Clinically Verified Hybrid Deep Learning System for Retinal Ganglion Cells Aware Grading of Glaucomatous Progression

Hina Raja, Taimur Hassan, Muhammad Usman Akram, Member, IEEE, and Naoufel Werghi, Senior Member, IEEE

Abstract—Objective: Glaucoma is the second leading cause of blindness worldwide. Glaucomatous progression can be easily monitored by analyzing the degeneration of retinal ganglion cells (RGCs). Many researchers have screened glaucoma by measuring cup-to-disc ratios from fundus and optical coherence tomography scans. However, this paper presents a novel strategy that pays attention to the RGC atrophy for screening glaucomatous pathologies and grading their severity. Methods: The proposed framework encompasses a hybrid convolutional neural network framework that extracts the retinal nerve fiber layer, ganglion cell with the inner plexiform layer and ganglion cell complex regions, allowing thus a quantitative screening of glaucomatous subjects. Furthermore, the severity of glaucoma in screened cases is objectively graded by analyzing the thickness of these regions. Results: The proposed framework is rigorously tested on publicly available Armed Forces Institute of Ophthalmology (AFIO) dataset, where it achieved the $F_1$ score of 0.9577 for diagnosing glaucoma, a mean dice coefficient score of 0.8697 for extracting the RGC regions and an accuracy of 0.9117 for grading glaucomatous progression. Furthermore, the performance of the proposed framework is clinically verified with the markings of four expert ophthalmologists, achieving a statistically significant Pearson correlation coefficient score of 0.9236. Conclusion: An automated assessment of RGC degeneration yields better glaucomatous screening and grading as compared to the state-of-the-art solutions. Significance: An RGC-aware system not only screens glaucoma but can also grade its severity and here we present an end-to-end solution that is thoroughly evaluated on a standardized dataset and is clinically validated for analyzing glaucomatous pathologies.

Index Terms—Retinal Ganglion Cells (RGCs), Retinal Nerve Fiber Layer (RNFL), Glaucoma, Deep Learning, Optical Coherence Tomography (OCT).

I. INTRODUCTION

The human retina is composed of various nerve cells. Among these, retinal ganglion cells (RGCs) are responsible for transmitting visual information to the brain. The axons of these ganglion cells collectively form the retinal never fiber layer (RNFL), their cell bodies are enclosed in the ganglion cell layer (GCL) and the inner plexiform layer (IPL) embodies their dendrites. The composition of these layers is commonly termed as ganglion cell complex (GCC). Glaucoma (a progressive optic neuropathy) severely degrades these RGCs, resulting in a thinning of RNFL, ganglion cell with the inner plexiform layer (GC-IPL), and the GCC profiles, as shown in Fig. 1. This RGC atrophy can cause permanent visual impairments and even blindness if left untreated [1]. Clinically, glaucoma can be identified through fundus and optical coherence tomography (OCT) based examinations. OCT imaging is generally preferred by the clinicians over other modalities due to its objective assessments of early and advanced staged glaucoma where early glaucoma refers to the condition when the RGCs start to degenerate due to increased intraocular pressure [2]. However, the progression of RGC dysfunction leads towards the advanced glaucomatous stage where the total cupping of the optic nerve and severe visual impairments can be observed.

II. RELATED WORK

Many researchers have worked on diagnosing glaucoma from retinal OCT images. These studies either emphasize the clinical significance of retinal OCT examination for analyzing glaucomatous severity, or they propose OCT-based autonomous systems for analyzing the glaucomatous pathologies.

A. Clinical Studies

Development in retinal imaging modalities (especially OCT) is making rapid strides in providing the objective visualization of early, mild, and severe ocular complications [4], particularly for the glaucoma [5], the second leading cause of irreversible blindness worldwide. Moreover, the detection and monitoring of glaucoma by measuring the velocity of RNFL thickness loss has been significantly highlighted in many recent state-of-the-art...
studies [6], [7]. Leung et al. [8] demonstrated the importance of RNFL thickness (generated through OCT and visual field tests) in determining the retinal pathological variations within different glaucomatous stages. Ojima et al. [9] signified the importance of RNFL thickness and macular volume, and declared that RNFL thickness has higher diagnostic power than a complete macular volume to detect glaucoma. Furthermore, Medeiros et al. [10] evaluated RGC loss using standard automated perimetry (SAP) and spectral-domain OCT (SD-OCT) examinations. They observed that the early pathological degeneration of RGC results in the drastic thinning of RNFL as compared to the RGC changes in the late glaucomatous stages. Likewise, El-Naby et al. [11] extracted the RNFL thickness from SD-OCT scans and compared it with the VF sensitivity to observe their correlation in screening primary open-angle glaucoma. They concluded that the mean RNFL thickness obtained through SD-OCT imagery is a very good indicator of screening glaucomatous subjects and also for monitoring the progression and severity of the disease.

