Eurographics Workshop on Visual Computing for Biology and Medicine (2020)
B. Kozlíková, M. Krone, and N. N. Smit (Editors)

GLANCE: Visual Analytics for Monitoring Glaucoma Progression

Astrid van den Brandt[1], Mark Christopher[4], Linda M. Zangwill[4], Jasmin Rezapour[2,4], Christopher Bowd [4], Sally L. Baxter[3,4], Derek S. Welsbie[4], Andrew Camp[4], Sasan Moghimi[4], Jiun L. Do[4], Robert N. Weinreb[4], Chris C.P. Snijders [1] and Michel A. Westenberg[1]

[1]Eindhoven University of Technology, The Netherlands
[2]University Medical Center Mainz, Department of Ophthalmology, Mainz, Germany
[3]Division of Biomedical Informatics, Department of Medicine, University of California San Diego, La Jolla, CA
[4]Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology and Shiley Eye Institute University of California San Diego, La Jolla, CA

Abstract

Deep learning is increasingly used in the field of glaucoma research. Although deep learning models can achieve high accuracy, issues with trust, interpretability, and practical utility form barriers to adoption in clinical practice. In this study, we explore whether and how visualizations of deep learning-based measurements can be used for glaucoma management in the clinic. Through iterative design sessions with ophthalmologists, vision researchers, and manufacturers of optical coherence tomography (OCT) instruments, we distilled four main tasks, and designed a visualization tool that incorporates a visual field (VF) prediction model to provide clinical decision support in managing glaucoma progression. The tasks are: (1) assess reliability of a prediction, (2) understand why the model made a prediction, (3) alert to features that are relevant, and (4) guide future scheduling of VFs. Our approach is novel in that it considers utility of the system in a clinical context where time is limited. With use cases and a pilot user study, we demonstrate that our approach can aid clinicians in clinical management decisions and obtain appropriate trust in the system. Taken together, our work shows how visual explanations of automated methods can augment clinicians' knowledge and calibrate their trust in DL-based measurements during clinical decision making.

1. Introduction

Glaucoma is one of the leading causes of irreversible blindness worldwide and its prevalence will likely continue to rise due to global aging populations [TLW*14, WAM14]. Glaucoma is a progressive eye disease that is characterized by loss of nerve fibers, resulting in visual field defects [AGS10]. Treatment can slow or even stop progression of the disease, which makes early detection crucial [SYC14]. However, timely detection of disease progression is challenging because glaucoma often remains asymptomatic until there is considerable visual field loss [WAM14].

The current standard-of-care for monitoring glaucoma involves both structural and functional measurements. Structural changes in the eye can be assessed by optical coherence tomography (OCT) and clinical examinations of the optic disc. Visual function is analyzed using a visual field (VF) test, which includes global summary statistics such as mean deviation (MD) and visual field index (VFI) [ZDF*17]. The VF test is essential to detecting and monitoring the disease, because it provides measurement of peripheral and central visual function of the patient. However, there are limitations associated with VF testing; VFs are subjective and variable [WDZ*13], and some people experience difficulties with taking the test. In contrast to VF, structural measurements performed by OCT are objective and have good reproducibility [PGI*12]. Moreover, glaucomatous structural damage found on OCT measurements often precedes VF defects [KZZ*15]. This, and the fact that they are less variable makes structural measurements powerful for detecting and monitoring glaucoma progression [ZDF*17].

Because the causes of glaucoma are multi-factorial and there exist various measurements (fundus photographs, visual fields, OCT of optic nerve and macula), deep learning (DL) methods have been introduced to analyze the disease [CBB*18, CBB*20, LHK*18]. Although many DL approaches show excellent performance for a variety of tasks, understanding these models and implementing them into clinical practice remains a challenge. Deep learning models are often regarded as black boxes because it is hard to grasp the rationale behind the nonlinear operations used for predictions. This is a problem for deployment in real-world applications, especially in the clinical domain where trust and interpretability are crucial [MWW*17]. Insufficient interpretability and trust are main barriers to adoption of DL models in clinical practice. Moreover, many experts warn that DL models may not be able to incorporate "outside" or "contextual" factors that are important for decision-making. Further, algorithmically determined features may not always be clinically familiar. In cases where there is strong disconnection between the two, clinicians may lose trust in the system, especially when no explanations are given [CRH*19].

Issues with trust and interpretability also affect other application domains, which has made explainable artificial intelligence (XAI) an important area of research. In XAI, visualization techniques are developed to enhance the collaboration between human and AI.

DOI: 10.2312/vcbm.20201175

State-of-the-art work in DL visualization focuses mainly on model development in research settings [PHVG*17, SCD*17, KAKC17, GCWGvW19]. However, for deep learning to be successfully integrated in clinical practices, we should also consider busy clinical settings where visualizations could improve the efficiency of healthcare providers. This is particularly relevant for ophthalmology, a high-volume specialty with complex clinical workflows that demand efficient clinical decision-making [BGC*20].

This paper explores how visualization of DL-based measurements can be used to facilitate disease management in the clinic. We hypothesize that clinicians can provide a recommendation using predicted VF MD *(H1)* and that visual explanations are helpful in clinical decision making *(H2)*. Furthermore, we hypothesize that visual explanations can help to assess the reliability of predictions *(H3)* and can influence patient-specific treatment decisions *(H4)*.

Through iterative design sessions with ophthalmologists, vision researchers, and manufacturers of OCT instruments, we designed a clinically-oriented visual analytics tool that incorporates a VF prediction model [CBB*19b] to provide clinical decision support in assessing and predicting disease progression. This model predicts function from structure using spectral domain (SD) OCT to help unravel the relationship between structural and functional damage. In summary, our work contributes the following:

- Development of a visual analytics approach to implement automated systems for glaucoma management in the clinic. While the approach is tailored to ophthalmology clinics, the described methods may also be generalizable to other clinical situations where experts rely on image analysis for decision making.
- Identification of key tasks that a visualization tool should support to provide assistance in clinical decision making with deep learning models.
- Demonstration of usability and effectiveness of the approach by evaluation with use cases and a pilot user study.

2. Background and Related Work

We consider this work at the intersection of Clinical Decision Support Systems, Deep Learning, Interpretability and Visual Analytics. As this section introduces several medical acronyms, we refer the reader to Table 3 in the supplementary material for a glossary.

2.1. Medical Background

Glaucoma is a progressive eye disease characterized by damage to the optic nerve, *i.e.*, thinning of the retinal nerve fiber layer (RNFL) and retinal ganglion cell-inner plexiform layer (GCIPL), and accompanying disease-related patterns of VF defects. Traditionally, clinicians based their diagnosis on clinical examination of the optic nerve head and VF tests. VF examinations can detect dysfunction in peripheral and central vision by measuring light sensitivity. Recently, SD-OCT has become increasingly important for diagnosis and monitoring of glaucomatous damage. The current practice typically involves fundus examination, VF testing, intraocular pressure (IOP) measurements and an optic disc centered OCT scan (see Fig. 1) [HDM18].

There are several drawbacks and challenges with the current practice patterns. First, visual fields are inherently variable, which poses challenges to progression monitoring because disease-related change can be masked by or erroneously attributed to variability of measurements. For 24-2 VF tests, Wall *et al.* [WDZ*13] reported average standard deviations of MD of 1.01 dB in glaucoma eyes. In another study, standard deviations of 24-2 VF MD of 1.0 to 1.5 dB were reported in early to moderate glaucoma eyes [YLDC16]. However, a recent study showed that long-term variability of VF MD is much higher than short term variability [UMJ*19] and that variability increases with the severity of disease. Further, early damage in the macula is hard to detect with a 24-2 VF which focuses on peripheral vision [Hoo17]. A 10-2 test that focuses on central vision can detect these defects, but because of time and financial constraints, 24-2 testing is often favored. Another drawback of VF testing is that it is taxing for some patients in terms of time and concentration. Furthermore, learning effects can limit the estimation of progression [ZDF*17]. In addition to these challenges, there are some shortcomings associated with OCT use. For instance, clinicians do not make optimal use of OCT technology [Hoo17] because in many clinics, only OCT scans of the optic disc are obtained routinely, even though it is now known that early damage can occur in the macula. Moreover, many clinicians fail to examine the RNFL circle scans to identify local damage because the software does not facilitate it or they were not trained to do so [Hoo17]. Above all, clinicians have limited time with patients, making employing all available features of OCT and VF impractical.

