyper-parameters. The final model was built via training the best architecture (optimised during the ICV stage) with the samples from both training and validation sets.

4.2.2. Validation phase

In this section, we detail and compare the structure and the hyper-parameters that compose the proposed architecture in relation to those of the state of the art. Specifically, we show the differences between the developed method and other basic LSTM structures to evidence the added value that the proposed sequential-weighting module (SWM) introduces for glaucoma detection via SD-OCT volumes. Note that RAGNet architecture (via fine-tuning the VGG16 network) was selected as the base model to extract the features from which to carry out all the experiments of this section.

Ablation experiments. To perform a comparison as reliable as possible between different approaches, we established some fixed conditions by means of an initial random exploration of several hyper-parameters. Firstly, we fixed the input and output dimensions, as detailed in Table 9, to elucidate which method extracted the most discriminatory features. Regarding the model hyper-parameters, a nested loop was sweeping different loss functions, gradient-based learning algorithms and sizes of batches to select those that achieved the best performance. In particular, we found weighted binary cross-entropy (WBCE) loss function, Adadelta optimiser and batch size of 16 the best hyper-parameter combination to address the next phase. Note that, in this case, we calculated an optimal weighting factor $\alpha = [1.11, 0.91]$ to balance the glaucomatous and healthy samples, respectively, during the training of the models (see Table 9). Otherwise, the number of training epochs and the learning rate were specified depending on the approach. Concerning the LSTM architecture, we fixed specific hyper-parameters, such as an input dropout of 0.3 to prevent overfitting, whereas the number of LSTM units was adapted to each approximation to provide a holistic feature map of size $T = 512$ before the classification stage. Also, we specified a constant top-model structure defined by a multi-layer perceptron (MLP) containing two fully-connected layers with 256 and 32 neurons, followed each one by a dropout layer with a coefficient of 0.25. Finally, a softmax layer with two neurons (healthy and glaucoma) was implemented to achieve the predictions per volume, as collected in Table 9.

Once the aforementioned hyper-parameters were established, we compared several useful LSTM options such as the shape of the final LSTM outputs or the use of Bidirectional layers, as proposed in [14], where the authors also performed a CNN-LSTM strategy to identify biomarkers associated with the age-related macular degeneration (AMD). Regarding the output shape, basic LSTM networks can provide 3D or 2D arrays depending on whether all LSTM outputs H_j are considered or just the last one h_P, which contains information from all the previous slides. Both approaches are analysed in this study, under the name of *OS3D* and *OS2D*, respectively (see Tables 10, 11). The use of bidirectional layers (Bi) was also contemplated for both previous approaches to determine the performance of this kind of layers. As evidenced in Tables 10, 11, bidirectional layers provide a slight outperforming, so we carried out another experiment, with the best-reported conditions, based on stacked LSTM layers in order to evaluate more complex and deeper architectures. Note that for all the aforementioned experiments, the models were trained during 60 epochs with a learning rate of 0.01, trying to optimise the compromise between accuracy and overfitting metrics. The number of LSTM units defined to adapt the size of the feature embedding space to $T = 512$ were 8, 512, 4, 256 and 1×(256, 512) for *OS3D*, *OS2D*, *Bi+OS3D*, *Bi+OS2D* and *stacked Bi+OS2D* approaches, respectively. Note that *OS3D*-based models re-

Table 9. List of hyper-parameters used to train the volume-based model for the different approaches. Remember that Q is the number of volumes with P B-scans that compose the *vol-DB-3* database.

Data shape	
Input shape	$Q \times P \times 248 \times 384 \times 3$
Output shape	$Q \times 512$ (before the top model)
Model hyper-parameters	
Loss function	WBCE
Weighting factor	α=[1.11, 0.91]
Optimiser	Adadelta
Batch size	16
Number of epochs	*Variable* → [50, 500]
Learning rate	*Variable* → $[5e^{-1}, 1e^{-5}]$
Architecture hyper-parameters	
Feature extractor	RAGNet (with VGG16)
Input dropout	0.3
LSTM units	*Variable* → 4, 8, 256, 512
Top model	
Dense units	1×(256, 32)
Dropout coefficients	2×(0.25)
Final activation	Softmax (2 neurons)

quire an extra layer to flattening the features extracted from all slides, unlike the *OS2D*-based methods which only output the features that come directly from the last slide. For this reason, the LSTM units related to *OS3D*-based models need to be lower than those associated with *OS2D*-based approaches. Worth noting that bidirectional layers are a kind of generative deep learning that allows creating a reversed copy of the input sequence. Therefore, when bidirectional layers are included, LSTM units must be halved to keep the output dimensionality.

Regarding the proposed strategy (RAGNet-VGG16 + LSTM + SWM), we kept most of the hyper-parameters constant, but thanks to the developed sequential-weighting module (SWM), it was possible to decrease the learning rate to perform a more stable training during more epochs, without reporting overfitting. Specifically, we trained the models during 150 epochs with a learning rate of 0.005. Additionally, since bidirectional layers worked better in the previous approaches, we performed an additional experiment to check them in combination with the designed SWM-based model (see Tables 10, 11). Following the same criteria as before, the number of LSTM units specified for the proposed method with and without bidirectional layers was 4 and 8, respectively.