B. Automated Glaucomatous Analysis

Initial methods developed for glaucomatous screening analyze cup-to-disc ratios from macula-centered and disc-centered fundus images [12]–[16]. However, observing the degeneration of RGCs through optic nerve head (ONH) SD-OCT scans can provide a superior and objective indication of early glaucoma, resulting in the timely prevention of non-recoverable visual impairments. Furthermore, due to the unprecedented clinical significance of retinal OCT examination [5]–[7], many researchers have developed autonomous systems to objectively screen glaucoma (especially in early-stage) using retinal OCT scans [17]. Moreover, Almobarak et al. [18] manually segmented ONH structures from SD-OCT scans to analyze pathological variations in healthy and glaucomatous pathologies. Kromer et al. [19] extracted eight retinal boundaries from 40 SD-OCT scans of healthy subjects using curve regularization. In [20], a generated model was presented to segment retinal layers from OCT images using a group-wise curve alignment. Niu et al. [21] proposed an automated method to segment the six retinal layers using correlation smoothness constraint and dual gradients. Apart from this, several methods have been proposed to quantify retinal layer thickness from SD-OCT scans depicting normal [22], [23], and abnormal retinal pathologies [24]–[28]. Ometto et al. [29] presented ReLayer, an automated framework to segment and estimate the thickness of ILM, inner/outer segment, and RPE from OCT scans to monitor retinal abnormalities. Gao et al. [30] extracted retinal layers through graph decomposition from Topcon SD-OCT scans and evaluated the mean macular thickness of RNFL regions. Afterward, they compared their obtained results with the thickness measurements from Topcon’s built-in layer extraction framework. Likewise, Mayer et al. [31] proposed an automated framework for extracting the retinal layers and computing the RNFL thickness by minimizing the energy obtained through scan gradients, local and regional smoothing filters. They validated their framework on a dataset containing both normal and glaucomatous affected OCT scans, and achieved a mean RNFL thickness of 94.1 ± 11.7 μm and 65.3 ± 15.7 μm, respectively for normal and glaucomatous pathologies. In addition to this, many researchers have proposed computer-aided diagnostic systems to diagnose glaucomatous pathologies from fundus [13]–[15], [32], OCT [17] and fused fundus and OCT imagery [12]. More recently, deep learning has been applied to analyze the glaucomatous pathologies through segmentation-free retinal layers extraction framework [33]. Zang et al. [34] used a convolutional neural network (CNN) and graph search to delineate the retinal boundaries and the optic disc region from ONH SD-OCT scans of normal and glaucomatous subjects, achieving the overall dice coefficient of 0.91 ± 0.04. Maetschke et al. [35] highlighted the significance of RNFL, GC-IPL profiles for diagnosing, and monitoring glaucoma progression and used the 3D CNN model to classify healthy and glaucomatous ONH SD-OCT scans. They outperformed conventional machine learning approaches by achieving the area under the curve (AUC) score of 0.94. Furthermore, Devalla et al. [36] proposed a dilated-residual U-Net architecture (DRUNET) for the extraction of six ONH tissues from SD-OCT scans to aid experts in analyzing glaucomatous progression. DRUNET achieved the overall dice coefficient of 0.91 ± 0.05 when accessed against manual tissue segmentation done by the expert observer. In addition to this, a joint retinal segmentation and classification pipeline was proposed in [37] to analyze healthy and glaucomatous pathologies from 1,004 locally acquired circular OCT scans, and also a severe diabetic macular edema (DME) pathology from selective 110 macular OCT scans of Duke dataset [28]. Pathological degeneration of RGCs observed in RNFL, GC-IPL, and GCC thickness profiles can objectively monitor the progression of glaucomatous severity. However, manual extraction of these regions is a subjective and time-consuming task. Several automated layers extraction methods have been proposed in the literature to address this shortcoming [23], [30], [31], [36], [37]. But, to the best of our knowledge, there is no clinically validated framework that utilizes the degraded RGC profiles to diagnose and grade glaucomatous progression using ONH SD-OCT scans. Moreover, as the ONH SD-OCT scans are considered to be more significant for detecting glaucoma progression [18], validating an automated framework on a publicly available standardized ONH SD-OCT dataset adds significant value to the body of knowledge.
In this paper, we present a fully automated diagnosis and grading of glaucoma from ONH SD-OCT images by analyzing pathological variations of RNFL, GC-IPL, and GCC regions. The proposed framework is unique as it employs a hybrid convolutional network for the RGC-aware diagnosis and grading of glaucoma, and it has been clinically validated with four expert clinicians. The main features of this paper are:

- A novel strategy for the classification and grading of glaucomatous progression by analyzing RNFL, GC-IPL, and GCC regions from ONH SD-OCT scans.
- A significantly improved and lightweight hybrid retinal analysis and grading network (RAG-Net\textsuperscript{v2}) for the simultaneous pixel-level segmentation of retinal regions and scan-level classification of glaucomatous pathologies.
- Rigorous clinical validation of the proposed framework with four expert ophthalmologists to screen, track, and grade glaucomatous progression from high-quality ONH SD-OCT scans. The complete dataset and the annotations from the expert observers are publicly available at https://data.mendeley.com/datasets/2rnnz5nz74/2.

The rest of the paper is organized as follows. Section III describes the proposed method. Section IV showcases the experimental setup. Section V presents the results, followed by detailed discussion and concluding remarks about the proposed framework in Section VI.

### III. PROPOSED METHOD

We present a novel framework that gives an RGC-aware diagnosis of glaucoma using ONH SD-OCT scans. Furthermore, it measures the severity of glaucomatous progression by analyzing the RNFL, GC-IPL, and GCC thickness profiles. The block diagram of the proposed framework is shown in Fig. 2. First of all, the input scan is preprocessed to retain the retinal area. Afterward, the preprocessed scan (containing only the retina and the ONH) is passed to the hybrid convolutional network that extracts the RNFL, GC-IPL, and GCC regions, and screens the scan against glaucoma. The thickness profiles of these extracted regions are computed and their mean values are passed as a feature vector to the supervised support vector machines (SVM) for grading the screened glaucomatous scan as either early suspect or a severe case. The detailed description of each block is presented below:

### A. Preprocessing

The purpose of the preprocessing is to remove the background artifacts and noisy content to obtain accurate extraction of RNFL, GC-IPL, and GCC regions. The preprocessing is performed through structure tensor [38], which highlights the predominant orientations of the image gradients within the specified neighborhood of a pixel. For each pixel of the input
We propose a hybrid convolutional network (HCN) for extracting the retinal regions and also for the classification of candidate scan as normal or glaucomatous. Using an HCN rather than a conventional classification model in this study aims to get an RGC-aware diagnosis of glaucoma. The HCN model proposed here is an improved version of the Retinal Analysis and Grading Network (RAG-Net) [39]. The RAG-Net and its improved version will be described next.

