

---

SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science  
Department of Computer Science  
MASTER THESIS

---



# Development of an AI-Driven Clinical Decision Support System for Treating Age-Related Macular Degeneration and Diabetic Retinopathy

submitted by  
Robert Andreas Leist  
Saarbrücken  
January 2025

---

**Advisor:**

Hans-Jürgen Profitlich  
German Research Center for Artificial Intelligence (DFKI)  
Saarland Informatics Campus  
Saarbrücken, Germany

**Reviewers:**

Prof. Dr. Daniel Sonntag  
German Research Center for Artificial Intelligence (DFKI)  
Saarland Informatics Campus  
Saarbrücken, Germany

Prof. Dr. Antonio Krüger  
German Research Center for Artificial Intelligence (DFKI)  
Saarland Informatics Campus  
Saarbrücken, Germany

Saarland University  
Faculty MI – Mathematics and Computer Science  
Department of Computer Science  
Campus - Building E1.1  
66123 Saarbrücken  
Germany

# **Declarations**

## **Statement in Lieu of an Oath:**

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis.

Saarbrücken, 17th June, 2024

## **Declaration of Consent:**

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken, 17th June, 2024

## Acknowledgements

I would like to thank everybody that supported me during the course of my studies and especially during the creation of this thesis even if they are not mentioned by name.

First, I want to sincerely thank my advisor, Hans-Jürgen Profitlich, who helped me in the creation of the idea for this thesis and supported me in every stage following.

Moreover, I want to thank Prof. Dr. Sonntag for giving me the chance to write the thesis at the research department for Interactive Machine Learning at the German Research Center for Artificial Intelligence in Saarbrücken. Additionally, I want to thank Prof. Dr. Krüger for reviewing this thesis.

I also want to thank Tim Hunsicker, who helped me conduct the user study and the qualitative analysis.

Special thanks goes to my family and friends, who accompanied and encouraged me during the development of this thesis. I want to especially thank the proofreaders amongst them.



## Abstract

As part of the OphthalmoAI project, this work aims to develop a Clinical Decision Support System (CDSS) to aid ophthalmologists in treating Age-related Macular Degeneration (AMD) and Diabetic Retinopathy (DR), the leading causes of vision loss in the elderly. Therapy for AMD and DR involves administering Anti-VEGF medication according to well-defined clinical guidelines. The core of the CDSS is a treatment recommendation algorithm based on these guidelines, utilizing real-world medical data from two German eye clinics, including Electronic Health Records (EHR) and Optical Coherence Tomography (OCT) scans. OCT scans are segmented using a Deep Learning (DL) model to compute semantic segmentation masks. These masks are used to reconstruct biomarkers in three dimensional space enabling fine-grained quantifications. Additionally, a Bidirectional Long Short-Term Memory (BiLSTM) network was trained to predict future biomarker developments from sequential patient data. Evaluation of the BiLSTM model reveals that it can reliably predict a patient's development even when generalizing to unseen patients. Shapley Additive Explanations (SHAP) were used to validate the forecast models clinical relevance. The CDSS integrates EHR data with computed quantifications and forecasts, presented through various enhanced visualization techniques. Feedback from eleven ophthalmologists indicated that the CDSS enhances efficiency, informedness, and user experience through its comprehensive data display, which fuses functionality from multiple old tools. Moreover, many participants stated to feel more informed through the quantification algorithms, titled the "future of indication" by one senior ophthalmologist. However, an investigation into trust in DL systems revealed initial skepticism, especially in the forecast system. The participants agreed that trust would only improve with prolonged use and control of the system. Feedback options and human-in-the-loop models could enhance trust according to ophthalmologists.