Additionally, an extra experiment was carried out to compare the performance for the end prediction when using the proposed RAGNet-VGG16 or the traditional VGG16 feature extractor. This experiment aims to know how large the gap in the final performance is between both slide-level discriminators when extracting the features of the latent space from each slide of the OCT volume. To accomplish this section, we fixed the same hyper-parameters and network structure as in the proposed method (RAGNet-VGG16 + SWM). The results reached using the traditional feature extractor during the cross-validation stage can be observed in Tables 10 and 11, under the name of *VGG16 + SWM*.

Quantitative results. In this case, we also show the results provided by the different approaches during the ICV stage of the volume-based predictive model. All figures of merit related to the training of the models are collected in Table 10, whereas the metrics corresponding to the validation learning curves are exposed in Table 11.

Table 10. Quantitative results reached from *vol-DB-3* database during the ICV of the volume-based predictive model with a confidence level of 95%. Metrics are reported in terms of average ± standard deviation followed by the CIs.

	SN	**SP**	**PPV**	**NPV**	**FS**	**AUC**
OS3D	0.68±0.04 (0.64-0.72)	0.77±0.09 (0.68-0.86)	0.72±0.08 (0.63-0.80)	0.75±0.03 (0.71-0.78)	0.70±0.065 (0.65-0.74)	0.73±0.05 (0.68-0.79)
OS2D	0.73±0.12 (0.60-0.86)	0.79±0.03 (0.77-0.82)	0.74±0.05 (0.69-0.79)	0.79±0.08 (0.71-0.87)	0.73±0.08 (0.64-0.82)	0.76±0.07 (0.70-0.81)
Bi+OS3D	0.72±0.05 (0.67-0.77)	0.75±0.08 (0.68-0.82)	0.71±0.08 (0.63-0.78)	0.77±0.05 (0.72-0.81)	0.71±0.06 (0.66-0.77)	0.74±0.05 (0.68-0.80)
Bi+OS2D	0.66±0.14 (0.51-0.81)	**0.85±0.11 (0.74-0.97)**	**0.82±0.12 (0.69-0.95)**	0.76±0.05 (0.71-0.82)	0.71±0.07 (0.63-0.79)	0.76±0.04 (0.72-0.80)
Stacked Bi+OS2D	0.75±0.13 (0.63-0.86)	0.71±0.20 (0.52-0.90)	0.71±0.11 (0.60-0.81)	0.79±0.06 (0.73-0.85)	0.72±0.06 (0.66-0.77)	0.73±0.06 (0.68-0.76)
Proposed method (RAGNet + SWM)	**0.76±0.11 (0.64-0.87)**	0.79±0.08 (0.70-0.87)	0.75±0.06 (0.68-0.82)	**0.81±0.07 (0.74-0.88)**	**0.75±0.05 (0.70-0.80)**	**0.78±0.04 (0.72-0.82)**
Proposed method (RAGNet + SWM) + Bi	0.75±0.09 (0.66-0.83)	0.77±0.15 (0.63-0.91)	0.75±0.12 (0.64-0.86)	0.76±0.06 (0.74-0.85)	0.74±0.07 (0.68-0.80)	0.76±0.05 (0.71-0.80)
VGG16 + SWM	0.70±0.11 (0.59-0.80)	0.69±0.05 (0.64-0.73)	0.64±0.04 (0.60-0.68)	0.74±0.06 (0.68-0.80)	0.67±0.07 (0.60-0.73)	0.69±0.04 (0.65-0.73)

Table 11. Results corresponding to the learning curves achieved during the validation of the trained models from *vol-DB-3* database with a confidence level of 95%. Metrics are reported in terms of average ± standard deviation followed by the CIs.

	ACC	**LOSS**	**OVFT**	**STBL**	**QLTY**
OS3D	0.71±0.06 (0.65-0.78)	0.57±0.09 (0.48-0.67)	0.28±0.08 (0.19-0.37)	0.04±0.02 (0.02-0.07)	0.82±0.21 (0.61-1)
OS2D	0.75±0.07 (0.68-0.82)	0.54±0.16 (0.38-0.72)	0.27±0.13 (0.14-0.40)	0.12±0.04 (0.08-0.16)	0.70±0.29 (0.42-0.98)
Bi+OS3D	0.74±0.05 (0.68-0.80)	0.53±0.08 (0.45-0.61)	0.21±0.08 (0.13-0.30)	0.05±0.01 (0.04-0.05)	0.91±0.16 (0.75-1)
Bi+OS2D	0.77±0.33 (0.73-0.80)	0.55±0.12 (0.42-0.67)	0.28±0.08 (0.19-0.37)	0.15±0.01 (0.14-0.16)	0.71±0.17 (0.55-0.88)
Stacked Bi+OS2D	0.73±0.07 (0.66-0.80)	0.59±0.18 (0.40-0.79)	0.20±0.10 (0.10-0.31)	0.10±0.02 (0.09-0.12)	0.66±0.35 (0.33-0.99)
Proposed method (RAGNet + SWM)	**0.77±0.04 (0.73-0.81)**	**0.49±0.09 (0.39-0.59)**	**0.13±0.09 (0.04-0.23)**	0.02±0.01 (0.01-0.03)	0.99±0.03 (0.96-1)
Proposed method (RAGNet + SWM) + Bi	0.77±0.07 (0.69-0.84)	0.53±0.15 (0.37-0.69)	0.18±0.11 (0.11-0.26)	0.05±0.01 (0.04-0.07)	0.92±0.13 (0.77-1)
VGG16 + SWM	0.69±0.04 (0.65-0.73)	0.64±0.07 (0.57-0.71)	0.41±0.06 (0.35-0.47)	**0.01±0.01 (0.01-0.02)**	**0.99±0.01 (0.99-1)**