1) Retinal Analysis and Grading Network: RAG-Net is a hybrid convolutional network specifically designed to extract retinal lesions and abnormalities from the macular OCT scans and give lesion-aware grading of retinal diseases by performing simultaneous pixel-level segmentation and scan-level classification. Architecturally, it contains 112 convolution layers, 111 batch normalization layers, 102 ReLUs, 6 pooling layers, 5 lambda layers, 2 softmax, and a fully connected layer [39]. Furthermore, it contains around 62.3M parameters from which around 0.1M are non-learnable. RAG-Net showed the capacity to generalize across multiple scanner specifications for retinal lesion extraction and lesion-aware grading of retinopathy and has also been applied to the multi-modal imagery for retinal lesions extraction, achieving superior performance among its competitors [40]. However, the original RAG-Net has been found limited in discriminating similar texture objects, like retinal boundaries, when their transitional variations are small. This is because RAG-Net possesses kernels with smaller receptive fields (field of view) that do not retain accurately the contextual information of small and similar textural regions. Although, the feature pyramid module within the RAG-Net...
architecture tends to overcome this limitation to some extent. But overall the performance is still capping as will be shown in the results section (Section V). Moreover, the source code of RAG-Net and its complete documentation is available at http://biomisa.org/index.php/downloads/ [39].

2) Modified Retinal Analysis and Grading Network: We propose a RAG-Net modified version, dubbed RAG-Netv2, whereby the contextual-aware unit is built upon atrous convolutions (also known as dilated convolution). The atrous convolutions are formulated in a residual fashion which greatly enhances the kernels receptive field to perform more broad and context-aware filtering while maintaining the same spatial resolution [41]. The 2D atrous convolution is expressed as:

\[
g(x, y) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} k(i, j) f(x - r \ast i, y - r \ast j) \quad (2)
\]

where \( f \) denotes input function (typically a feature map from the previous layer), \( k \) represents the \( N_1 \times N_2 \) kernel, \( r \) is the dilation rate and \( g \) denotes the convolution output (a feature map produced in the current layer). It should be noticed in the above equation that we have introduced a common dilation rate \( r \) in both image dimensions to ensure a consistent receptive field at both of them. When \( r = 1 \), then atrous convolution is simply a linear convolution as shown in Fig. 4 (A), and when \( r > 1 \), the receptive field is enlarged so the kernel captures wider contextual area from the input function to produce more distinctive feature maps. However, increasing the \( r \) also introduces gridding artifacts [41] due to large gaps between convolving pixels in the input function, causing a cascading effect in the consecutive convolution layers which may result in a significant decrease of network performance. The gridding artifacts are also illustrated in Fig. 5 (top row) for the stacked convolution layers. In order to compensate this, we perform atrous convolution with variable dilation factors [42] in the RAG-Netv2 architecture. For a block of \( n \) consecutive convolution layers within the network having the dilation rate \( r^v \) where \( n > r \), the dilation factors in the proposed framework are generated through \( \text{round}(r - \frac{n}{2} + i) \) where \( i \) varies from 0 to \( n - 1 \). For example, for a block containing 5 cascaded convolution layers having the dilation rate \( r = 3 \), the dilation factors will be \([1, 2, 3, 4, 5]\), meaning that the first convolution layer within the block will perform standard convolution (as \( r = 1 \)), the second layer will have \( r = 2 \) and so on as shown in Fig. 5 (bottom row). Similarly, for \( n = 3, r = 3 \), the dilation factors will be \([2, 3, 4]\). The second major benefit of RAG-Netv2 is that it is extremely lightweight and contains 91.04% fewer parameters than original RAG-Net architecture (having 62,352,188 parameters in total) while achieving the better segmentation and classification performance. The detailed architectural description and hyper-parameters of RAG-Netv2 are reported in Table I from which we can see that it contains 5,573,161 learnable and 8,960 non-learnable parameters. Moreover, rather than training RAG-Netv2 from scratch, it is fine-tuned on RAG-Net weights (which are already adjusted in [39] for lesion-aware retinal image analysis) to achieve faster convergence.

### Table I

| Layers            | Number of Layers | Parameters |
|-------------------|------------------|------------|
| Convolution       | 16               | 4,847,369  |
| Pooling           | Average, 10 Max  | 0          |
| Batch Normalization| 15               | 17,920     |
| Activation        | 13 ReLU, 2 Softmax| 0          |
| Lambda            | 5                | 0          |
| Input             | 2                | 0          |
| Zero-Padding      | 10               | 0          |
| Concatenation     | 1                | 0          |
| Reshape           | 1                | 0          |
| Fully Connected and Flatten | 2 | 716,810 |
| Classification    | 1                | 22         |
| Learnable Parameters | 5,573,161     |            |
| Non-learnable Parameters | 8,960     |            |
| Total Parameters  | 5,582,121        |            |

C. Estimation of Retinal Profiles

The severity of glaucoma can be effectively graded by analyzing the RGC atrophy. RGCs primarily exists within the GCC region that consists of RNFL, the ganglion cellular layer (GCL),
and the inner plexiform layer (IPL) [35]. To objectively evaluate the glaucoma progression, the proposed system computes the RNFL, GC-IPL, and the GCC thickness profiles. The RNFL thickness is computed by taking the absolute difference between the ILM and the GCL. Moreover, GC-IPL thickness is computed by taking the absolute difference between GCL and IPL. Also, the GCC thickness is computed by taking the absolute difference between ILM and IPL. Afterward, the mean RNFL, GC-IPL, and GCC thickness values are computed from the extracted thickness profiles which are then passed to the supervised SVM model for grading the glaucomatous progression. The reason for choosing these thickness values as features for the SVM is because they reflect the pathological degeneration of RGCs which can be used for grading glaucoma (predicted through the RAG-Netv2 classification unit) as early or advanced (more severe).