---

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Ophthalmology AI . . . . .	2
1.3	Optical Coherence Tomography . . . . .	3
1.4	Structure of the eye . . . . .	4
1.5	Eye Diseases . . . . .	5
1.5.1	Diabetic Retinopathy . . . . .	5
1.5.2	Age-related Macular Degeneration . . . . .	6
1.5.3	Treatment & Guidelines . . . . .	6
1.5.4	Indication & Biomarkers . . . . .	7
1.6	Research Goals . . . . .	7
<b>2</b>	<b>Related Work</b>	<b>9</b>
2.1	Explainable Artificial Intelligence . . . . .	9
2.1.1	Model-centric Explanations . . . . .	9
2.1.2	Data-centric Explanations . . . . .	11
2.2	Clinical Decision Support Systems . . . . .	11
2.2.1	Usability Guidelines . . . . .	12
2.2.2	Examples . . . . .	12
2.3	OCT Segmentation . . . . .	13
2.4	Time Series Forecasting in Healthcare . . . . .	14
2.4.1	Recurrent Neural Networks . . . . .	14
2.4.2	Bidirectional LSTMs . . . . .	14
2.4.3	Predicting Visual Acuity . . . . .	15
2.5	User Study Evaluation . . . . .	16
2.5.1	System Usability Scale . . . . .	16
2.5.2	Thematic Analysis . . . . .	16
2.5.3	Cohens Kappa . . . . .	17
<b>3</b>	<b>Implementation</b>	<b>18</b>
3.1	Technology Stack . . . . .	18
3.2	User Centered Design . . . . .	20

3.3	Data . . . . .	20
3.4	Alignment . . . . .	22
3.5	Segmentation model . . . . .	23
3.6	3D Reconstruction and Quantification of fluids . . . . .	23
3.6.1	Algorithm . . . . .	25
3.6.2	Quantification . . . . .	25
3.7	Prognosis model . . . . .	25
3.7.1	Data set . . . . .	27
3.7.2	Bias . . . . .	29
3.7.3	Architecture and Hyperparameter Tuning . . . . .	32
3.7.4	Results . . . . .	33
3.7.5	Explaining model predictions with SHAP values . . . . .	35
3.8	Treatment recommendation . . . . .	36
3.8.1	Algorithm . . . . .	39
3.8.2	Evaluation & Results . . . . .	39
3.9	Visual Components . . . . .	40
3.9.1	Top bar . . . . .	41
3.9.2	OCT Viewer . . . . .	42
3.9.3	History graphs . . . . .	48
3.9.4	Metrics . . . . .	48
3.9.5	Recommendation . . . . .	49
3.9.6	Infobox . . . . .	49
<b>4</b>	<b>User Study</b>	<b>53</b>
4.1	Preliminary Workflow Study . . . . .	53
4.1.1	Findings . . . . .	53
4.1.2	Low Fidelity Prototype . . . . .	55
4.2	Study Protocol . . . . .	55
4.3	Results . . . . .	56
4.3.1	Demographics . . . . .	56
4.3.2	Systems Usability Scale . . . . .	56
4.3.3	Interview Questions . . . . .	57
4.3.4	Qualitative Analysis . . . . .	57
4.3.5	Feature specific Feedback . . . . .	60
4.4	Final Improvement Iteration . . . . .	61
<b>5</b>	<b>Conclusion</b>	<b>62</b>
5.1	Prognosis model . . . . .	62

5.1.1	Discussion . . . . .	62
5.1.2	Future Work . . . . .	63
5.2	Recommendation System . . . . .	63
5.2.1	Discussion . . . . .	63
5.2.2	Future Work . . . . .	63
5.3	User Study . . . . .	64
5.3.1	Discussion . . . . .	64
5.3.2	Future work . . . . .	65
5.4	Potential Applications . . . . .	65
<b>Bibliography</b>		<b>67</b>
<b>A 3D Reconstruction algorithm &amp; Code</b>		<b>75</b>
<b>B Questionnaire</b>		<b>78</b>
B.1	Demographics . . . . .	78
B.2	Experience with AI and Software . . . . .	78
B.3	System Usability Scale . . . . .	78