4.2.3. Prediction stage

This section consists of two parts. In the first one, we compare the results achieved on the *vol-DB-3* test set from the different approaches (see Table 12) and secondly, we show the computed Class Activation Maps (CAMs), which highlight the regions of interest obtained from the SD-OCT volumes. Specifically, CAMs allow identifying the areas in which the proposed deep-learning model was paying attention for classifying each OCT volume, as healthy or glaucomatous, during the prediction stage. Since the proposed SWM-based method outperforms the rest of approaches, we report the CAMs extracted from the test set making use of the developed volume-based predictive model (without bidirectional layers). The combination of CAMs and LSTM units allows detecting the ROIs of each B-scan, as well as the key slides inside the volume to address the glaucoma diagnosis. In Figure 6, we show several examples of the heat maps generated by the model from random SD-OCT volumes of the test set. In particular, a sweep of several heat maps of representative slides I_i corresponding to four randomly selected volumes from each class v_r^c, being $r \in [1, Q]$ a random integer and c the class, are exposed to elucidate the discriminating OCT regions depending on the class. It should be remarked that the hot-coloured areas indicate greater discriminatory power to detect the class under study.

4.3. Comparison with the state of the art

In this section, we aim to compare the proposed method with those used in other studies of the state of the art (SotA). It is important to highlight that, to the best of the authors' knowledge, there are no public databases of SD-OCT volumes to make possible a direct comparison. For this reason, in order to accomplish this section, we have faithfully replicated the experiments carried out in those SD-OCT volume-based

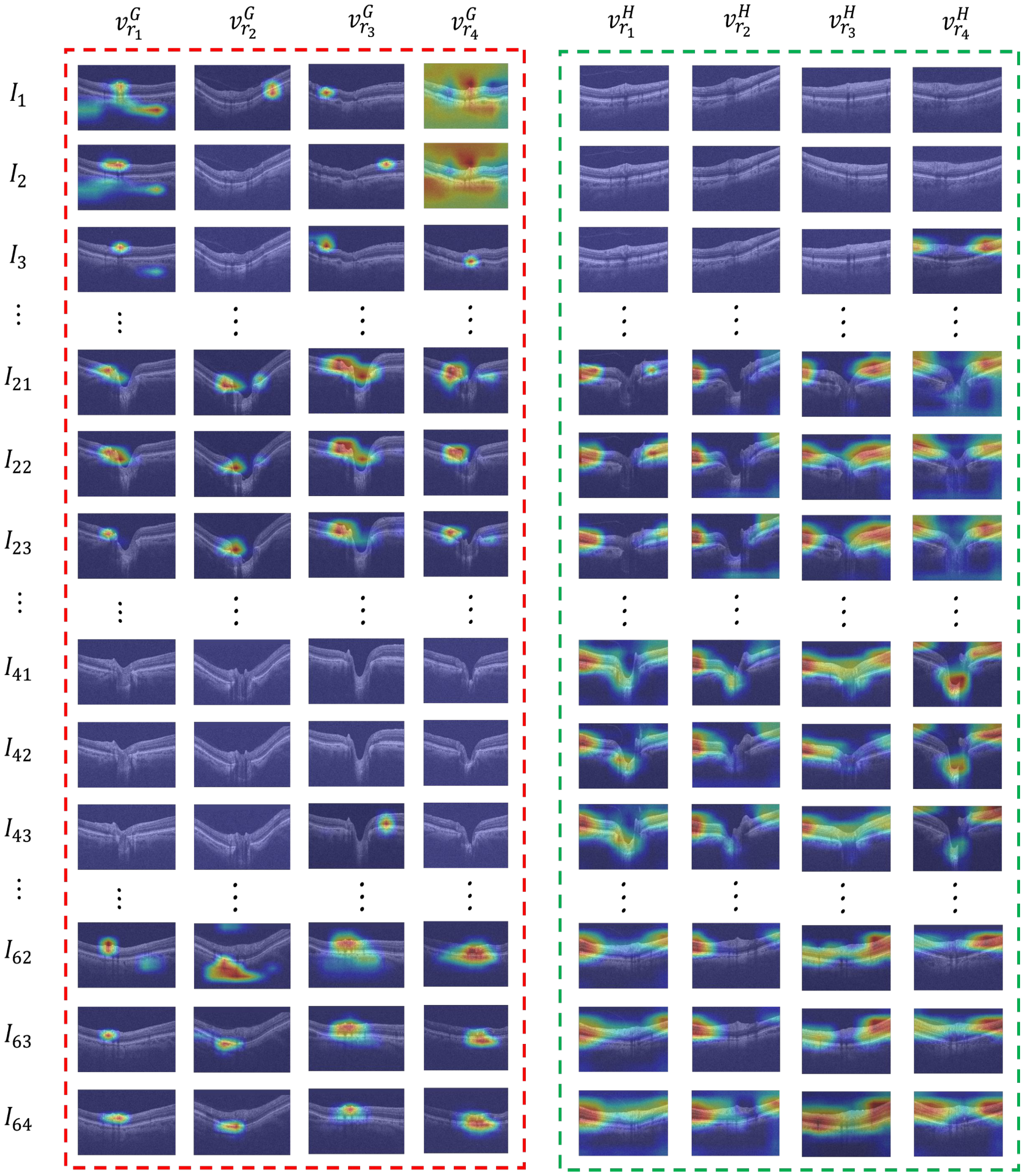


Figure 6. Heat maps generated by the proposed LSTM network to highlight the most important regions of each B-scan when classifying specific random volumes as glaucomatous or healthy samples. Four first columns of volumes $v_{r_1}^G, ..., v_{r_4}^G$ bounded by the red rectangle correspond to glaucomatous eyes, whereas the four last columns of volumes $v_{r_1}^H, ..., v_{r_4}^H$ marked with the green rectangle correspond to healthy samples.