D. Modified Dice-Entropy Loss Function

In order to effectively extract the retinal regions and use them for the RGC-aware classification of glaucomatous scans, RAG-Netv2 jointly optimized the dice-entropy loss which is a linear combination of both dice and cross-entropy loss as expressed below:

\[
L_{de} = \alpha_1 L_d + \alpha_2 L_e
\]

(3)

\[
L_d = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{2 \sum_{j=1}^{C} t_{i,j} p_{i,j}}{\sum_{j=1}^{C} t_{i,j}^2 + \sum_{j=1}^{C} p_{i,j}^2}\right)
\]

(4)

\[
L_e = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} t_{i,j} \log(p_{i,j})
\]

(5)

where \(L_d\) denotes the dice loss, \(L_e\) represents the multi-category cross-entropy loss, \(t_{i,j}\) represents the true labels for \(i\)th sample belonging to \(j\)th class, \(p_{i,j}\) denotes the predicted probability for the \(i\)th sample belonging to \(j\)th class, \(\alpha_1, \alpha_2\) represent the loss functions weights, \(N\) represents the total number of samples in a batch, and \(C\) represents the total number of classes. The dice-entropy loss penalizes RAG-Netv2 to accurately segment the RNFL and GC-IPL pixels from the other retinal pixels and at the same time enables RAG-Netv2 to robustly classify the ONH SD-OCT scans as healthy or glaucomatous.

IV. EXPERIMENTAL SETUP

This section presents a detailed description of the dataset and the training protocol. It also presents the evaluation metrics which have been used to validate the performance of the proposed framework.

A. AFIO Dataset

Armed Forces Institute of Ophthalmology (AFIO) dataset, firstly introduced in [3], is a publicly available repository containing high-resolution ONH SD-OCT scans of healthy and glaucomatous subjects. The dataset has been acquired from AFIO Hospital, Rawalpindi, Pakistan. To the best of our knowledge, it is the only dataset that contains OD centered fundus and ONH centered SD-OCT scans for each subject along with the detailed cup-to-disc markings and annotations from four expert ophthalmologists. The scans within the AFIO dataset are acquired using Topcon 3D OCT-1000 sampled over four years. Furthermore, all the scans within AFIO datasets have been thoroughly graded by four expert ophthalmologists in a blind-manner (i.e. each grader does not know the grading done by his/her colleagues). Also, all four clinicians were very senior (having 20 to 25 years of professional experience in clinical ophthalmology). Moreover, the detailed specifications of the AFIO dataset are presented in Table II.

B. Training Details

RAG-Netv2 in the proposed framework is implemented using Keras APIs on Anaconda Python 3.7.4 platform\(^1\). The training is conducted for 40 epochs where each epoch lasted for 512 iterations on a machine with Intel Core i7-9750H@2.6 GHz processor and 32 GB RAM with a single NVIDIA RTX 2080 Max-Q GPU having cuDNN v7.5 and a CUDA Toolkit 10.1.243. The optimization during the training is performed through ADADelta [43] having a default learning rate of one with the decay factor of 0.95. Moreover, 70% of the dataset is used for the training and the rest of 30% were used for testing as per the dataset standard [3]. To compensate for the low number of training scans within the AFIO dataset, we fine-tuned the weights of the original RAG-Net architecture (obtained after training on more than 0.1 million macular OCT scans [39]) and also performed augmentation of the training scans. The data augmentation we performed is as follows: first, all the scans were horizontally flipped, and then they were rotated between -5 to 5 degrees. Then, we added a zero-mean white Gaussian noise with 0.01 variance. The augmentation procedure resulted in a total of 6,028 training scans to fulfill the training requirements.

C. Evaluation Metrics

The proposed framework has been evaluated using a number of metrics described below:

1) Confusion Matrix: The performance of the proposed framework for accurately classifying and grading glaucomatous subject is measured through the confusion matrix and its associated metrics such as accuracy (\(A_C = \frac{T_P}{T_P + T_N + T_F + T_V}\)), recall (\(T_{PR} = \frac{T_P}{T_P + F_P}\)), specificity (\(T_{NR} = \frac{T_N}{T_N + F_P}\)), falspositive rate (\(F_{PR} = \frac{F_P}{F_P + T_N}\)), precision (\(F_{PV} = \frac{T_P}{F_P + T_P}\)), and \(F_1\) score (\(F_1 = \frac{2P_V T_P}{P_V + T_P}\)), where \(T_P\) denotes true positives, \(T_N\) denotes the true negatives, \(F_P\) denotes the false positives, and \(F_V\) denotes the false

\(^1\)The source code is available at https://github.com/taimurhassan/rag-net-v2.
\(F_N\) denotes the false negatives. To measure classification performance, \(T_P\), \(F_P\), \(T_N\) and \(F_N\) are calculated scan-wise.

2) Receiver Operating Characteristics (ROC) Curve: ROC curve indicates the capacity of the proposed framework to correctly classify and grade healthy and glaucomatous pathologies at various classification thresholds. Moreover, the performance through ROC curves is quantitatively measured through AUC scores.