# List of Figures

1.1	Overview of the high-fidelity dashboard prototype. . . . .	2
1.2	Overview of the OphthalmoAI architecture in german. . . . .	3
1.3	Example of an OCT from a patient of the eye clinic in Münster, recorded as part of the OphthalmoAI project [25]. A) shows the IRSLO view and B) shows a single slice from this OCT. . . . .	4
1.4	Overview of all retinal layers visible on an OCT. Taken from [31]. . . . .	5
1.5	Examples of lesions shown on OCTs from different forms of AMD. A = Dry AMD; B = Non-active wet AMD; C-F = Active AMD. Dark and yellow arrows indicate drusen and pigment epithelial detachment (PED), respectively. Blue and red dots indicate sub-retinal fluids (SRF) and intra-retinal fluids (IRF). The green triangle indicates scarring. Image taken from Fu et al. [26] . . . . .	8
3.1	Entity Relationship Diagram of the used data base. . . . .	21
3.2	Example of a slice from an older OCT (Slice 1; left) and a newer OCT (Slice 2; middle). One can see that the two images are positioned differently. Aligning slice 1 (right) gets rid of those differences and the images can be easily compared. . . . .	23
3.3	Example of a segmentation mask. . . . .	24
3.4	Visualization of the 3D reconstruction algorithm through steps A to F. Pigment epithelial detachments (PEDs) are reconstructed in this example. In step A the algorithm starts with iterating through the masks. In step B the algorithm finds the first annotation of PED. In the next step, step C, the algorithm cannot find another annotation. Hence, in the following step D it marks this as one PED object, which is completely reconstructed. Step E shows a multitude of completely reconstructed PEDs shown as volumes as well as two large WIP reconstructions shown as lines through their contours. Step F is the finalized reconstruction of PEDs. . . . .	26
3.5	Convex hull of a random set of points in a circle. . . . .	27
3.6	Linegraph of actual datapoints (blue) of visual acuity and volumes versus the interpolated data created by smoothed interpolation (green) versus the data predicted by the three months forecast model of the specific metric (red). A) shows the visual acuity. B) and C) show the volume of fluids. One can see that A) and B) accurately predict the development. In C) we can see that the model keeps predicting the presence of fluids, while none can actually be observed, although notably on a very small scale. . . . .	30
3.7	Violin plots of visual acuity (top), volume of fluids (middle) and number of fluids (bottom) values. Note the logarithmic scale for volumes and number of fluids. The violins show the distribution of data along the metric's possible values. All metrics are lower bound by zero. . . . .	31

3.8	Boxplots of the absolute error of test set predictions for visual acuity (top), total volume (middle) and number of fluids (bottom). Please note the logarithmic scale on the volume and number of fluids plot. The dashed line represents the mean and the continuous line the median of absolute errors. Dots above the boxplot represent outliers. Note that the data is lower bound by 0, hence outliers will only be above the boxplot. . . . .	34
3.9	SHAP values of the impact of the five most impactful features on the three month prediction model of visual acuity. Each datapoint shows the SHAP value of one single prediction. Positive SHAP values impacted the predicted value positively, while negative values impacted it negatively. The x axis shows how many visits ago that datapoint lies in the sequence. Red dots indicate high feature values, while blue dots indicate low feature values. For example: High visual acuity values on the second most recent visits indicate a positive impact on the prediction, whereas low values indicate a negative impact. The more visits ago, the less influence they have on the prediction except for the first visit, which has smaller impact. . . . .	37
3.10	SHAP values of the impact of the five most impactful features on the three month prediction model of total volume of fluids. Each datapoint shows the SHAP value of one single prediction. Positive SHAP values impacted the predicted value positively, while negative values impacted it negatively. The x axis shows how many visits ago that datapoint lies in the sequence. Red dots indicate high feature values, while blue dots indicate low feature values. For example: High volume values on the most recent visits indicate a positive impact on the prediction, whereas low values indicate a negative impact. The more visits ago, the less influence they have on the prediction. . . . .	38
3.11	Scheme of the recommendation algorithm. . . . .	40
3.12	Confusion matrix of the binary classification of treatment versus no treatment. . . . .	41
3.13	The six visual components (VC) of the dashboard: VC1 = Top bar, VC2 = OCT viewer, VC3 = History graphs, VC4 = Metrics, VC5 = Recommendation, VC6 = Infobox. Moreover, the sidebar can be seen, which yields functionality about patient selection. . . . .	42
3.14	Example of VC2's IRSLO view with comparison and segmentation tool turned on. The slider on the image can be moved to make the right IRSLO overlap more or less over the left IRSLO. A segmentation mask is overlayed on top of the IRSLO. It shows only the segmentations of drusen, PEDs and fluids. Its transparency can be controlled by the transparency slider in the bottom right. Segmentation, the comparison slider and the alignment can be turned on and off. The left side IRSLO can be selected via the dropdown at the top and the right side IRSLO can be selected via the dropdown at the middle of the toolbar. . . . .	44