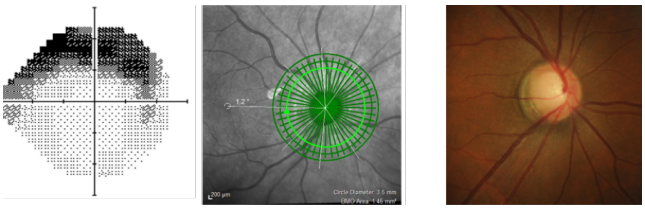


Figure 1: *Examples of images that are obtained during a clinical visit. From left to right: a printout of a patient's visual field measured in a 24-2 Humphrey Visual Field Analyzer (HFA) test, an optical coherence tomography (OCT) circle scan of the retina, an optic nerve head centered fundus image.*

These drawbacks in the current clinical practice call for approaches to improve the usefulness of OCT-based assessments and to facilitate more efficient VF testing. For example, Hood [Hoo17] proposes a one-page OCT report that encourages visual evaluation of circle scans as well as macular scans and probability plots. Based on this report, the clinician can decide if, when, and how (peripheral versus central testing) a VF test should be performed. Inspired by this work, we combine deep learning and visualization to aid clinicians in a similar fashion.

2.2. Clinical Decision Support Systems

Clinical Decision Support (CDS) systems provide clinicians with person- and situation-specific information to improve healthcare decisions and outcomes [OTM*07]. Broadly speaking, CDS comes in three varieties: (1) systems that provide extra information relevant to the current clinical situation, (2) alerts, reminders and recommendations for direct action and (3) systems that organize

and present information in a way that enhances decision making [MMG14]. We focus on the latter. Despite successes of CDS at reducing errors and improving outcomes, issues with user acceptance and trust have prevented their adoption in healthcare [KMCAS18, CRH*19]. For glaucoma, some CDS systems for screening and early detection of glaucoma have been developed [ADD*11, ANE*15]. In this paper, we propose a system that provides CDS for disease progression monitoring and demonstrate that some of the issues related to acceptance and trust can be overcome.

2.3. Deep Learning in Glaucoma Research

In the past years, DL methods have been used in different application domains, including healthcare. DL models, especially convolutional neural networks (CNN), have permeated tasks in medical image analysis such as image classification, segmentation, and object recognition. Specifically, in ophthalmology and dermatology, CNNs have made significant impacts [LKB*17]. For example, the first autonomous DL application approved by the United States Food and Drug Administration (FDA) is a CNN that reviews fundus images to detect referable diabetic retinopathy [Rat18]. Compared to other retinal diseases, DL applications for glaucoma have been limited. Prior work has primarily focused predicting and identifying disease using fundus photos [CBB*18, LHK*18]. More recent studies report models using OCT for detection, segmentation, and diagnosis [DFLRP*18, CBB*20].

In this work, we make use of two DL strategies in our visualization approach. One DL model predicts peripheral 24-2 visual function using RNFL circle scans (*i.e.*, scans that measure RNFL thickness surrounding the optic nerve head (ONH)), the most common method to assess structural damage in glaucoma [CBB*19b, CPB*20]. The other estimates central 10-2 visual function using SD-OCT imaging of the macula, which has recently been shown to be effected early in glaucoma [CBB*19a].

2.4. Trust and Interpretability

In recent years black box decision systems have appeared in various application domains. These systems typically use DL models [GMR*19], but the complicated nature of these models limits how much we can understand or interpret them, creating a lack of trust [RSG16]. This poses a problem for healthcare, because safety and reliability must be guaranteed. In this work, we focus on interpretability of explanations and trust in predictions.

Trust. Whether a system will be used is greatly dependent on the user's trust. Ribeiro *et al.* [RSG16] define trust in two ways: (1) a user can trust individual predictions to take some action based on them or (2) a user can trust the complete model to behave reasonably when deployed. Further, these authors explain that trusting predictions is especially important when a system is used for decision making. Because our study is focused on a decision support system, we will continue with the first definition of trust.

Interpretability. Interpretability of a model enables users to understand the system. It is, therefore, an important prerequisite for trust [RSG16]. In the context of AI, interpretability is defined as the ability to explain or to provide meaning in human-understandable terms. Interpretability may be further specified according to different user needs (*e.g.,* trust or causality) and properties of models that make them interpretable (*e.g.,* transparency or post-hoc interpretability). Because we focus on trusting predictions, we consider *post-hoc interpretability*, which is the ability to explain predictions without elucidating precisely how a model works [Lip16].

2.5. Visual Analytics for Deep Learning

Over the past years, visualization approaches have been developed for understanding DL models, especially for CNNs. These approaches serve different purposes of interpretation. A popular class of tools that provide post-hoc interpretations are saliency methods [ZF14, ZKL*16, ZCAW17], which highlight aspects in the input that were relevant to a given prediction. These methods produce heat maps (saliency maps or activation maps) that can be used for determining qualitatively what a model has learned. In addition to saliency methods, there are approaches that visualize learned features in each neuron to support more low-level algorithmic interpretation of the workings of a model [ZF14, YCN*15].

Most DL visualization tools in the literature assist expert users who have a background in machine learning and help them develop better performing models [HKPC18]. Some examples are ActiVis [KAKC17], DeepEyes [PHVG*17] and LSTMVis [SGPR17]. Some recent approaches explicitly address the needs of end users of AI-powered systems [CRH*19, GCWGvW19]. However, both of these tools are purposed for asynchronous medical examinations or analysis tasks where immediate diagnosis or decision-making requiring real-time communication with patients are not required.

Our work focuses on real-life scenarios where clinical decision-making should be performed immediately. Clinicians (glaucoma specialists, general ophthalmologists and optometrists) should be able to assess whether a prediction from an imperfect DL model can be trusted and thus used for monitoring progression in the clinic.

3. Problem Definition

Similar to prior work [KAKC17, GCWGvW19], we define the problem of diagnosing disease progression using automated systems by illustrating tasks that our visualization approach should support. First, we describe the models and data used for development of our approach. Next, we define four tasks (labeled **T1** - **T4**) identified by clinicians and industry experts that serve as design goals for our system.

3.1. Data and Model Descriptions

We use three existing DL models (see Fig. 2) that predict quantitative VF measurements (here: VF MD) using Spectralis (Heidelberg Engineering, Heidelberg) OCT scans [CBB*19a, CBB*19b, CPB*20]. Although these models use different inputs, they all can be obtained using two OCT scanning protocols.

The first model is a regression model that uses unsegmented B-scans from RNFL circle scans at 3.4-mm-diameter location to predict the MD measured on a 24-2 VF. B-scans are 2-D images

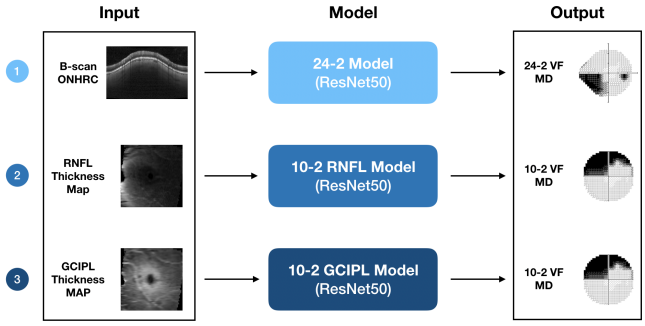


Figure 2: *Overview of the three models used in our approach, showing the input images (b-scan, RNFL and GCIPL thickness maps) and their targets (24-2 and 10-2 MD). Here the 24-2 and 10-2 VF scan patterns are added as reference.*

created by many 1-D scans (A-scans) performed at several locations. For analyzing thickness of the RNFL in clinical practice, B-scans from 3.4-mm-diameter circles centered on the optic nerve are most commonly used. We call this model the 24-2 RNFL model, although it should be noted that the B-scan did not include segmentation of the RNFL explicitly.

The second and third model predict 10-2 VF MD from RNFL and GCIPL thickness maps that were extracted from macula-centered SD-OCT scans. The values of these thickness maps denote the distance between inner limiting membrane (ILM) and the RNFL for RNFL thickness maps, and between the RNFL and inner nuclear layer (INL) for GCIPL thickness maps. The order of these layers is depicted in Fig. 3. We call these models the 10-2 RNFL model and 10-2 GCIPL model, respectively.