Table 12. Comparison of the results achieved during the prediction of the primary test set from *vol-DB-3* volume-based database.

	SN	SP	PPV	NPV	FS	AUC	ACC
OS3D	0.7586	0.7982	0.7591	0.8000	0.7547	0.7793	0.7813
OS2D	0.6897	0.6857	0.6451	0.7272	0.6667	0.6877	0.6875
Bi+OS3D	0.7586	0.8286	0.7857	0.8056	0.7719	0.7936	0.7969
Bi+OS2D	0.7586	0.7714	0.7333	0.7941	0.7458	0.7650	0.7656
Stacked Bi+OS2D	0.72414	0.8286	0.7778	0.7838	0.7500	0.7764	0.7813
Proposed method (RAGNet + SWM)	0.7586	**0.8571**	**0.8148**	0.8108	**0.7857**	**0.8079**	**0.8125**
Proposed method (RAGNet + SWM) + Bi	0.7586	0.7714	0.7333	0.7941	0.7458	0.7650	0.7656
VGG16 + SWM	**0.8276**	0.6857	0.6857	**0.8276**	0.7500	0.7567	0.7500

works of the literature intended to glaucoma detection. As we mentioned in Section 1.1, two recent studies [27,29] could be directly compared with ours since they focused just on raw SD-OCT volumes to address the glaucoma detection, without including other variables extracted from the patient, such as visual field (VF), intraocular pressure (IOP), mean deviation (MD) or fundus images, among others.

Worth noting that the state-of-the-art experiments were replicated making use of the same deep-learning architecture and maintaining the original hyper-parameters during the training of the models (see Table 13). However, some specific conditions were adapted in each case to make possible the comparison between methods. In particular, batch size and loss function hyper-parameters were not reported in [27], so we defined them according to our work, i.e. a batch size of 16 and the weighted binary cross-entropy (WBCE) as a loss function. Otherwise, since the batch size hyper-parameter was also not reported in [29], we specified it to 4 in order to face the GPU memory problems associated with such a deep neural network. Regarding the architectures, we also conserved the same original structures to provide a comparison as objective as possible. In Table 14, we expose the test results via comparing the proposed methodology with the different state-of-the-art approaches intended to glaucoma detection just from raw SD-OCT volumes. Note that all the experiments were performed on an Intel i7 @4.00 GHz of 16 GB of RAM with a Titan V GPU of 12 GB of RAM.

Table 13. Hyper-parameters specified to perform the models' training from *vol-DB-3* database.

	Optimiser	Learning rate	Loss function	Batch size	Epochs
Maetschke et al.	Nadam	0.0001	WBCE	16	100
Ran et al.	Adam	0.0001	WBCE	4	40
Proposed Method	Adadelta	0.005	WBCE	16	150

5. Discussion

5.1. Discussion about the feature extractor

In contrast to the state-of-the-art studies, which performed the glaucoma detection from SD-OCT volumes through 3D architectures, in this paper, we propose a new

Table 14. Results reached by the state-of-the-art approaches in comparison with the proposed method when predicting the test set from *vol-DB-3* volume-based database.

	SN	SP	PPV	NPV	FS	AUC	ACC
Maetschke et al.	0.6207	0.8000	0.7200	0.7180	0.6667	0.7103	0.7188
Ran et al.	0.6348	0.7286	0.7771	0.7756	0.5858	0.6817	0.6873
Proposed method	**0.7586**	**0.8571**	**0.8148**	**0.8108**	**0.7857**	**0.8079**	**0.8125**

way of addressing this task by using the spatial dependencies between 2D images, instead of operating in the three-dimensional space. Thereby, we have developed a new slide-level discriminator able to extract the features from the slides of the SD-OCT volumes. At this point, it should be remarked the importance of using pre-trained networks when addressing small databases in order to leverage the information acquired by the weights of the architecture when it was trained on larger databases. The proposed *Residual Attention Glaucoma Network* (RAGNet) method was compared with other validated architectures, which reported the best performance in our previous work [18] for glaucoma detection using circumpapillary OCT samples. In Table 5, different figures of merit extracted from the five cross-validated iterations are exposed to compare the different approaches. The proposed RAGNet model, characterised by the combination of residual connections and convolutional attention blocks, reported a significant outperforming with respect to the rest of networks. Specifically, RAGNet model (via fine-tuning VGG16 architecture) achieved the best results for the most of metrics, except for SP and PPV measures, whose highest values were reached by the RAGNet model (with the fine-tuned VGG19). However, taking into account that this approach showed a significantly worse performance for the rest of metrics, and RAGNet model (with VGG16) provided SP and PPV values closely similar to the best approach, the proposed network could be considered superior. In any case, the inclusion of the proposed residual and attention modules outperforms the popular pre-trained architectures of the state of the art.