3.15	Example of VC2's IRSLO view with comparison and thickness map tool turned on. The heatmap shows the differences in thickness of the ELM layer between the two OCTs. Turning off compare will only show the thickness map of that layer from the main OCT. The "Align" button aligns the heatmaps according to the alignment of the underlying IRSLO images. The layer to be displayed can be selected through a dropdown menu at the bottom of the toolbar. Additionally, one can change the transparency of the heatmap through a slider. The mean thickness change of the selected layer will also be shown in a color coded info box below the plot. Green colors indicate thickening, while red colors indicate thinning of retinal layers. . . . .	45
3.16	Example of VC2's Slice view with comparison and segmentation tool turned on. The slider on the image can be moved to make the right OCT overlap more or less over the left OCT. A segmentation mask is overlayed on top of the OCT. Its transparency can be controlled by the transparency slider in the bottom right. . . . .	46
3.17	Example of VC2's 3D graph feature. The graph shows the top side of the segmented retinal layers as transparent layers colored by the same color scheme as the segmentation. Fluids, drusen and PEDs are shown as volume objects colored in the same scheme. Through the slider on top, one can move the OCT slice inside the 3D graph. The 3D graph can be rotated and zoomed using the mouse. . . . .	47
3.18	Example of VC3. The visual acuity history graph of the right eye for a randomly selected AMD patient. The orange markers are the recordings of visual acuities connected by the orange line. They are not interpolated in this graph. The green vertical lines indicate treatment with the medication ranibizumab shown in the legend. Moreover, this patient has been treated with bevacizumab according to the legend. These treatments are out of scope and can not be seen. By panning the user can make these treatments visible. The green dashed line shows the expected change when continuing treatment with this medicament. The transparent lines in the legend can be clicked to turn their prognosis visible in the graph. . . . .	48
3.19	Example of VC4. The metrics show the difference between current and last visit in percent. . . . .	49
3.20	Four examples of VC5. A: Recommendation to abort is highlighted in red. B: Recommendation to switch medication. C: Recommendation to continue started series. D: Recommendation to not treat as no fluids are present. . . . .	50
3.21	Example of VC6' "Reasoning" feature. The reasoning tab shows, how critical metrics changed from the last to the current visit and how the system expects these values to change in the future. Through the "Select treatment" dropdown menu the user can select, for which treatment they want to see the expected values. . . . .	50
3.22	Example of VC6' "Visit Diff" feature. The "Visit Diff" tab shows the difference in annotations of the last and current visit and how they changed. The differences are highlighted through color coding and phrasing. . . . .	51

3.23	Example of VC6' Mean Thickness feature. The Mean Thickness feature shows a table of the development of retinal layers from the last to the current visit as well as by how much they increased or decreased. . . . .	52
4.1	Low fidelity prototype developed for this thesis and its visual components (VC): VC1 = Infobox, VC2 = OCT viewer, VC3 = History graphs, VC4 = Metrics, VC5 = Recommendation . . . . .	54
4.2	Overview of the final dashboard prototype. . . . .	61
5.1	Scheme of the fully integrated Clinical Decision Support System with Feedback functionality. Dashed, dotted and continuous lines indicate feedback, interaction and data flow, respectively. The data comes from the clinic and enter the databank of the CDSS. This data is fed to the models for training, to the recommendation system for treatment recommendation and to the visual components for visualization. The models provide predictions for both the recommendation algorithm and the visual components. The recommendation is also displayed in the visual components. The user (medical expert) is interacting with the visual components, the feedback system and the annotation tool. The feedback system captures the users feedback and can trigger a retraining of the models, select training data or provide bad model predictions to the annotation tool. The user can interact with the annotation tool to segment OCTs from the eye clinic or from the feedback system to feed segmentation data into the databank. . .	66



---

## List of Tables

3.1	Main programming language and libraries used for this thesis. This list might not be complete. . . . .	19
3.2	Classes, their number of samples in the dataset, what they annotate and their class dependent test dice scores. . . . .	24
3.3	Features that are generated from multiple annotations and the number of annotations they are generated from. For example the new feature "Bleeding" was generated by merging all features that contained "bleeding" or "hemorrhage" in their name, which accounts for a total of 501 features. . . . .	28
3.4	Target timeframes and their respective number of datapoints in the dataset. . . . .	32
3.5	All tuned parameters included in the hyperparameter tuning and their respective possible values. . . . .	32
3.6	Performance of prognosis models in mean and median absolute errors. Bold numbers indicate that the model was finetuned. AE = Absolute Error . . . . .	33