All 3 DL models were able to explain a large proportion of the variance (0.5 to 0.8) in predicting VF MD from OCT. The 3 models had mean absolute errors (MAE) ranging from 1.6 decibels (dB) to 2.1 dB for the 24-2 RNFL model and from 1.29 dB to 2.23 dB for the other two models. On an individual eye level, predictions were more accurate. For example, in the 24-2 RNFL model, 43% percent of eyes in the test set had a MAE less than 1.5 dB, which is comparable to the moderate variability of MD values reported in prior studies where standard deviation of MD from 24-2 VFs varies between 1.0 dB and 1.5 dB in eyes with disease severity similar to those included in the current study [YLDC16]. This is where our tool will play an important role in helping clinicians determine the model accuracy for the individual patient.

The data used to develop these models included 1081 OCT images from 665 healthy participants and 828 images from 529 glaucoma participants. Glaucoma was defined based on VF measurements. The data were adopted from two longitudinal studies designed to evaluate function and structure in glaucoma: The African Descent and Glaucoma Evaluation Study (ADAGES clinicaltrials.govidentifier: NCT00221923) and the University of California, San Diego (UCSD) based Diagnostic Innovations in Glaucoma Study (DIGS, clinicaltrials.govidentifier:NCT00221897). Table 1 summarizes some characteristics of these datasets.

Table 1: *Characteristics of the participants with glaucomatous visual field damage (GVFD+) and without (GVFD-).*

Parameter	GVFD-	GVFD+	P-value
Number of participants	665	529	-
Number of eyes	1,081	828	-
Number SDOCT-VF Pairs	4,261	5,504	-
VF Mean Deviation (dB)	-0.04 ± 1.6	-5.2 ± 6.5	<0.001
Age (years)	54.8 ± 20.6	58.0 ± 26.1	0.02

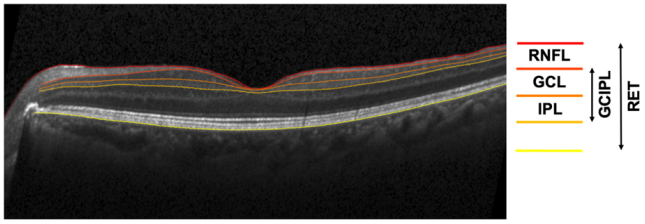


Figure 3: *A b-scan showing a cross sectional image of the retinal layers: a segmentation of the retinal nerve fiber layer (RNFL) is indicated in red, orange segmentations indicate the thickness of the ganglion cell layer (GCL) and inner plexiform layer (IPL), which together form the ganglion cell-inner plexiform layer (GCIPL).*

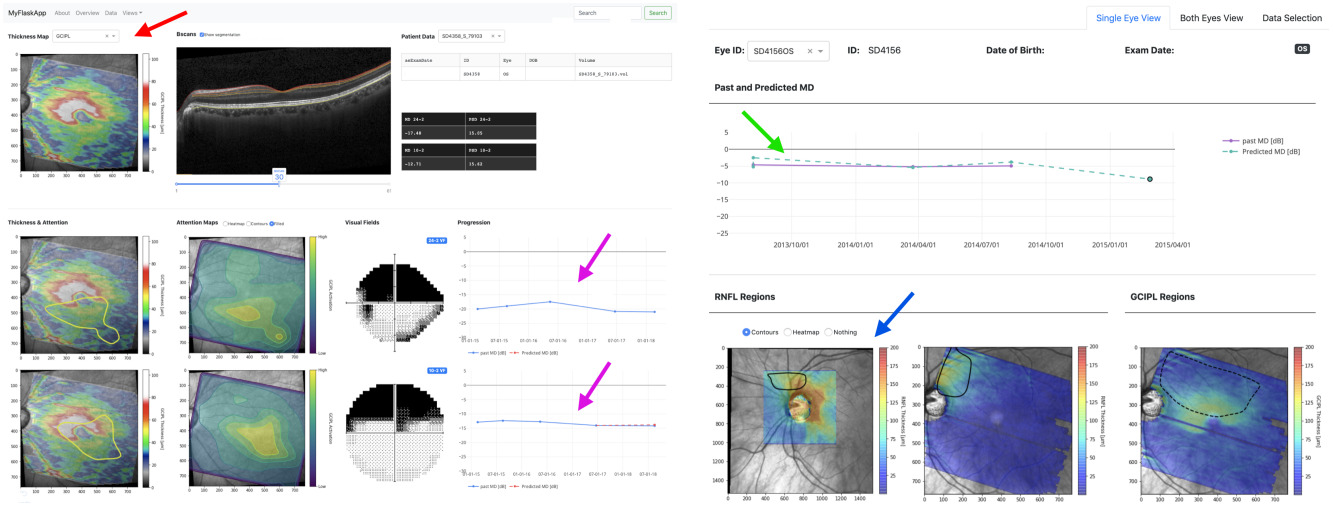
Besides these data, saliency maps were extracted from the models to highlight discriminative regions in the input scans that were important for prediction. These maps were computed using a gradient-based approach to produce class activation maps or attention maps (Grad-CAM) [SCD*17]. In the original implementation, a rectified linear unit (ReLU) activation is applied to retain only features that have a positive influence on a class of interest. Because we are working with regression models, we used a variation of Grad-CAM that computes absolute values of the gradients. This variation highlights pixels that drive the prediction to "more glaucoma-like" and pixels that drive the prediction to "less glaucoma-like".

3.2. Tasks

Through semi-structured interviews, shadowing sessions in the clinic and participatory design sessions with three ophthalmologists and two industry experts, we identified needs, challenges and suggestions that are relevant for the analysis of glaucoma progression using DL models. We summarize them into four tasks (**T1** - **T4**). In Section 4, we show how our tool supports these tasks.

T1 Assess reliability of a prediction. Before incorporating information from automated predictions into clinical decisions, the clinician wants to know how reliable this prediction is. Comparisons with previous measurements of the same eye, which are available in ophthalmology clinics, might provide some insight into reliability. Clinicians should be supported in exploring these predictions to get an appropriate level of trust in the model in a way that can be easily integrated into their busy clinical routine.

T2 Understand why the model made a prediction. Without understanding the reasons for a prediction, clinicians cannot use it for diagnosis. The system should explain to its users why the model



(a) First prototype **(b)** Second prototype

Figure 4: *Two earlier prototypes of GLANCE:* ***(a)*** *the first prototype that had options for multiple layers (red arrow) and showed predictions (pink arrows),* ***(b)*** *the second prototype that showed predicted versus actual values (green arrow) and ONH-Cube scans (blue arrow).*

made a prediction and show which parts of the input contributed to a certain prediction outcome, without going into detail about the model methods or mathematical operations. This can aid clinicians to further assess reliability but also to gain insight in the relation between structural measurements and functional loss.

T3 Alert the clinician to features that are relevant. Once a prediction is determined to be sufficiently reliable, explanations might be provided to discover new features for progression tracking. This would be helpful in cases where changes are hard to notice with traditional measurements, for example in tracking early progression.

T4 Guide future scheduling of VF testing. In current clinical practice, a 24-2 VF test is most commonly used, regardless of the severity of disease. However, in cases of severe damage or paracentral loss, a 10-2 VF test may be more appropriate. The 24-2 VF testing pattern assesses only 4 of 54 points in the macular region, thus loss in this area might be missed due to sparse sampling. However, tracking loss in this area is important because it accounts for 30% of the total retinal ganglion cells and represents over 60% of the visual cortex [SWT07]. Due to billing and time reasons, clinicians often cannot request both testing modalities. Moreover, the frequency of VF testing is often standardized and may not be optimal for some patients. Another benefit of a DL-based CDS system would be to assist clinicians in determining the testing protocol and frequency of VF testing (*e.g.*, is current testing imperative or can it wait) to achieve a personalized approach for each patient.

4. GLANCE

Here we describe the design process and key components of GLANCE, a visualization tool to help clinicians make DL-based glaucoma progression management decisions efficiently (*i.e.* at a "glance").

4.1. User-Centered Design Process

The current design of GLANCE resulted from 6 months of iterations involving domain and industry experts. Our domain situation is clinicians analyzing glaucoma progression using DL models for decision support. To understand their needs, both regarding traditional and DL-enabled workflows, one-hour semi-structured interviews with 3 clinicians and two-hour shadowing sessions (direct observations) with 2 clinicians were completed. Supplemental Table 1 provides an overview of the interview questions. From these interviews and shadowing sessions, we obtained important insights and initial design requirements: (1) the design should be simple and intuitive due to clinicians' time constraints, and (2) clinicians preferred progression information to be on the same page as the scan.