3) Dice Coefficient (DC): The dice coefficient \((D_C)\) measures how well the proposed framework segments the RNFL, GC-IPL, and GCC regions as compared to their ground truths, and it is computed through: \(D_C = \frac{2TP}{2TP + FN + FP}\). Here \(TP\), \(FP\), and \(FN\) are calculated pixel-wise where \(TP\) indicates the correct extraction of positives (RNFL, GC-IPL, and GCC regions), \(FP\) indicates the misclassified background pixels, and \(FN\) denotes those positive pixels which have been missed by the proposed framework. Afterward, the mean dice coefficient \((\mu_Dc)\) is computed by taking an average of \(DC\) scores scan-wise across the whole dataset.

4) Mask Precision: To further validate the performance of the proposed framework for extracting RNFL, GC-IPL, and GCC regions, we used the mask precision \((m_p)\) metric. Unlike the dice coefficient, \(m_p\) measures both the capacity of the proposed framework in accurately recognizing the RNFL, GC-IPL, and GCC regions as well as extracting their corresponding masks. First of all, the dice coefficient \(D_C\) of the extracted regions is computed in each image using their ground truths. If \(D_C \geq 0.5\), then the \(m_p\) for each region is computed pixel-wise through \(m_p = \frac{TP}{TP + FP}\). However, if the dice coefficient is below 0.5, then the whole region is considered as \(FP\), resulting in a \(m_p\) score of 0. Moreover, the mean mask precision \(\mu_{mp}\) is computed by taking the average of \(m_p\) for each region as \(\mu_{mp} = \sum_{k=0}^{c} m_p(k)\), where \(c\) denotes the number of classes (regions).

5) Clinical Validation: Apart from using performance metrics, we clinically validated the glaucomatous screening and grading performance of the proposed framework with four expert ophthalmologists using the standardized Pearson correlation coefficient \((r_c)\) and its statistical significance measured through the \(p\)-value.

V. RESULTS

The proposed framework has been thoroughly evaluated on the AFIO dataset for the RGC-aware diagnosis and grading of glaucoma. First, we present an ablative analysis to evaluate different segmentation models for the extraction of RNFL, GC-IPL, and GCC regions. Afterward, we present a detailed comparison of the proposed framework with state-of-the-art solutions for extracting the RGC regions as well as screening glaucomatous subjects. Apart from this, we also present the clinical validation of our RAG-Net\textsubscript{v2} driven grading system with four expert ophthalmologists.

A. Ablation Study

The ablative aspect of this research involves choosing the segmentation framework that can accurately extract the retinal regions such as the RNFL and the GC-IPL regions to compute the GCC profiles and grade glaucomatous subjects accordingly. For this purpose, we compared the performance of the proposed RAG-Net\textsubscript{v2} segmentation unit with the popular state-of-the-art PSPNet [44], SegNet [45], U-Net [46], FCN-(8, 32) [47], as well as our original RAG-Net architecture [39]. The extraction performance is shown in Table IV, where we can observe that RAG-Net\textsubscript{v2} achieved the overall best \(\mu_{DC}\) score of 0.8697 leading the second-best FCN-8 [47] by 2.78%. Moreover, the performance of PSPNet [44] and FCN-8 [47] are extremely comparable as the FCN-8 [47] is leading PSPNet [44] by 0.035% only. Similarly, FCN-8 [47] is leading RAG-Net [39] by 7.12%. If we look at the performance of each model for extracting individual regions, we can see that the best score for extracting RNFL is achieved by FCN-8 [47], though it’s leading PSPNet [44] by 0.011% and RAG-Net\textsubscript{v2} by 0.651%. For extracting GC-IPL, the best performance is achieved by RAG-Net\textsubscript{v2}, leading the second-best FCN-8 [47] by 6.28%. The best performance for extracting GCC regions is also achieved by the RAG-Net\textsubscript{v2} i.e. \(\mu_{DC}\) score of 0.8698. In terms of \(\mu_{mp}\), the overall best performance is also achieved by RAG-Net\textsubscript{v2} as shown in Table V, leading the second-best FCN-8 by 1.78%. It leads the second-best FCN-8 by 3.19% and 3.38%, respectively for extracting GC-IPL and GCC regions. However, for the RNFL extraction, it lags from both FCN-8 and PSPNet by 1.05% and 0.76%. But overall, RAG-Net\textsubscript{v2} outperforms FCN-8 and PSPNet with a larger margin in extracting RNFL, GC-IPL, and GCC regions. Fig. 6 showcases some qualitative comparison of all the hybrid and conventional segmentation models. Here, scan (A), (J), and (S) depict glaucomatous pathologies whereas scan (AB) depicts a healthy pathology. Both RAG-Net\textsubscript{v2} and FCN-8 produced good performance (comparing it with the ground truths) in extracting the RNFL and GC-IPL regions. But RAG-Net\textsubscript{v2} has an upper hand in extracting GC-IPL boundaries from both healthy and glaucomatous scans (highlighted in green color). Furthermore, the original RAG-Net [39] was found limited in accurately extract RGC regions especially in scans (D), (M), (V), and (AE). Such limitation is because RAG-Net [39] cannot well differentiate similar textural patterns (often depicted in the retinal layers and boundaries). Moreover, because RAG-Net\textsubscript{v2} achieves the overall best performance for extracting RGC regions (as evident from Table IV, Table V and Fig. 6), besides having the capacity to classify and grade glaucomatous pathologies, we chose it in the proposed framework for further analysis.