As a novelty, besides the accuracy and loss values, we also introduced new metrics, such as OVFT, STBL and QLTY related to the learning of the models, as observed in Table 6. These additional measures provide information about the generalisation ability of the models to predict new samples. In this case, RAGNet (with VGG16) was consolidated as the best network since it reported higher values for accuracy and loss metrics, besides those aforementioned. Additionally, for OVFT, STBL and QLTY measures the proposed model achieves closely similar results in relation with those reached by the best architecture (pre-trained VGG16). The small reported differences (0.01, 0.06 and 0.002 for OVFT, STBL and QLTY indicators, respectively) are negligible in the model's performance since all of them represented a stable learning, which is associated with a reliable predictive system, as can be deduced in Tables 7 and 8.

To verify how the models would work with new OCT samples, we carried out a prediction stage in which we evaluated the models on a primary test set from the *circ-DB-1* database (see Table 7). In line with the results obtained during the internal cross-validation stage, RAGNet-based approaches also surpassed the results for all figures of merit. In particular, RAGNet with VGG16 worked as a more specific model since it achieved higher values for SP and PPV metrics, whereas RAGNet with VGG19 provided better SN and NPV results. The rest of metrics (FS, AUC and ACC) offer information about the general behaviour of the model. In this case, RAGNet with VGG16 stood out for FS and ACC measures, but it reported lower values estimating

AUC. Nevertheless, both RAGNet-based methods showed excellent performance, with results around 0.9 for all figures of merit. Also, the results of the experiments make evident that the proposed RAGNet approach based on residual and skip-connections improved of the performance of the traditional networks previously validated in [18].

Moreover, in order to check the generalisation ability of the models, we performed an additional validation from an external database. It should be highlighted that this part is really important because the final success of the volume-based predictive model largely depends on the feature extractor performance. For this reason, it is necessary to validate the proposed slide-level discriminator with independent databases to ensure that it can predict independent OCT samples centred in the optic nerve head (ONH). Results corresponding to this stage are exposed in Table 8, where it can be appreciated that the proposed RAGNet (with fine-tuned VGG16 architecture) clearly achieved the best performance for all figures of merit. Furthermore, results reported closely similar values to those reached in the primary test set, which indicates that the proposed feature extractor is perfectly applicable to other databases. In addition, the designed slide-level discriminator is robust to the acquisition machine, which can be deduced from the exposed volume-level results. After this complete validation process, we found the proposed RAGNet model (with the fine-tuned VGG16) as the best feature extractor, surpassing the performance reported by the previously trained models in [18]. For this reason, we made use of the proposed slide-level discriminator to extract the latent space features from the B-scans of the SD-OCT volumes.

5.2. Discussion about the LSTM-based model

First of all, it is important to mention the data-volume conditioning stage performed to accomplish the experiments. In line with the state-of-the-art works [27, 28], a pre-processing stage was necessary to face the GPU memory problems due to a large amount of data contained in the SD-OCT volumes. However, unlike the aforementioned studies where each OCT scan of the database was a cube of resolution of $200 \times 200 \times 1024$, the proposed approach was addressed from volumes of $885 \times 512 \times 128$. Specifically, [27] and [28] applied a down-sampling step to obtain volumes of dimensions $64 \times 64 \times 128$ and $100 \times 100 \times 128$ voxels, respectively. Contrary, the method proposed in this paper allows taking better advantage of the useful information from each slide by focusing on the retina regions around the ONH, as detailed in Figure 4.

Regarding the predictive model development, to the best of the authors' knowledge, we are the first that propose the use of LSTM networks to address the glaucoma detection just from raw SD-OCT volumes. In addition, we introduce some novelties with respect to the basic LSTM networks. In particular, we propose a sequential-weighting module (SWM) which allows refining the LSTM outputs to control the overfitting and improve the learning of the models via skip-connections. SWM block makes possible that each LSTM output h_i directly contributes to the volume classification, but in a weighted way. To demonstrate the outperforming of the proposed SWM-based approach, we compared it with other basic LSTM structures, as observed in Table 10. Since LSTM networks can output 3D or 2D arrays depending on the specified output shape, we analysed both options (OS3D and OS2D, respectively) and we also included bidirectional layers (Bi), according to the architecture used in [14]. As appreciated in Table 10, the use of bidirectional layers surpasses in both cases the results achieved with the same models without including these layers. For this reason, we based on the best model from the four previous experiments to build deeper LSTM networks via stacking two LSTM memory cells. Finally, the results reached by the proposed methods with and without bidirectional layers can be appreciated in Table 10. In such table, we observe that the *Bi+OS2D* model reports better SP and PPV, but in exchange for compromising the rest of the metrics. However, the proposed model provides a more sensible behaviour outperforming the SN, NPV, FS and AUC results and keeping stable the rest of the metrics. Otherwise, when the traditional VGG16 architecture is used as a feature extractor, the performance of the model notably decreases for all figures of merit in contrast to the proposed method.

In Table 11, corresponding to the analysis of training and validation curves, the superiority of the proposed model is accentuated since it achieves the best results related to the learning stage for most of the metrics. Note that, in line with the results reported during the evaluation of the feature extractors, when the traditional VGG16 architecture is used, the stability (STBL) and the quality (QLTY) of the models are better. However, in this case, the end-to-end system with VGG16 as a feature extractor reports a lot of overfitting (OVFT), which explains the poor performance of the model affecting to the rest of figures of merit. Additionally, the proposed method also stands out for STBL and QLTY metrics, besides the OVFT, being the differences in this case remarkably enough to affect the future behaviour of the model when predicting new SD-OCT volumes. Specially, OVFT metric shows a notorious better performance, which allowed training the model during more epochs with a lower learning rate to achieve a more robust model. This resulted in a higher quality (QLTY) of the training model since greater and more stable learning (higher LRNG and lower STBL values) was accomplished. Moreover, in Table 11, we can also see that the proposed model reaches the best results for validation accuracy and loss measures. From this rigorous analysis, we considered the proposed method as the best volume-based predictive model, which is characterised by using the RAGNet-VGG16 model as a feature extractor and the SWM block to refine the LSTM outputs before the top model.