These insights were used to formulate abstract tasks (see Section 3). After some iterations, we decided to keep the data close to its original form to reduce cognitive load and ensure real-world generalizability (Fig. 4b blue arrow). For example, early prototypes had options for viewing activation and thickness maps of segmented layers besides RNFL and GCIPL (Fig. 4a, red arrow). Further, multiple predictions were shown on the same page, which was deemed confusing (Fig. 4a, pink arrows). Clinicians' feedback informed our decision to revise the tool to be narrower in scope. Later in the process, however, we added another encoding for presenting true versus predicted values (Fig. 4b, green arrow) based on clinicians' feedback.

4.2. Final Design

We now introduce the key components of GLANCE (see Fig. 5 for an overview). Similar to existing CDS systems for OCT-based assessment in glaucoma, GLANCE enables clinicians to select patients, inspect demographics, and view OCT scans and derived data (*e.g.*, thickness maps). What is new is the MD prediction that informs the clinician about the anticipated VF loss and corresponding visual explanations for this prediction. These visual explana-

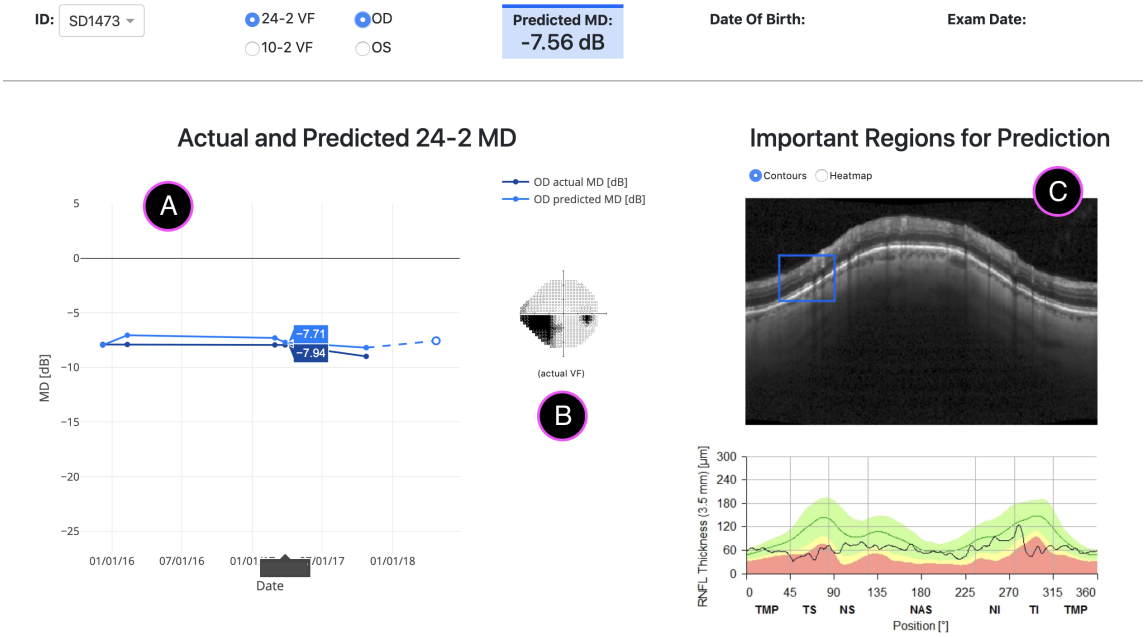


Figure 5: *The user interface of GLANCE. Here the **24-2 view** is shown, which reports the predicted MD for a 24-2 VF based on circle scans. Capital letters depict explanation types and supporting encodings of our approach: line plot **(A)**, hover **(B)**, and important regions **(C)**.*

tions are the core components of our approach and are applied to all views.

Explanation type 1. This explanation consists of a line plot where all previous measured (true) and predicted MD values are displayed chronologically (see Fig. 5.A, for example). Colors encoding predicted versus true values were chosen in consultation with the clinicians. A dashed line indicates the transition to the current predicted value. In addition, by hovering over the traces, the clinician can inspect previous VF results to evaluate progression (Fig. 5.B). The goal is to help clinicians assess the reliability of the most recent predicted value (**T1**). Further, this information may lead to a different testing frequency (**T4**). When the predictions approximate true values, the clinician might decide that the current predicted value can be trusted and used for diagnosing progression. Alternatively, when the model has been unstable in the past, or when it systemically seems to be incorrect, the clinician might not trust the predicted MD and request the patient do an actual VF test.

Explanation type 2. The second explanation consists of a bounding box overlaid on the actual scan to highlight spatial features salient for the current predicted MD (see Fig. 5.C, for example). Activation maps were binarized with a threshold of 20% of the maximum intensity (empirically determined), resulting in contours of pixels around which a bounding box is drawn. Bounding boxes were chosen instead of contours to overcome fragmentation in small activation regions (causing clutter and confusion). The underlying OCT scan data is displayed in a familiar manner to the clinician. For example, the red-green color scheme indicates deviation (Fig. 5.C, bottom) and the rainbow color map depicts thickness (Fig. 6.C), similar to existing reports. Radio buttons above the scan may be used to display the original activation map (viridis heat map), instead of the bounding box (contours). The viridis color scheme is used for activation maps in order to avoid confusion with thickness maps. For 24-2 VF data, a deviation map is added under the scan that denotes areas of abnormal RNFL thinning compared to age-matched controls (indicated by red and yellow, Fig. 5.C). The combined figure helps explain why the current prediction was made (**T2**). Highlighted areas either correspond to known clinical features or indicate new landmarks that make the prediction more or less glaucoma-like (**T3**). Further, if the highlighted parts are located in areas with noise or artifacts (*e.g.*, missing local image information) visible in the underlying scan or where no abnormalities are expected, the clinician might decide that the prediction is not reliable (**T1**) and may subsequently adjust the testing frequency (**T4**). Lastly, when both the highlighted areas on the OCT scan and the past VFs show central loss, and only a 24-2 test was obtained, the clinician may order a 10-2 VF test (**T4**).

Based on the clinicians' feedback, two separate views were designed for 10-2 VF MD and 24-2 VF MD prediction. The explanations are the same for both views, but the type of visual field (10-2 vs 24-2) and scan data (circle scan or thickness map) differ. In practice, a clinician often only has results from one test (usually the 24-2 test). Having both 24-2 and 10-2 predictions on one page would create confusion and visual clutter. Similarly, left eye (OS) and right eye (OD) are separated within these views.

24-2 view. Here the clinician can inspect a patient's VF progression by MD measured on a 24-2 spaced grid (Fig. 5). To clarify: a more negative MD dB value corresponds to more glaucomatous damage. The model's behavior is shown for previous 24-2 MD predictions (*i.e.*, the first explanation type; Fig. 5.A). An ONH-centered circle scan at 3.4 mm is displayed for inspecting RNFL thinning. Further, salient features are visualized over this RNFL