---

# Chapter 1

## Introduction

### 1.1 Motivation

The rise of Artificial Intelligence (AI) and Machine Learning (ML) technologies has initiated a transformative era across various domains. From utilizing Large Language Models (LLMs) to enhance customer support interactions to employing Stable Diffusion algorithms for generating artistic creations, and harnessing Deep Learning (DL) models for autonomous vehicle navigation, the influence of AI and ML is omnipresent.

This paradigm shift is increasingly permeating the medical field, where numerous ML models, including those for segmentation, classification, and LLM applications, have demonstrated performance comparable to or surpassing that of human experts [8, 74, 75, 49, 83]. Particularly within ophthalmology, the discipline dedicated to the study of the eye, ML models have exhibited remarkable efficacy [24, 43, 44].

For instance, RetInsight<sup>1</sup>, a company from Vienna, has developed two Clinical Decision Support Systems (CDSS) tailored for monitoring geographic atrophies and fluid-related volumes in ophthalmic patients using AI technologies. Although the internal workings of their algorithm is not publicly available, a three-dimensional reconstruction of fluids must be computed in order to assess volumetric quantities. These systems currently exist separately from the standard software systems utilized by medical professionals on a daily basis. There exists a considerable opportunity for enhancing patient care through the integration of standard systems, that provide medical patient records and imaging data, and the computational capabilities of AI.

Additionally, time series forecasting ML methodologies like Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) networks have been developed but are not yet widely applied in the medical field. Predicting the course of a disease from patient data could enable preventive therapies, helping maintain patients' health before symptoms even appear. While these models are highly effective, their application in Clinical Decision Support Systems (CDSS) for ophthalmology has not yet been explored.

---

<sup>1</sup><https://retinsight.com/>

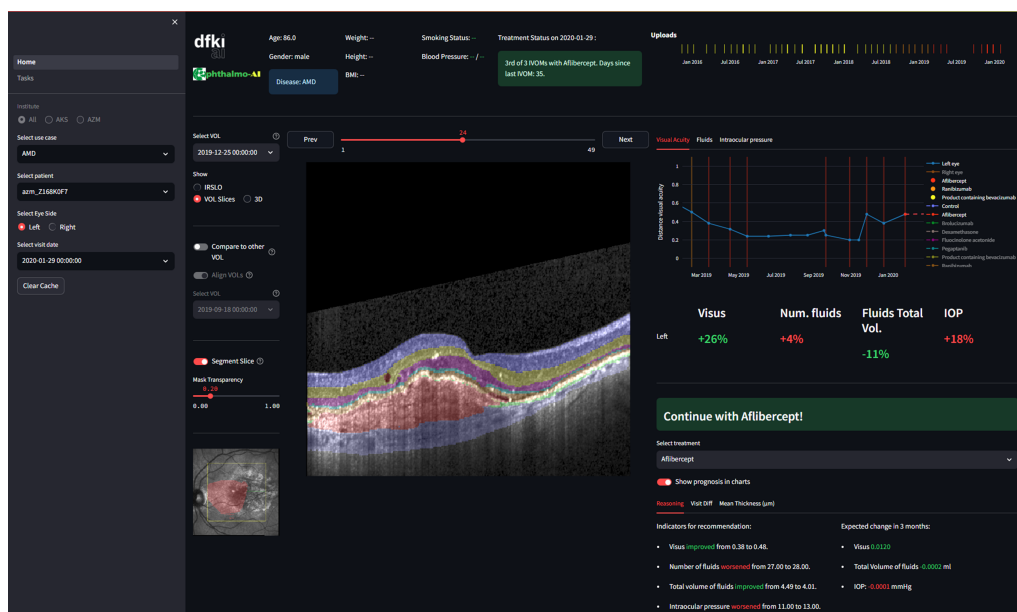


Figure 1.1: Overview of the high-fidelity dashboard prototype.

The performance of these algorithms and the trust of medical experts in their use still need to be investigated.