Finally, we carried out a prediction stage to evaluate the models' performance making use of the test set. Specifically, during the evaluation of the test set, SWM-based models stand out for all metrics, as expected. The end-to-end system using VGG16 as a feature extractor reports a more sensible behaviour since it outperforms for SN and NPV metrics, whereas the proposed method (using RAGNet-VGG16 as a feature extractor) is a more specific model since it highlights for SP and PPV metrics. Also, in line with the findings obtained during the internal cross-validation stage (Tables 10 and 11), the proposed method reaches the best performance for general metrics such as FS, AUC and ACC (see Table 12). It is important to note that, although VGG16+SVM model achieves better SN and NPV values, the rest of metrics are greatly affected, which does not correspond to a reliable model in order to predict new samples. However, the proposed method reports more stable values for all figures of merit, being the SN and NPV differences very small with respect to the VGG16+SWM model (0.0690 and 0.0168, respectively). Therefore, based on the results reported during both training and testing phases, it leads to thinking that the proposed model arises like the best system to provide an added value for glaucoma detection, taking into account that SD-OCT volumes are not being currently traced due to the workload involved. Nevertheless, it would be necessary to validate the volume-based model on external databases to verify the robustness of the proposed system.

From the final results reached by the end-to-end system, we can also conclude that the proposed feature extractor is not camera-specific because, although it was developed using circumpapillary OCT images, the features of the latent space extracted from each cross-sectional slide of the volume allow obtaining a high performance after including LSTM and SWM architectures. This fact evidences that the proposed feature extractor, besides to be non-camera specific, is robust against different types of acquisition cut-offs (circumpapillary or linear) since it founds the relevant information around the optic nerve head of the retina, independently of the extraction mode.

Additionally, the computed Class Activation Maps (CAMs) extracted from the proposed model could help to determine the relevance of the models since, according to Figure 6, notorious differences are evident in the classification of healthy and glaucoma SD-OCT volumes. In this paper, unlike the rest of the state-of-the-art works which reported CAMs from single slides of each volume, we expose different representative slides for several random volumes to determine which B-scans become more relevant for glaucoma diagnosis. In particular, the first slides of the volumes $(I_1, I_2, I_3, ...)$ do not seem to matter much when predicting the healthy class, in contrast to the glaucoma label since the heat maps highlight specific areas around the RNFL. Central slides $(..., I_{21}, I_{22}, I_{23}, ...)$ are more interesting because, in the case of healthy volumes, the LSTM model pays attention to the left and right bounds of the retina areas, whereas central regions corresponding to the optic disc cupping seem more discriminative for the glaucoma class. Additionally, a prominent activation usually appears highlighting the neuroretinal rims for glaucomatous volumes, especially on the left part. Otherwise, more advanced central slides, i.e. $..., I_{41}, I_{42}, I_{43}, ...$, reports a clear discriminatory ability to determine the healthy class, but no obvious signs of glaucoma are evidenced by the proposed model. Concerning the last slides $(..., I_{62}, I_{63}$ and $I_{64})$, the heat maps also manifest differences depending on the class since the proposed model tends to highlight the external areas of the retina for healthy slides, and the central zones for glaucomatous samples. In summary, the findings achieved by the CAMs are directly in line with those reported in the literature [27, 29] since the heat maps focus on the edges of the retinal layers in the normal volumes, whereas retinal structures such as RNFL, neuroretinal rims and lamina cribrosa are evident in the glaucomatous cases.

5.3. Discussion about the SotA comparison

As we have previously mentioned, we are the first that propose the use of CNN-LSTM networks to address the glaucoma detection from SD-OCT volumes. In particular, we expose in this paper a comparison between our method and other works proposed in the literature which addressed the problem via 3D deep-learning architectures. Since no public SD-OCT databases are available, we replicated the experiments performed by the state-of-the-art studies [27, 29] to objectively contrast the differences reported in Table 14. Results show a clear outperforming of the proposed method with respect to the rest of the state-of-the-art models for all figures of merit. In Table 14, we can observe the superiority of the proposed RAGNet+SWM method for OCT glaucoma detection, which specially stands out for SN, FS, AUC and ACC metrics, where the differences between models exceeds more than 10%. At this point, it is important to highlight that the results achieved by the state-of-the-art networks in this study could be underperformed since, with the aim of reporting a direct comparison, we trained their architectures on our database, but they were originally intended to be trained on larger data sets.

From the state-of-the-art comparison carried out, which is very limited by the absence of public databases, it can be concluded that the proposed method, based on a new combination of CNN and LSTM networks, outperforms the glaucoma detection results achieved from other state-of-the-art studies focused on 3D architectures.