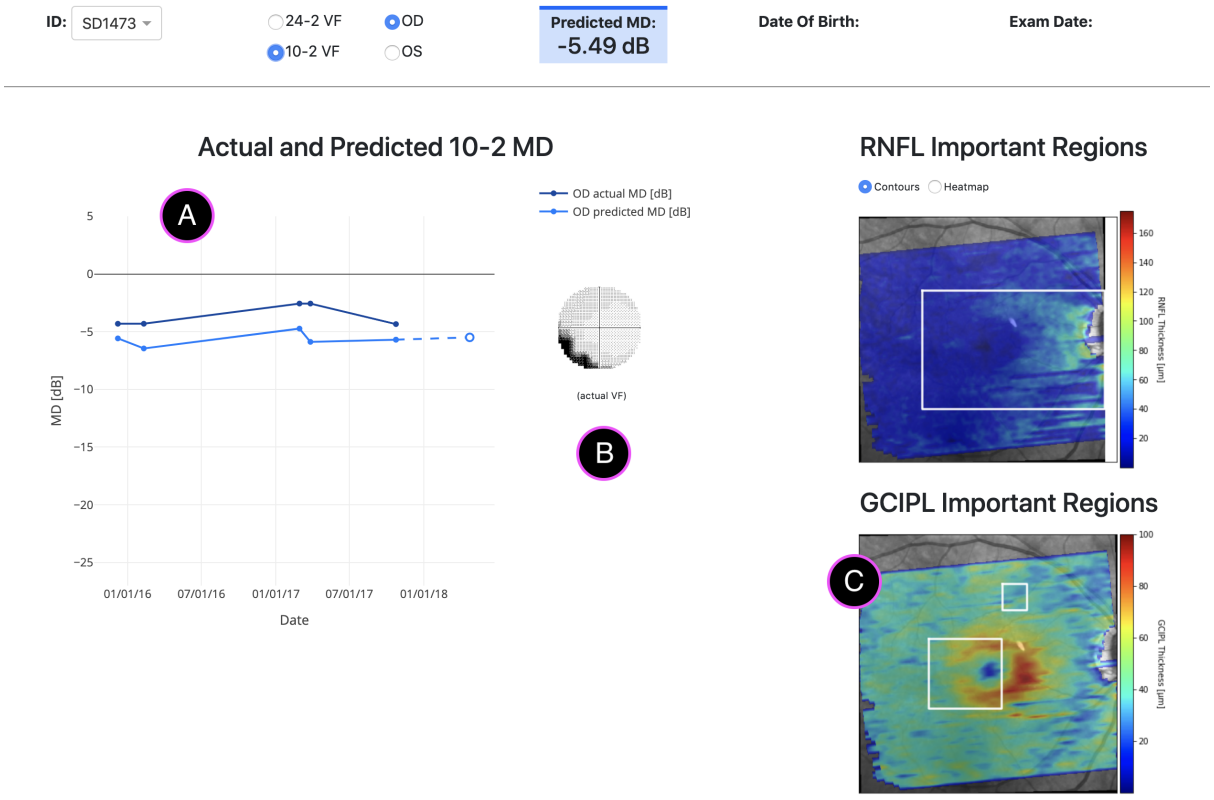


Figure 6: *The **10-2 view** of GLANCE. This view shows predicted MD of a 10-2 spaced visual field based on GCIPL and RNFL thickness maps of the macula. Capital letters depict explanation types and supporting encodings: line plot **(A)**, hover **(B)** and important regions **(C)**.*

circle scan (second explanation type; Fig. 5.C). Below it, a deviation plot aligned with the scan enables the clinician to compare the location of the salient features with areas of abnormal thinning for this eye. Furthermore, the salient areas can be compared to the previous 24-2 loss patterns that are visible when hovering over the past MD trace in the line plot (Fig. 5.B). When the salient regions in the circle scan align with the abnormal area's deviation plot and also with patterns in the previous VFs, a clinician may feel more confident about the prediction, especially if these areas are known to be related to glaucoma progression. For example in Fig. 5, the RNFL circle scan on the right (Fig. 5.C) indicates an area of thinning that is in the temporal superior (TS) region (blue box), the deviation map below displays abnormal thinning in this region (*i.e.*, the black line crosses the red area), and one of the last real VFs also shows a defect (bottom left dark area) in the topographically corresponding nasal inferior region (Fig. 5.B). This information together may help the clinician assess the reliability of the prediction.

10-2 view. The clinician may decide to perform a 10-2 visual field test in cases of paracentral loss. The 10-2 view provides a prediction for the 10-2 VF MD and corresponding explanations, in a manner similar to the 24-2 view (Fig. 6). In this view, bounding boxes highlight salient features in the RNFL and GCIPL thickness maps (Fig. 6.C). For example, the important regions of the GCIPL mainly indicate an area that is in the superior region of the visual field (Fig. 6.C, the large white bounding box in the bottom thickness map). Comparing this area to the previous (actual) tests, we see inferior defects on the VF pattern (Fig. 6.B, bottom left dark area). Because the inferior region in the VF is topographically related to the superior region in the macula, and this region is highlighted (large white bounding box), the current prediction may be considered reliable.

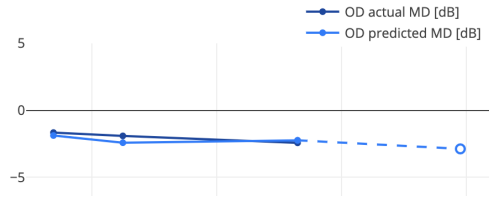
4.3. Implementation Details

GLANCE consists of a web-based system which is implemented using Python, Plotly Dash and Bootstrap. Image data is manipulated using and Matplotlib, OpenCV and PIL. Plotly is used as main graphing library. All computations are performed with Python.

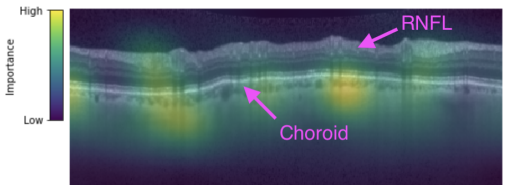
5. Use Cases

We now demonstrate our approach by two clinical use cases. Three clinicians were recruited, and asked to think out loud while interacting with GLANCE. The subjects considered in the case studies were not used for training the DL model. Mild, moderate, and advanced glaucoma patients were explored. We ended the sessions with some questions about future use and usage scenarios. Although we gathered input from three clinicians on an array of clinical scenarios, for illustrative purposes, we highlight the thought process and feedback of an individual clinician for two cases.

Patient 1. For the first exploration, we selected a patient with mild glaucoma that has not been progressing (based on VF) for a cou-

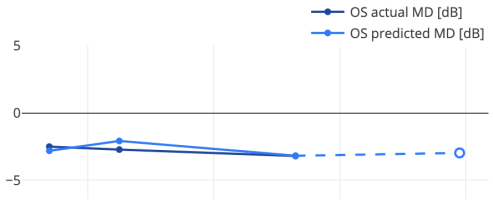


(a) Actual and predicted MD

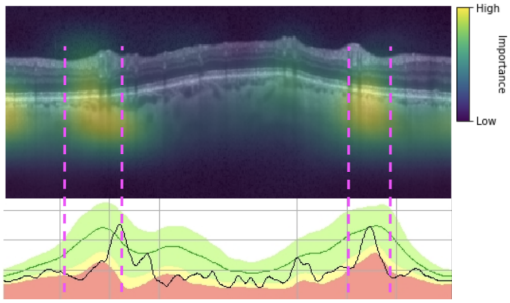


(b) Heat map of important regions

Figure 7: *The right eye (OD) of patient 1:* ***(a)*** *indicates that predictions aligned with actual measurements;* ***(b)*** *illustrates that important regions were not located on the RNFL but on the choroid.*



(a) Actual and predicted MD



(b) Heat map of important regions

Figure 8: *The left eye (OS) of patient 1:* ***(a)*** *predictions are consistent, but the current prediction failed to follow progression in the actual and previous predicted progression,* ***(b)*** *regions between the pink dotted lines show that important regions align with defects observed in the deviation plot.*

ple of years. The clinician started with the right eye. Looking at the graph with past and predicted MD (Fig. 7a), she noted that the model's predictions were fairly accurate (**T1**). Next, she checked the important regions for the prediction. She switched to the heat map to see more detail. She explained that this plot was confusing her, because the highlighted regions were not located on the RNFL layer, but more on the choroid (**T2**). This is strange from a clinical perspective because this area is generally not considered important (Fig. 7b). However, she also noted that these might be potential new landmarks (**T3**). Based on this information only, she would not trust the prediction. The predicted MD for this patient is -2.88 dB, which is slightly worse than the previous -2.43 dB. Because there is a progression trend both in the real measurements and the model predictions, she would want to see this patient again in 6 months instead of 1 year (**T4**). She justified her choice mainly by the progressive and irreversible character of glaucoma; it is important to track progression accurately to reduce further loss.

The model also showed consistent and accurate predictions (**T1**) for all measurements of the left eye (Fig. 8a). It can be seen that according to the model, there was no progression since the last visit. The important regions for this prediction were also located beneath the RNFL, but the RNFL deviation plot (Fig. 8b) showed that they were in line with defects on the RNFL (**T2**). The clinician concluded that a follow up in 6 months would not be necessary for the left eye (**T4**). This use case demonstrates that GLANCE enables a clinician to determine reliability of a prediction, and based on that, decide when the next visit and/or VF test should be planned.

Patient 2. For the second exploration, we looked at a patient with mild to moderate glaucoma. The clinician started with the right eye (Fig. 5). Looking at the graph with past and predicted MDs (Fig. 5.A), the clinician noticed that the model was always underestimating the magnitude of the actual MD, and that the current prediction even indicated an improvement or at least stability. Based on this information alone, the clinician would not trust the prediction (**T1**). To obtain more information, the clinician inspected the important regions on the right side of this view (Fig. 5.C). The important region (bounded by the blue rectangle) aligned clearly with the loss in the RNFL deviation plot beneath it (area in red). By hovering over the past points in the graph, the clinician confirmed that previous patterns also showed a clear loss (Fig. 5.B, black area) in the corresponding region (*i.e.*, the temporal-inferior (TI) region in the VF related to the defects in the nasal-superior (NS) region in the scan and deviation plot; addressing **T2**). The important regions looked fine, thus the clinician decided to explore further in the 10-2 view of this patient (Fig. 6). Because the clinician used the important regions information while reasoning about the defects, we conclude that the visualization is interpretable.