B. Extraction of RNFL, GC-IPL and GCC Regions

To the best of our knowledge, all the existing methods (except [35]) which have been proposed for the extraction of RNFL, GC-IPL and GCC regions have been validated on either local in-house datasets or on publicly available datasets which only contains macular OCT scans. Moreover, the framework proposed in [35] is designed for screening normal and glaucomatous pathologies without paying attention to RGC atrophy. The comparison of RAG-Net\textsubscript{v2} with state-of-the-art literature [36], [37] for the extraction of RNFL, GC-IPL and GCC regions is indirect as the experimental protocols and the datasets were different.
Fig. 6. Performance comparison of deep segmentation models for the extraction of RNFL (shown in red color) and GC-IPL regions (shown in green color). Left to right, column-wise: Original scans, ground truths, the performance of RAG-Net_v2, RAG-Net, PSPNet, SegNet, UNet, FCN-8, and FCN-32.

### TABLE III
COMPARISON OF THE PROPOSED FRAMEWORK WITH DRUNET [36] AND BI-DECISION [37]. THE ABBREVIATIONS ARE DS: DATASET SIZE, DPA: DATASET PUBLICLY AVAILABLE, EP: EXPERIMENTATION PROTOCOL, GS: GLAUCOMATOUS SCREENING, GG: GLAUCOMATOUS GRADING, TR: TRAINING, TE: TESTING, CV: CROSS-VALIDATION

|                  | Proposed | DRUNET [36] | Bi-decision [37] |
|------------------|----------|-------------|------------------|
| DS               | 196* (ONH) | 107* (ONH)  | 1.114*           |
| DPA              | Yes      | No          | Yes              |
| EP               | TR: 137, TE: 59 | TR: 40, TE: 60 | 3-Fold CV         |
| GS               | Yes      | No          | Yes              |
| GG               | Yes      | No          | No               |
| µDC              | 0.86     | 0.91        | 0.72             |

* 1,004 are circular OCT scans from the local dataset, and 110 selected macular scans are taken from Duke dataset [28] representing DME pathologies. # This count represents the dataset size excluding the augmented scans.

### TABLE IV
COMPARISON OF SEGMENTATION MODELS IN TERMS OF µDC FOR EXTRACTING THE RNFL, GC-IPL, AND GCC REGIONS

| Framework       | RNFL   | GC-IPL  | GCC    | Mean  |
|-----------------|--------|---------|--------|-------|
| RAG-Net_v2      | 0.8692 | 0.8703  | 0.8698 | 0.8697|
| RAG-Net [39]    | 0.8192 | 0.7508  | 0.7860 | 0.7853|
| PSPNet [44]     | 0.8748 | 0.8151  | 0.8457 | 0.8452|
| SegNet [45]     | 0.8111 | 0.6945  | 0.7555 | 0.7537|
| UNet [46]       | 0.8216 | 0.8253  | 0.8234 | 0.8234|
| FCN-32 [47]     | 0.8638 | 0.7470  | 0.8083 | 0.8064|
| FCN-8 [47]      | 0.8749 | 0.8156  | 0.8460 | 0.8455|

### TABLE V
COMPARISON OF SEGMENTATION MODELS IN TERMS OF µmp FOR EXTRACTING THE RNFL, GC-IPL, AND GCC REGIONS

| Framework       | RNFL   | GC-IPL  | GCC    | Mean  |
|-----------------|--------|---------|--------|-------|
| RAG-Net_v2      | 0.8410 | 0.8108  | 0.7915 | 0.8144|
| RAG-Net [39]    | 0.7661 | 0.6701  | 0.6016 | 0.6792|
| PSPNet [44]     | 0.8475 | 0.7685  | 0.7229 | 0.7796|
| SegNet [45]     | 0.7402 | 0.6978  | 0.6630 | 0.7003|
| UNet [46]       | 0.8082 | 0.7791  | 0.7291 | 0.7723|
| FCN-32 [47]     | 0.8500 | 0.7849  | 0.7647 | 0.7999|
| FCN-8 [47]      | 0.7623 | 0.7080  | 0.6607 | 0.7190|

(see Table III). Nonetheless, this indirect comparison highlights the capacity of RAG-Net_v2 for extracting the retinal regions while simultaneously diagnosing and grading the glaucomatous pathologies as compared to competitive methods. As reported in Table III, our proposed framework lags from DRUNET [36] by 5.49% in terms of µDC. However, their dataset is not public and its size is almost the half of AFIO dataset. Furthermore, DRUNET is only a conventional segmentation model that does not possess glaucomatous screening and grading capabilities as compared to RAG-Net_v2. The joint segmentation and classification pipeline, termed bi-decision [37] achieved the mean dice coefficient of 0.72 for extracting the retinal regions. However, for glaucoma screening, bi-decision [37] is only validated on the local dataset (although the authors verified bi-decision on the selective 110 DME affected scans as well from the publicly available Duke dataset [28]). Here, we want to highlight that we have already validated the original RAG-Net [39] on 43,613 macular OCT scans from five publicly available datasets (including the Duke dataset [28]), and here our emphasis is on extending our hybrid framework for the RGC-aware diagnosis and grading of glaucoma using high-quality ONH SD-OCT scans from the publicly available dataset, and for reproducing the clinical trials.