6. Conclusion

In this paper, we have proposed an artificial-intelligence predictive model based on a new deep-learning strategy to address the glaucoma detection just from raw SD-OCT volumes. Specifically, the proposed model consists of a novel combination of CNN and LSTM networks that allows taking into account spatial dependencies between the B-scans of the volumes. For the first time, we have combined fine-tuning techniques with other convolutional blocks in parallel to build a slide-level feature extractor from circumpapillary OCT samples. In addition, skip-residual connections were included to improve the discriminator performance through an attention module intended to refine the features of the latent space. Also, in order to keep the spatial information along with the three-dimensional data, we have proposed the use of LSTM networks with an innovative sequential-weighting module (SWM) that allows optimising the LSTM outputs to enable a more stable and efficient model learning. From the developed model, we computed the class activation maps, whose results could suppose a promising tool for an easier 3D scans analysis by the specialists, who could scroll the heat maps that highlight the areas of interest to determine the class of each sample. Additionally, the proposed method outperforms the results reported by other state-of-the-art works, which also focused on the raw SD-OCT volumes to address the glaucoma detection via 3D deep-learning architectures. Based on the obtained results and taking into account that we did not include visual field, intraocular pressure or other external tests to develop the predictive models, we can conclude that SD-OCT volumes could provide great added value for glaucoma diagnosis. Furthermore, artificial intelligence techniques, such as the proposed in this paper, could help ophthalmologists to face the workload associated with the analysis of the cross-sectional OCT images.

As future research lines, better results for SD-OCT volumes could be reported by training a slide-level discriminator focused on the specific knowledge of the SD-OCT B-scans, instead of the circumpapillary images. Although the proposed method was intended to evidence the possible added value that SD-OCT cubes provide for glaucoma detection, it could be considered as a good starting point to build a reliable computer-aided diagnosis system. Additionally, significant improvements could be reached by increasing the number of samples of the database.

Acknowledgement(s)

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used here.

Funding

This work has been funded by GALAHAD project [H2020-ICT-2016-2017, 732613], SICAP project (DPI2016-77869-C2-1-R) and GVA through project PROMETEO/2019/109. The work of Gabriel García has been supported by the State Research Spanish Agency PTA2017-14610-I.

References

[1] R. N. Weinreb, P. T. Khaw, Primary open-angle glaucoma, The Lancet 363 (9422) (2004) 1711–1720.

[2] J. B. Jonas, T. Aung, R. R. Bourne, A. M. Bron, R. Ritch, S. Panda-Jonas, Glaucoma–authors' reply, The Lancet 391 (10122) (2018) 740.

[3] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, C.-Y. Cheng, Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis, Ophthalmology 121 (11) (2014) 2081–2090.

[4] G. A. U. National, Glaucoma: diagnosis and management.

[5] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, et al., Optical coherence tomography, science 254 (5035) (1991) 1178–1181.

[6] F. A. Medeiros, L. M. Zangwill, L. M. Alencar, C. Bowd, P. A. Sample, R. Susanna, R. N. Weinreb, Detection of glaucoma progression with stratus oct retinal nerve fiber layer, optic nerve head, and macular thickness measurements, Investigative ophthalmology & visual science 50 (12) (2009) 5741–5748.

[7] C. Sinthanayothin, J. F. Boyce, T. H. Williamson, H. L. Cook, E. Mensah, S. Lal, D. Usher, Automated detection of diabetic retinopathy on digital fundus images, Diabetic medicine 19 (2) (2002) 105–112.

[8] T. Walter, P. Massin, A. Erginay, R. Ordonez, C. Jeulin, J.-C. Klein, Automatic detection of microaneurysms in color fundus images, Medical image analysis 11 (6) (2007) 555–566.

[9] A. Diaz-Pinto, A. Colomer, V. Naranjo, S. Morales, Y. Xu, A. F. Frangi, Retinal image synthesis and semi-supervised learning for glaucoma assessment, IEEE transactions on medical imaging 38 (9) (2019) 2211–2218.

[10] I. I. Bussel, G. Wollstein, J. S. Schuman, Oct for glaucoma diagnosis, screening and detection of glaucoma progression, British Journal of Ophthalmology 98 (Suppl 2) (2014) ii15–ii19.

[11] P. R. Lichter, Variability of expert observers in evaluating the optic disc., Transactions of the American Ophthalmological Society 74 (1976) 532.

[12] R. Varma, W. C. Steinmann, I. U. Scott, Expert agreement in evaluating the optic disc for glaucoma, Ophthalmology 99 (2) (1992) 215–221.

[13] G. J. Jaffe, J. Caprioli, Optical coherence tomography to detect and manage retinal disease and glaucoma, American journal of ophthalmology 137 (1) (2004) 156–169.

[14] T. Kurmann, P. Márquez-Neila, S. Yu, M. Munk, S. Wolf, R. Sznitman, Fused detection of retinal biomarkers in oct volumes, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 255–263.

[15] D. C. Hood, A. S. Raza, On improving the use of oct imaging for detecting glaucomatous damage, British Journal of Ophthalmology 98 (Suppl 2) (2014) ii1–ii9.

[16] D. Bizios, A. Heijl, J. L. Hougaard, B. Bengtsson, Machine learning classifiers for glaucoma diagnosis based on classification of retinal nerve fibre layer thickness parameters measured by stratus oct, Acta ophthalmologica 88 (1) (2010) 44–52.

[17] S. J. Kim, K. J. Cho, S. Oh, Development of machine learning models for diagnosis of glaucoma, PLoS One 12 (5) (2017) e0177726.

[18] G. García, R. del Amor, A. Colomer, V. Naranjo, Glaucoma detection from raw circumapillary oct images using fully convolutional neural networks, arXiv preprint arXiv:2006.00027.