The model predictions for 10-2 VF MD looked reasonably consistent but overestimated the magnitude of MD (Fig. 6.A). Further, the clinician noted that the improvements in the actual MD may reflect a learning effect. There was progression visible in the last two actual values, but the predicted values were stable. The clinician explained that there are two possible trajectories for the predicted MD: either the current prediction is correct because it did follow the progression in the actual MD, or it is overestimating the actual MD as in the previous predictions (**T1**). Also, the important regions in the GCIPL (Fig. 6.C, large white square) corresponded to the loss in the past VF patterns (Fig. 6.B, black area), performing **T2** and **T3**. Based on all 10-2 VF information, the clinician expected the real MD to be on or above the predicted value. The predictions in the 24-2 view did not show progression, but actual values showed the opposite. The clinician decided that a real VF test would be needed to confirm whether or not there is progression and to be able to make a management decision (**T4**). For this patient, the left eye was not examined. From this use case, we can conclude that

information on the actual versus predicted values seems to be the main driver for determining reliability, even if the important regions indicate correct areas. Moreover, it shows that the MD prediction is not used for management decisions if it seems unreliable.

6. Pilot User Study

Building on the use cases, we conducted a pilot study with three independent clinicians regarding how GLANCE might inform clinical decision-making.

6.1. Study Setup

The clinicians (two males, one female) were recruited from the UCSD Shiley Eye Institute and were all familiar with the process of managing glaucoma progression using OCT technology. Their working experiences ranged from 2 to >10 years. We embedded three variants of GLANCE in an online survey. Variant 1 provided only the current prediction and standard OCT information, serving as a baseline without any explanations. In variant 2a, a line graph was added showing actual vs predicted points while variant 2b showed an activation map over the OCT scan. Variant 3 showed both the graph and activation map (see supplemental Fig. 1). Every participant was asked to provide a clinical recommendation for 12 patients using variants 1, 2a or 2b (randomly assigned per case), and 3, in that order, for each patient. For 6 additional patients they only viewed variant 3 to provide insights into learning effects that can occur while reviewing the initial 12 patients. Thus, in total each clinician reviewed 42 cases from 18 patients. Immediately after every case, they were asked to answer questions about trust in the prediction (supplemental Table 2). The trust measures were adapted from Jian *et al.* [JBD00]. We concluded the survey with demographic questions and open-ended items for general remarks.

6.2. Methods of Assessing Usability

User-centered design (UCD) is a widely adopted method to incorporate human factors, information science, and computer science into product and data interface design and is increasingly used in healthcare [LQG*15]. UCD processes are highly iterative and start with qualitative analyses involving a small number of participants in the early stages to explore themes and design needs prior to quantitative evaluations with a larger group. Given the novel nature of GLANCE, our initial usability assessments focused on qualitative analyses of clinicians' experiences with the tool.

First, to analyze whether a clinical decision could be based on predicted MD instead of actual VF MD (*H1*), we recorded the number of cases where clinicians provided a decision based on predicted MD. We further determined whether and how visual explanations provided added value (*H2, H3*) by assessing whether a recommendation changed when moving from variant 1 to 2a/2b to 3. Furthermore, we assessed clinician-reported trust ratings and confidence levels. We also recorded whether clinicians changed their treatment decisions after viewing the visualizations (*H4*). Finally, we performed a thematic analysis of clinicians' open-ended responses. Representative comments were identified for illustration.

6.3. Results

In the majority of cases, (68 of 126 cases (54%)), clinicians made a management decision based on the predicted MD. In the other cases, it was not possible to provide a recommendation without additional information, such as an IOP measurement or the true VF. Further, we found that in cases with advanced glaucoma, it is relatively difficult to provide recommendations based on predicted MD only because the VF is inherently more variable, whereas for mild cases the opposite seems to hold (Table 2).

Specifically, in 16 cases (22%), the initial recommendation was changed after reviewing variants 1 and 2 (see Table 3). For variants 2a and 2b, we see that 5 (14%) recommendations were changed compared to reviewing variant 1. For variant 3, 11 (31%) recommendations were changed compared to reviewing variants 2a and 2b (Table 3). Further, in 8 of the influenced cases (50%) clinicians decided not to obtain a VF, while in 30% of these cases they chose to obtain one. In the remaining 3 cases, they for example changed from "continue present management" to "increase treatment" or vice versa. In summary, the visualizations provided a way to determine the reliability of the prediction and in some cases helped clinicians to be more certain about their initial recommendation based on predicted MD *(H3)*. Moreover, the level of confidence and trust in the predictions also slightly increased with variants 2 and 3 which provided the historical accuracy of the model predictions. More importantly, we found that clinicians changed their treatment plan and trust ratings after viewing variant 3, indicating that the visual explanations provided them with new insights *(H3)* for individualizing treatment based on MD predictions *(H4)*. Furthermore, no learning effect was found, as the assessments for the 6 cases where only variant 3 was used were similar to the 12 cases which were reviewed with increasing level of visual explanation (*i.e.*, going through the pipeline 1-2a/2b-3). Below, we discuss these findings in more detail.

Determining reliability. Visual explanations helped to determine reliability of the prediction *(H3)*. All clinicians emphasized that whenever the previous predictions lined up with the actual, this rendered the prediction more reliable. For example, *"I do not trust the predicted MD because while there is good correlation with actual MD in the beginning, it is unable to predict slow decline in actual MD"* (clinician #2). The same was observed regarding explanation type 2, *i.e.* when the saliency regions corresponded to known clinical regions, the prediction was considered more reliable: *"I trust the predicted MD because it corresponds with actual MDs and appears to be focusing on important regions on RNFL"* (clinician #2). In another case, clinician #3 related level of trust in the prediction:*"Somehow not. The MD is predicted based on regions [...]*

Table 2: *Medical recommendations per disease level.*

	Recommendation based on prediction	Unclear/request actual VF
All disease levels	**68** (54%)	**58** (46%)
Early and mild disease	38 (90%)	4 (10%)
Moderate disease	25 (45%)	31 (55%)
Advanced disease	5 (18%)	23 (82%)

which are not the main regions that are affected by glaucoma and have influence on RNFL and VF.". This indicates that visualization of both past predictions and important regions help to determine reliability.

Progression and variability in results. In many cases, the clinicians reported that the combination of true versus predicted progression and variability was important. One combination is when the past predictions corresponded with the actual MDs: *"I trust the predicted MD because it appears to be within range of the actual MDs and the trend from previous predictions follows the actual trend"* (Clinician #2). Another combination occurs when previously predicted MDs follow the trend of the actual, but the magnitude is different: *"Although I cannot trust the predicted MD, now at least I know that the model is consistently underestimating the magnitude of the MD for the last several visits [...] I can deduce that it is underestimating again on this visit, and that this patient most likely has progression of his/her VF that merits escalation in treatment"* (Clinician #1). Lastly there is the situation when past predictions are not following any trend, making the current predicted value hard to interpret: *"There has been some discrepancy between the predicted and the observed in the past, so it's hard to know whether the predicted MD will reflect the true value on this visit"* (Clinician #2). This clinician recommended obtaining a VF test for this visit. From these examples we conclude that visualization can still support management decisions, even when the prediction seems unreliable.

Disease level matters. In advanced disease, all clinicians indicated that predicted MD was likely to be inaccurate and therefore risky. They explained that from a clinical perspective, the RNFL had likely "hit the floor" (*i.e.*, is extremely thin where further thinning likely would not be detectable), making predictions doubtful [BZW*17]. In advanced cases, clinicians take more "subjective" factors into account: *"It is difficult to give advice/recommendation for this patient, who has advanced glaucoma, based on this information alone. The OCT shows that the RNFL is essentially at the floor. I don't think any predicted MD based on OCT data will be useful [...], and no visualization will really have added value in this context either. Like one of the other earlier patients, I would need to depend on the patient's subjective symptoms, IOP, overall life expectancy, and co-morbid conditions"* (Clinician #1). However, for early disease, the MD measurement is a good indicator: *"I would trust the predicted MD given early disease, reasonable correspondence with the general MD trend, and measured RNFL thickness"* (Clinician #2). For another mild case, clinician #1 explained: *"I think the predicted value looks reasonable and follows the prior trend. Also, overall MD changes are very mild for this patient, so there is more room for error regarding the prediction."*. These findings suggested that the tool is mainly applicable to cases with mild to moderate glaucoma.