In the OphthalmAI project [25], discussed in detail in the next section, ophthalmology experts were consulted about their desired features in a CDSS. The ophthalmologists expressed the desire for a CDSS that provides an overview of metric trends through line graphs, projects layer segmentations onto retinal scans, quantifies fluid levels, recommends treatment and offers treatment efficacy prognosis. Figure 1.1 illustrates the CDSS developed based on these specifications and for which an evaluation was conducted in the following thesis.

## 1.2 OphthalmAI

The OphthalmAI project, funded by the German Federal Ministry of Education and Research (BMBF, funding label: 16SV8638)<sup>2</sup>, is a collaborative effort involving several esteemed institutions: the Fraunhofer Institute for Biomedical Technology (IBMT)<sup>3</sup>, the Interactive Machine Learning (IML) research department at the German Research Center for Artificial Intelligence (DFKI)<sup>4</sup>, the eye clinic at St. Franziskus Hospital in Münster<sup>5</sup> and the eye clinic in Sulzbach<sup>6</sup>, Heidelberg Engineering GmbH<sup>7</sup>, LangTec<sup>8</sup>, and the University of Saarland<sup>9</sup> [25]. The project's primary objective is to develop advanced

<sup>2</sup>[https://www.bmbf.de/bmbf/de/home/home\\_node.html](https://www.bmbf.de/bmbf/de/home/home_node.html)

<sup>3</sup><https://www.ibmt.fraunhofer.de/>

<sup>4</sup><https://dfki.de/web/forschung/forschungsbereiche/interaktives-maschinelles-lernen/>

<sup>5</sup><https://www.augen-franziskus.de/>

<sup>6</sup><https://www.augenklinik-sulzbach.de/>

<sup>7</sup><https://www.heidelbergengineering.com/de/>

<sup>8</sup><https://www.langtec.de/?lang=en>

<sup>9</sup><https://www.uni-saarland.de/start.html>

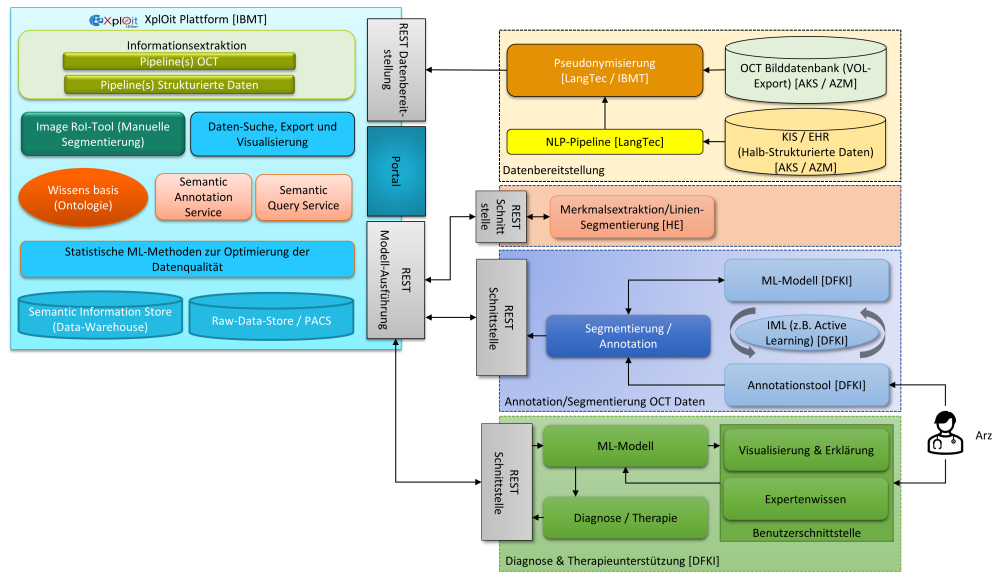


Figure 1.2: Overview of the OphthalmolAI architecture in German.

decision support systems to enhance the diagnostics and treatment of ophthalmological conditions. This involves combining clinical guidelines and medical expertise with machine learning (ML) and deep learning (DL) techniques. Additionally, the project aims to create visual, explanatory components that bridge the gap between black-box algorithms and medical professionals, ensuring a reliable and transparent support system for diagnosis and therapy. The focus is on two prevalent eye diseases: Age-related Macular Degeneration (AMD) and Diabetic Retinopathy (DR). This thesis is a contribution to the OphthalmolAI project.