### C. Classification of Glaucomatous Scans

Since the thinness of the RNFL, GC-IPL and GCC profiles highlight the degeneration of RGCs, which directly reflects glaucomatous progression. We have utilized the encoder end of the RAG-Net_v2 to perform RGC-aware classification of healthy and classification pathologies. After training the RAG-Net_v2 for extracting the RGC regions, the trained weights are used for screening healthy and glaucomatous pathologies through the RAG-Net_v2 classification unit. The performance of RAGNet_v2 classification unit for screening glaucoma is measured through standard metric such as $A_C$, $TPR$, $TN_H$, $PPV$, $AUC$ and $F_1$ score. As reported in Table VI, the proposed framework achieves 0.958% better results as compared with [17] and [12], and 12.5%
TABLE VI

| Metric | Proposed | [17] | [12] | [32] | [35] | [37] |
|--------|----------|------|------|------|------|------|
| AP     | 0.9491   | 0.9400 | 0.9420 | 0.8300 | -    | 0.8140 |
| TPR    | 0.9714   | 0.9545 | 0.9285 | 0.8846 | -    | -    |
| TNR    | 0.9166   | 0.9285 | 0.9545 | 0.7708 | -    | -    |
| FPR    | 0.0834   | 0.0714 | 0.0455 | 0.2292 | -    | -    |
| PPV    | 0.9444   | 0.9629 | 0.9629 | 0.8604 | -    | -    |
| F1     | 0.9577   | 0.9453 | 0.9453 | 0.8723 | -    | -    |
| AUC    | 0.9871   | -     | -     | -     | 0.9400 | 0.8640 |

Fig. 7. ROC curve highlighting the performance of RAG-Net_v2 for classifying and grading glaucoma.

better results as compared to [32] in terms of accuracy. The comparison with [32] is indirect as the authors only used fundus imagery for the classification of glaucomatous pathologies. However, the comparison with [17] and [12] is fair and direct as both of these frameworks were tested on the AFIO dataset using the same experimental protocols. We can further observe the capacity of the proposed framework in screening glaucomatous pathologies through the TPR ratings in Table VI, where the RAG-Net_v2 achieves 1.73% better performance than [17], and also leading [12] by 4.41%. This performance gain is achieved because RAG-Net_v2 pays attention to the pathological variations of RGCs related to the progression of glaucoma. The classification performance of RAG-Net_v2 is also evaluated through ROC curve as shown in Fig. 7, and compared with [35] in terms of AUC as shown in Table VI, where we can see that RAG-Net_v2 leads [35] by 4.77%. However, this comparison is also indirect as both frameworks are tested on different datasets.

D. Profile Extraction

The novel aspect of the proposed framework is that it can grade glaucomatous progression by analyzing pathological variations of RGCs through RNFL, GC-IPL, and GCC regions represented within the ONH SD-OCT scans. Table VII reports the mean RNFL, GC-IPL, and GCC region thickness for the early and advanced glaucomatous pathologies. We can observe here that the RAG-Net_v2 achieved the mean RNFL thickness of 93.50 μm for early glaucomatous suspects and 69.46 μm for the advanced glaucomatous stage. These RNFL thickness ranges were obtained from scans of publicly available AFIO datasets and confirmed by expert clinicians. They deviate from the ranges defined by [11] by 2.67% and 3.25%, respectively, for the early and advanced stage glaucoma. We can also observe that GC-IPL and GCC profiles provide a unique distinction between glaucomatous severity, and can contribute positively towards grading it. In Fig. 8, we report some qualitative examples highlighting the segmentation performance of RAG-Net_v2. The scans in this figure (except the original ones in the first column) are intentionally converted to grayscale so that the extracted regions can be visualized. Scans (A), (E), and (I) represent normal, early glaucomatous suspect and advanced-stage glaucoma, respectively from which RAG-Net_v2 has accurately extracted the RNFL, GC-IPL, and GCC profiles. For example, we can see RNFL thinning in the scan (F) as compared to (B), and how precisely it is picked by RAG-Net_v2. Also, the yellow color in the scans of Fig. 8 indicates complete overlap with the ground truths whereas other colors indicate incorrect extractions. Please zoom-in to best see the results.

Fig. 8. Examples of normal, early and advanced stage glaucomatous scans from which the RNFL, GC-IPL, and GCC regions are extracted. The scans are intentionally made grayscale to highlight the extraction results as compared to the ground truths. The yellow color indicates complete overlap with the ground truths while other colors indicate incorrect extractions. Please zoom-in to best see the results.

TABLE VII

| Mean Thickness | Early       | Advanced   |
|----------------|-------------|------------|
| RNFL           | 93.50±9.84  | 69.46±13.77|
| GC-IPL         | 62.23±5.67  | 33.96±7.53 |
| GCC            | 155.73±13.10| 103.42±10.27|
| RNFL [11]     | 91.00±7.28  | 67.20±7.06 |

In Fig. 9 we report plots of the class distribution for the early and advanced glaucomatous classes in the feature space defined...
by RNFL, GC-IPL, and GCC profiles. We can observe that the thinning of these profiles can clearly distinguish glaucoma progression. We notice that the training samples in Fig. 9 are highly clustered compared to the testing samples. This is due to the fact that these training samples are generated through data augmentation and, thus, highly correlated with the actual samples. Overall, these profiles can show clear discrimination and great potential to be used with an SVM classifier for grading glaucoma progression.