[19] F. A. Medeiros, A. A. Jammal, A. C. Thompson, From machine to machine: An oct-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs, Ophthalmology 126 (4) (2019) 513–521.

[20] K. A. Thakoor, X. Li, E. Tsamis, P. Sajda, D. C. Hood, Enhancing the accuracy of glaucoma detection from oct probability maps using convolutional neural networks, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 2036–2040.

[21] G. An, K. Omodaka, K. Hashimoto, S. Tsuda, Y. Shiga, N. Takada, et al., Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images, Journal of healthcare engineering 2019.

[22] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, S. Farsiu, Automatic segmentation of nine retinal layer boundaries in oct images of non-exudative amd patients using deep learning and graph search, Biomedical optics express 8 (5) (2017) 2732–2744.

[23] M. Pekala, N. Joshi, T. A. Liu, N. M. Bressler, D. C. DeBuc, P. Burlina, Deep learning based retinal oct segmentation, Computers in Biology and Medicine 114 (2019) 103445.

[24] K. A. Barella, V. P. Costa, V. Gonçalves Vidotti, F. R. Silva, M. Dias, E. S. Gomi, Glaucoma diagnostic accuracy of machine learning classifiers using retinal nerve fiber layer and optic nerve data from sd-oct, Journal of ophthalmology 2013.

[25] V. G. Vidotti, V. P. Costa, F. R. Silva, G. M. Resende, F. Cremasco, M. Dias, E. S. Gomi, Sensitivity and specificity of machine learning classifiers and spectral domain oct for the diagnosis of glaucoma, European journal of ophthalmology 23 (1) (2013) 61–69.

[26] J. Xu, H. Ishikawa, G. Wollstein, R. A. Bilonick, L. S. Folio, Z. Nadler, L. Kagemann, J. S. Schuman, Three-dimensional spectral-domain optical coherence tomography data analysis for glaucoma detection, PloS one 8 (2) (2013) e55476.

[27] S. Maetschke, B. Antony, H. Ishikawa, G. Wollstein, J. Schuman, R. Garnavi, A feature agnostic approach for glaucoma detection in oct volumes, PloS one 14 (7).

[28] E. Noury, S. S. Mannil, R. T. Chang, A. R. Ran, C. Y. Cheung, S. S. Thapa, H. L. Rao, S. Dasari, M. Riyazuddin, S. Nagaraj, et al., Detecting glaucoma using 3d convolutional neural network of raw sd-oct optic nerve scans, arXiv preprint arXiv:1910.06302.

[29] A. R. Ran, C. Y. Cheung, X. Wang, H. Chen, L.-y. Luo, P. P. Chan, M. O. Wong, R. T. Chang, S. S. Mannil, A. L. Young, et al., Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep-learning analysis, The Lancet Digital Health 1 (4) (2019) e172–e182.

[30] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, et al., Clinically applicable deep learning for diagnosis and referral in retinal disease, Nature medicine 24 (9) (2018) 1342.

[31] X. Wang, H. Chen, L. Luo, A.-r. Ran, P. P. Chan, C. C. Tham, C. Y. Cheung, P.-A. Heng, Unifying structure analysis and surrogate-driven function regression for glaucoma oct image screening, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 39–47.

[32] X. Wang, H. Chen, A.-R. Ran, L. Luo, P. P. Chan, C. C. Tham, R. T. Chang, S. S. Mannil, C. Y. Cheung, P.-A. Heng, Towards multi-center glaucoma oct image screening with semi-supervised joint structure and function multi-task learning, Medical Image Analysis 63 (2020) 101695.

[33] A. R. Ran, J. Shi, A. K. Ngai, W.-Y. Chan, P. P. Chan, A. L. Young, H.-W. Yung, C. C. Tham, C. Y. Cheung, Artificial intelligence deep learning algorithm for discriminating ungradable optical coherence tomography three-dimensional volumetric optic disc scans, Neurophotonics 6 (4) (2019) 041110.

[34] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, et al., Clinically applicable deep learning for diagnosis and referral in retinal disease, Nature medicine 24 (9) (2018) 1342–1350.

[35] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[36] J. Jiang, X. Liu, L. Liu, S. Wang, E. Long, H. Yang, F. Yuan, D. Yu, K. Zhang, L. Wang, et al., Predicting the progression of ophthalmic disease based on slit-lamp images using a deep temporal sequence network, PloS one 13 (7).

[37] A. Khan, A. Sohail, U. Zahoora, A. S. Qureshi, A survey of the recent architectures of deep convolutional neural networks, arXiv preprint arXiv:1901.06032.

[38] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: Full training or fine tuning?, IEEE transactions on medical imaging 35 (5) (2016) 1299–1312.

[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[40] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE transactions on pattern analysis and machine intelligence 31 (5) (2008) 855–868.

[41] H. Sak, A. W. Senior, F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling.

[42] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, G. Toderici, Beyond short snippets: Deep networks for video classification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[43] S. C. Wong, A. Gatt, V. Stamatescu, M. D. McDonnell, Understanding data augmentation for classification: when to warp?, in: 2016 international conference on digital image computing: techniques and applications (DICTA), IEEE, 2016, pp. 1–6.

[44] T. Calders, S. Jaroszewicz, Efficient auc optimization for classification, in: European Conference on Principles of Data Mining and Knowledge Discovery, Springer, 2007, pp. 42–53.