Table 3: *Influence of variant on medical recommendation.*

	No influence	Some influence
Variant 2a or 2b (compared to 1)	31 (86%)	5 (14%)
Variant 3 (compared to 2a or 2b)	25 (69%)	11 (31%)

Individualize treatment. All clinicians found that visual explanations increase trust and confidence about management decisions. Thus, if they became less trustful because of the visualizations, and therefore decided to obtain a VF, this was a positive result as the visualizations helped clinicians determine when to trust the DL model. For example, for one patient, clinician #1 stated *"Obtain a visual field for this visit"* for all three variations, rated trust in the prediction lower after the last, and explained: *"Now that I can see the prior predictions, it is clear that the predicted MD is consistently underestimating the magnitude of the true MD."*. For another case, after viewing more visualizations, clinician #2 changed the decision to continue the treatment plan to obtaining an actual VF to make his management decision, rated his trust in the prediction lower and explained: *"I do not trust the predicted MD because while there is good correlation with actual MD in the beginning, it is unable to predict slow decline in actual MD."*. Checking the ground truth MD, the prediction turned out to be very inaccurate. From these examples, it can be concluded that visualizations help to individualize treatment (*i.e.*, to determine when a VF is needed soon or when it can be delayed).

7. Discussion and Limitations

We have presented GLANCE, a visual analytics approach for monitoring glaucoma progression in the clinic. This type of analysis is novel in the field of glaucoma. While the use cases and pilot user study show promising results, there are also some limitations.

Validation based on use cases indicates that our approach can guide the frequency of future VF tests. In some cases (*e.g.*, very stable patients) the clinician might decide to skip the VF test. In other cases more frequent VF tests might be needed to ensure timely detection of progression and adjustment of the treatment. These results do not suggest that GLANCE can replace VF testing in itself; rather, it can help individualize the frequency to each patient.

The pilot user study provides initial insights on how clinicians change their judgement of glaucoma progression using GLANCE and whether different visual explanations affect their understanding and trust. Based on the pilot results, we can confirm that GLANCE can assist clinicians to assess reliability of MD predictions and decide whether a prediction can be used for disease management decisions. Moreover, we found that clinicians calibrated their trust and made decisions with slightly more confidence after viewing more informative visualizations. For the pilot, cases with all disease levels were considered. However, the results showed that GLANCE may primarily be useful for mild and moderate cases. Clinicians were excited about the possibilities, especially when the cases showed stability and VF testing could be postponed, thus saving the patient and health care system time and money. Because many of the patients in the clinic are stable [SRK*14], this indicates that GLANCE holds potential for efficient disease management by creating the opportunity to individualize treatment based on appropriate trust and confidence. This pilot study based on feedback from a small number of users provides interesting insights to support future investigations involving a larger number of users. Future work should preferably execute the user test in the clinic.

GLANCE was evaluated by human- and application-grounded

methods. DL models may be too complex to explain globally, but methods such as saliency maps can give insight in local behavior. Our approach used these methods, and they were tested for robustness and explanation adequacy using randomly permuted labels [AGM*18]. The results of this "sanity check" show that our saliency maps are sensitive to the permuted labels, meaning that they are able to express the relationship between the original input instances and their labels. In future work, additional investigation of faithfulness should be performed by comparing the outputs to brute-force methods such as occlusion testing.

The DL models in our approach all have fairly good accuracy in predicting the severity of functional loss, but we should keep in mind that the predicted MD values are inherently noisy [PKY*17], and that there is no ground truth for measuring or predicting visual function. The models can average out some artifacts and visual explanations can indicate incorrect-looking model predictions, but the knowledge and expertise of ophthalmologists are essential for interpreting the results in a safe manner.

GLANCE was designed to be simple and intuitive for in-clinic use. Some interactivity is built in, such as hovering or zooming, but the overall visualization is primarily static. The rationale to begin with a static visualization was based on multiple studies demonstrating cognitive load and burden of information processing imposed by health information technologies on clinicians [AL12, ZKS15]. Additionally, static visualizations are more amenable to being exported as PDFs for picture archiving and communication systems (PACS). However, at this stage of development we did not formally evaluate whether this design facilitated efficiency for clinical decision-making. Future analyses may help evaluate whether GLANCE can help ophthalmologists achieve time savings in clinical settings, whether in individual patient interactions or in overall clinic flows with reduced VF testing. These analyses can be accomplished via time-motion observations and/or analyses of electronic health record audit log data [BGM*20].

Acknowledgements

We thank Ali Tafreshi and Patricia Manalastas (MD) for their valuable feedback, comments and time. Grant support: United States National Eye Institute EY11008, EY19869, EY14267, EY027510, EY026574, EY029058, P30EY022589, T15LM011271, 5K12EY024225, T32EY026590 Unrestricted grant from Research to Prevent Blindness, NY. Research fellowship grant of the German Research Foundation (DFG), Grant-Nr: RE 4155/1-1 and German Ophthalmological Society (DOG) grant.

References

[ADD*11] Acharya U. R., Dua S., Du X., Chua C. K., et al.: Automated diagnosis of glaucoma using texture and higher order spectra features. *IEEE Transactions on Information Technology in Biomedicine 15*, 3 (2011), 449–455. 3

[AGM*18] Adebayo J., Gilmer J., Muelly M., Goodfellow I., Hardt M., Kim B.: Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems* (2018), pp. 9505–9515. 11

[AGS10] Abràmoff M. D., Garvin M. K., Sonka M.: Retinal imaging and image analysis. *IEEE Reviews in Biomedical Engineering 3* (2010), 169–208. 1

[AL12] Avansino J., Leu M. G.: Effects of CPOE on provider cognitive workload: a randomized crossover trial. *Pediatrics 130*, 3 (2012), e547–e552. 11

[ANE*15] Acharya U. R., Ng E., Eugene L. W. J., Noronha K. P., Min L. C., Nayak K. P., Bhandary S. V.: Decision support system for the glaucoma using gabor transformation. *Biomedical Signal Processing and Control 15* (2015), 18–26. 3

[BGC*20] Baxter S. L., Gali H. E., Chiang M. F., Hribar M. R., Ohno-Machado L., El-Kareh R., et al.: Promoting quality face-to-face communication during ophthalmology encounters in the electronic health record era. *Applied Clinical Informatics 11*, 01 (2020), 130–141. 2

[BGM*20] Baxter S. L., Gali H. E., Mehta M. C., Rudkin S. E., Bartlett J., Brandt J. D., Sun C. Q., Millen M., Longhurst C. A.: Multi-center analysis of electronic health record use among ophthalmologists. *Ophthalmology* (2020). 11

[BZW*17] Bowd C., Zangwill L. M., Weinreb R. N., Medeiros F. A., Belghith A.: Estimating optical coherence tomography structural measurement floors to improve detection of progression in advanced glaucoma. *American Journal of Ophthalmology 175* (2017), 37–44. 10

[CBB*18] Christopher M., Belghith A., Bowd C., Proudfoot J. A., Goldbaum M. H., Weinreb R. N., Girkin C. A., Liebmann J. M., Zangwill L. M.: Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Scientific Reports 8*, 1 (2018), 16685. 1, 3

[CBB*19a] Christopher M., Belghith A., Bowd C., Fazio M. A., Goldbaum M. H., Weinreb R. N., Girkin C. A., Liebmann J. M., Zangwill L. M.: Deep learning models predict visual function from macula thickness map. *Investigative Ophthalmology & Visual Science 60*, 9 (2019), 5600–5600. 3

[CBB*19b] Christopher M., Bowd C., Belghith A., Goldbaum M., Weinreb R., Fazio M., Girkin C., Liebmann J., Zangwill L.: Deep learning approaches predict glaucomatous visual field damage from oct optic nerve head en face images and retinal nerve fiber layer thickness maps. *Ophthalmology* (2019). 2, 3

[CBB*20] Christopher M., Bowd C., Belghith A., Goldbaum M. H., Weinreb R. N., Fazio M. A., Girkin C. A., Liebmann J. M., Zangwill L. M.: Deep learning approaches predict glaucomatous visual field damage from oct optic nerve head en face images and retinal nerve fiber layer thickness maps. *Ophthalmology 127*, 3 (2020), 346–356. 1, 3