E. RGC Aware Grading of Glaucoma

Using the RNFL, GC-IPL, and GCC profiles as distinctive features, the proposed system provides RGC-aware grading of glaucoma using the SVM classification model. Both the classification and grading performance of the proposed framework is shown through confusion matrices in Fig. 10 (A) and Fig. 10 (B), respectively. Moreover, Fig. 10 (C) reports the grading performance using only the mean RNFL thickness threshold (see Table VII). The classification between normal and glaucomatous scans is performed through RAG-Net, whereas the RGC-aware grading of glaucoma as early suspect or advanced stage (reported in Fig. 10-B) is performed through SVM based on RNFL, GC-IPL, and GCC profiles. In Fig. 10 (B), we can see that the proposed framework achieved an accuracy of 0.9117 for screening early and advanced stage glaucoma cases which is 16.123% superior compared to grading approach based on analyzing only the RNFL thickness. Also, out of 34 test samples in Fig. 10 (B), three are falsely graded i.e. two early cases are graded as severe and one severe case is graded as having early glaucoma symptoms. But, these three misclassifications have less impact on the overall performance of the proposed system because both of them have been correctly classified as having glaucoma. In Fig. 10 (A), we notice that one misclassification (out of 35) of a glaucomatous scan predicted as normal by our system. This is one challenging boundary case showing very early glaucomatous symptoms i.e. having the cup-to-disc ratio of 0.325. All the four ophthalmologists have considered this as an early suspect (please see the recommendation of the ophthalmologists in the patient record sheet within the dataset for more detail. The misclassified case has a name ‘149155_20150914_095855_Color_R_001.jpg’).

F. Clinical Trials

We have also performed a series of clinical trials in which we cross-validated the predictions of the proposed framework with the recommendations from the expert clinicians. Table VIII reports samples of the clinical trials along with the recommendations of the four expert ophthalmologists (publicly available in the dataset package [3]). Furthermore, we report in Table IX the Pearson correlation analysis along with its statistical significance showcasing the clinical validation of the proposed framework with each clinician. \( r_c \) ranges between -1 to +1 where -1 indicates the strong negative association between the two entities, +1 shows the strong positive association and 0 depicts that both entities are not related. Furthermore, \( p \)-value < 0.05 indicates that the obtained \( r_c \) score is statistically significant. From Table IX, we can observe that although the recommendations contradict with each other (as they are marked by each clinician based on his/her own experience), the proposed framework achieves the highest correlation with Clinician-4 (i.e. \( r_c = 0.9236 \)), having \( p \)-value of 4.40 \( \times 10^{-58} \). Moreover, the minimum \( r_c \) score is achieved with the grading of Clinician-3 (i.e. \( r_c = 0.6863 \)), but it is still quite significant. In 8th row of Table VIII, we have an exception where all clinicians have marked the scan as having early glaucomatous symptoms, and the proposed framework grades it as depicting advanced stage pathology. This indicates the proposed framework is tuned in a way to prioritize advanced stage subjects that need immediate attention and treatment to prevent vision loss and blindness. We also report in Table VIII the fundus scan with each case to help the readers (and other clinicians) in cross-verifying the predictions made by the proposed framework by correlating the cup-to-disc ratios. Through these fundus scans, we can also notice that the ONH SD-OCT scans graded as having early or advanced stage glaucoma typically contains a cup-to-disc ratio of 0.6 or above which is normally considered to imply glaucoma [48].
VI. DISCUSSION AND CONCLUSION

In this work, we proposed a fully automated system for the classification and grading of glaucoma from ONH SD-OCT scans. Unlike other frameworks that rely on cup-to-disc ratios for screening glaucoma, the proposed system analyzes the pathological changes related to the degeneration of RGCs through RNFL, GC-IPL, and GCC thickness profiles. We propose a hybrid convolutional network (RAG-Net_v2) for the extraction of these profiles, coupled with an SVM classifier for the classification and the grading of healthy and glaucomatous pathologies. The experiments evidenced the superiority of our framework in screening early and advanced glaucoma cases as compared to the state-of-the-art solutions relying on the cup-to-disc ratios as evidenced by the $F_1$ score of 0.9577. The preeminence of our system emanated from the newly proposed architecture variants in RAG-Net_v2, integrating contextual-aware modules, built on residual atrous convolutions, along with the feature pyramid block. This proposed variants boosted the capacity of the RAG-Net_v2 for discriminating the retinal regions as reflected by the $\mu_{DC}$ score of 0.8697, outperforming popular deep segmentation models. Apart from this, RAG-Net_v2 significantly reduces the total number of trainable and non-trainable parameters by 91.04% as compared to the original RAG-Net architecture. This improvement also relates to the addition of contextual-aware modules replacing the standard convolutional blocks with the atrous convolutions, making the RAG-Net_v2 a lightweight architecture for screening and monitoring the glaucoma progression. The introduction of contextual-aware modules in RAG-Net_v2 addresses the limitations of the original RAG-Net architecture which, while outperforms state-of-the-art frameworks for lesions extraction, showed restraints in differentiating similar textural patterns in retinal layers and boundaries as shown Fig. 6. However, the implications of introducing contextual-aware modules need to be thoroughly tested for the lesions extraction from the macular pathologies, if we want to extend the modified RAG-Net architecture to the analysis of lesion-aware maculopathy. This investigation will be part of our next future work.

ACKNOWLEDGMENT

The authors would like to thank to the four expert ophthalmologists for providing the clinical validations of the retinal scans within the AFIO dataset.

REFERENCES

[1] R. N. Weinreb, T. Aung, and F. A. Medeiros, “The pathophysiology and treatment of glaucoma: A review,” JAMA, vol. 311, no. 18, pp. 1901–1911, 2014.
[2] C. F. Burgoyne et al., “The optic nerve head as a biomechanical structure: A new paradigm for understanding the role of IOP-related stress and strain in the pathophysiology of glaucomatous optic nerve head damage,” Prog. Retinal Eye Res., vol. 24, pp. 39–73, 2005.
[3] H. Raja et al., “Data on OCT and fundus images for the detection of glaucoma,” Data Brief, vol. 29, pp. 1–4, Apr. 2020.
[4] T. Hassan et al., “Review of OCT and fundus images for detection of Macular Edema,” in Proc. IEEE 12th Int. Conf. Imag. Syst. Techn., 2015, pp. 1–4.
et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.

et al.