[CPB*20] Christopher M., Proudfoot J. A., Bowd C., Belghith A., Goldbaum M. H., Rezapour J., Moghimi, et al.: Deep learning models based on unsegmented oct rnfl circle scans provide accurate detection of glaucoma and high resolution prediction of visual field damage. *Investigative Ophthalmology & Visual Science 61*, 7 (2020), 1439–1439. 3

[CRH*19] Cai C. J., Reif E., Hegde N., Hipp J., Kim B., Smilkov D., Wattenberg M., Viegas F., Corrado G. S., Stumpe M. C., et al.: Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), ACM, p. 4. 1, 3

[DFLRP*18] De Fauw J., Ledsam J. R., Romera-Paredes B., Nikolov S., Tomasev N., Blackwell S., Askham H., Glorot X., O'Donoghue B., Visentin D., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine 24*, 9 (2018), 1342. 3

[GCWGvW19] Garcia Caballero H. S., Westenberg M. A., Gebre B., van Wijk J. J.: V-awake: A visual analytics approach for correcting sleep predictions from deep learning models. In *Computer Graphics Forum* (2019), vol. 38, Wiley Online Library, pp. 1–12. 2, 3

[GMR*19] Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D.: A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR) 51*, 5 (2019), 93. 3

[HDM18] HOOD D. C., DE MORAES C. G.: Challenges to the common clinical paradigm for diagnosis of glaucomatous damage with oct and visual fields. *Investigative Ophthalmology & Visual Science 59*, 2 (2018), 788–791. 2

[HKPC18] HOHMAN F. M., KAHNG M., PIENTA R., CHAU D. H.: Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics* (2018). 3

[Hoo17] HOOD D. C.: Improving our understanding, and detection, of glaucomatous damage: an approach based upon optical coherence tomography (oct). *Progress in Retinal and Eye Research 57* (2017), 46–75. 2

[JBD00] JIAN J.-Y., BISANTZ A. M., DRURY C. G.: Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics 4*, 1 (2000), 53–71. 9

[KAKC17] KAHNG M., ANDREWS P. Y., KALRO A., CHAU D. H. P.: A cti v is: Visual exploration of industry-scale deep neural network models. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2017), 88–97. 2, 3

[KMCAS18] KHAIRAT S., MARC D., CROSBY W., AL SANOUSI A.: Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR Medical Informatics 6*, 2 (2018), e24. 3

[KZZ*15] KUANG T. M., ZHANG C., ZANGWILL L. M., WEINREB R. N., MEDEIROS F. A.: Estimating lead time gained by optical coherence tomography in detecting glaucoma before development of visual field defects. *Ophthalmology 122*, 10 (2015), 2002–2009. 1

[LHK*18] LI Z., HE Y., KEEL S., MENG W., CHANG R. T., HE M.: Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology 125*, 8 (2018), 1199–1206. 1, 3

[Lip16] LIPTON Z. C.: The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016). 3

[LKB*17] LITJENS G., KOOI T., BEJNORDI B. E., SETIO A. A. A., CIOMPI F., GHAFOORIAN M., VAN DER LAAK J. A., VAN GINNEKEN B., SÁNCHEZ C. I.: A survey on deep learning in medical image analysis. *Medical Image Analysis 42* (2017), 60–88. 3

[LQG*15] LUNA D., QUISPE M., GONZALEZ Z., ALEMRARES A., RISK M., GARCIA A. M., OTERO C.: User-centered design to develop clinical applications. literature review. *Studies in Health Technology and Informatics 216* (2015), 967. 9

[MMG14] MUSEN M. A., MIDDLETON B., GREENES R. A.: Clinical decision-support systems. In *Biomedical Informatics*. Springer, 2014, pp. 643–674. 3

[MWW*17] MIOTTO R., WANG F., WANG S., JIANG X., DUDLEY J. T.: Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics 19*, 6 (2017), 1236–1246. 1

[OTM*07] OSHEROFF J. A., TEICH J. M., MIDDLETON B., STEEN E. B., WRIGHT A., DETMER D. E.: A roadmap for national action on clinical decision support. *Journal of the American Medical Informatics Association 14*, 2 (2007), 141–145. 2

[PGI*12] PIERRO L., GAGLIARDI M., IULIANO L., AMBROSI A., BANDELLO F.: Retinal nerve fiber layer thickness reproducibility using seven different oct instruments. *Investigative Ophthalmology & Visual Science 53*, 9 (2012), 5912–5920. 1

[PHVG*17] PEZZOTTI N., HÖLLT T., VAN GEMERT J., LELIEVELDT B. P., EISEMANN E., VILANOVA A.: Deepeyes: Progressive visual analytics for designing deep neural networks. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2017), 98–108. 2, 3

[PKY*17] PHU J., KHUU S. K., YAPP M., ASSAAD N., HENNESSY M. P., KALLONIATIS M.: The value of visual field testing in the era of advanced imaging: clinical and psychophysical perspectives. *Clinical and Experimental Optometry 100*, 4 (2017), 313–332. 11

[Rat18] RATNER M.: FDA backs clinician-free AI imaging diagnostic tools, 2018. 3

[RSG16] RIBEIRO M. T., SINGH S., GUESTRIN C.: Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), ACM, pp. 1135–1144. 3

[SCD*17] SELVARAJU R. R., COGSWELL M., DAS A., VEDANTAM R., PARIKH D., BATRA D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 618–626. 2, 4

[SGPR17] STROBELT H., GEHRMANN S., PFISTER H., RUSH A. M.: LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2017), 667–676. 3

[SRK*14] SAUNDERS L. J., RUSSELL R. A., KIRWAN J. F., MCNAUGHT A. I., CRABB D. P.: Examining visual field loss in patients in glaucoma clinics during their predicted remaining lifetime. *Investigative Ophthalmology & Visual Science 55*, 1 (2014), 102–109. 10

[SWT07] SCHIRA M. M., WADE A. R., TYLER C. W.: Two-dimensional mapping of the central and parafoveal visual field to human visual cortex. *Journal of Neurophysiology 97*, 6 (2007), 4284–4295. 5

[SYC14] SHAIKH Y., YU F., COLEMAN A. L.: Burden of undetected and untreated glaucoma in the united states. *American Journal of Ophthalmology 158*, 6 (2014), 1121–1129. 1

[TLW*14] THAM Y.-C., LI X., WONG T. Y., QUIGLEY H. A., AUNG T., CHENG C.-Y.: Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology 121*, 11 (2014), 2081–2090. 1

[UMJ*19] URATA C. N., MARIOTTONI E. B., JAMMAL A. A., OGATA N. G., THOMPSON A. C., BERCHUCK S. I., ESTRELA T., MEDEIROS F. A.: Comparison of short-and long-term variability on standard perimetry and spectral domain optical coherence tomography in glaucoma. *American Journal of Ophthalmology* (2019). 2

[WAM14] WEINREB R. N., AUNG T., MEDEIROS F. A.: The pathophysiology and treatment of glaucoma: a review. *Jama 311*, 18 (2014), 1901–1911. 1

[WDZ*13] WALL M., DOYLE C. K., ZAMBA K., ARTES P., JOHNSON C. A.: The repeatability of mean defect with size III and size V standard automated perimetry. *Investigative Ophthalmology & Visual Science 54*, 2 (2013), 1345–1351. 1, 2

[YCN*15] YOSINSKI J., CLUNE J., NGUYEN A., FUCHS T., LIPSON H.: Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015). 3

[YLDC16] YU S., LEE G. C., DURBIN M. K., CALLAN T.: Repeatability of SITA standard and SITA fast visual fields. *Investigative Ophthalmology & Visual Science 57*, 12 (2016), 3926–3926. 2, 4

[ZCAW17] ZINTGRAF L. M., COHEN T. S., ADEL T., WELLING M.: Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595* (2017). 3

[ZDF*17] ZHANG X., DASTIRIDOU A., FRANCIS B. A., TAN O., VARMA R., GREENFIELD D. S., SCHUMAN J. S., HUANG D.: Comparison of glaucoma progression detection by optical coherence tomography and visual field. *American Journal of Ophthalmology 184* (2017), 63–74. 1, 2

[ZF14] ZEILER M. D., FERGUS R.: Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* (2014), Springer, pp. 818–833. 3

[ZKL*16] ZHOU B., KHOSLA A., LAPEDRIZA A., OLIVA A., TORRALBA A.: Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 2921–2929. 3

[ZKS15] ZAHABI M., KABER D. B., SWANGNETR M.: Usability and safety in electronic medical records interface design: a review of recent literature and guideline formulation. *Human Factors 57*, 5 (2015), 805–834. 11