Figure 1.2 shows an overview of the OphthalmolAI architecture and the areas of responsibility for each partner. IBMT developed and maintained the data warehouse XplOit and engineered major parts of the information extraction, integration and analysis. They were also responsible for assuring data quality. LangTec delivered tools for the pseudonymization and information extraction from medical text data. Semantic segmentation models were developed by the DFKI and Heidelberg Engineering GmbH. Semantic segmentation models are ML models, that segment images by classifying each pixel in the image. DFKI also realised several Artificial Intelligence (AI) components such as DL models and intelligent user interfaces including explanation and visualization tools. The eye clinics offered clinical knowledge and delivered the data. The university of Saarland was responsible for ethical, law and social aspects and the evaluation of the demonstrator prototypes.

### 1.3 Optical Coherence Tomography

Optical Coherence Tomography (OCT) is a non-invasive imaging technique widely used in ophthalmology to obtain high-resolution cross-sectional and en-face images of the retina. By using light waves to capture detailed images, OCT allows clinicians to visualize the layers of the retina. This technology works by measuring the echo time delay and intensity of reflected light, similarly to ultrasound imaging techniques, and is also called

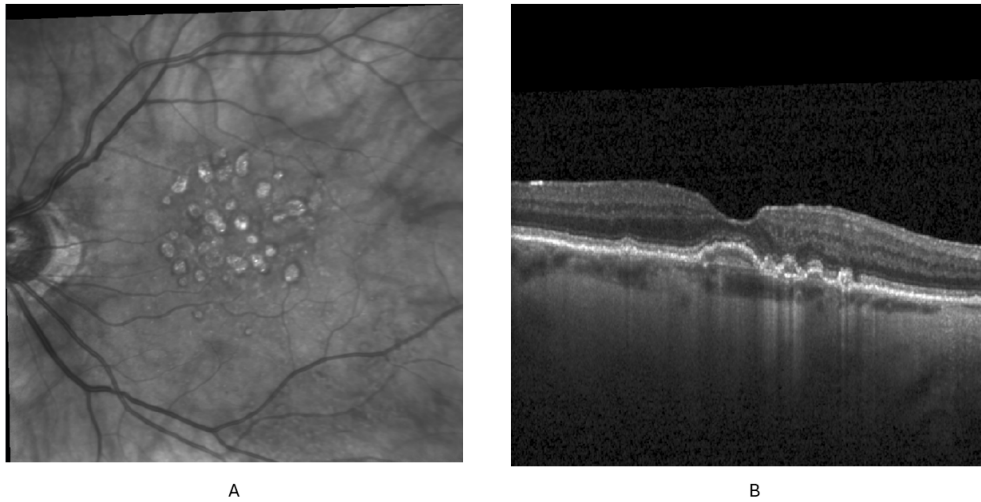


Figure 1.3: Example of an OCT from a patient of the eye clinic in Münster, recorded as part of the OphthalmoAI project [25]. A) shows the IRSLO view and B) shows a single slice from this OCT.

interferometry [37].

In general, an ophthalmologist has two options for assessing the health status of the eye. They either consult a fundus image or an OCT scan. The first one offers a top view onto the retina by essentially photographing the retina, which affords the patient to take dilation drops in preparation [55]. The second one offers both a top view -also known as Infra Red Scanning Laser Ophthalmoscopy (IRSLO)- and cross sectional views. It becomes apparent that the OCT yields more information and, additionally, it does not need preparation of the subject unlike fundus images [54]. Figure 1.3 shows an example of an OCT featuring the IRSLO view (A) and one slice (B).

## 1.4 Structure of the eye

The eye is an intricate organ responsible for capturing light and converting it into neural signals, which the brain processes to form images. Central to this function is the retina, a layered structure at the back of the eye. The retina contains several important layers, which can be seen on figure 1.4. The most important layers are: the Inner Plexiform Layer (IPL), which facilitates synaptic interactions between bipolar and ganglion cells; the Outer Plexiform Layer (OPL), where photoreceptors connect with bipolar cells; and the External Limiting Membrane (ELM), a thin barrier that supports photoreceptor cells. The Ellipsoid Zone (EZ), also known as the inner segment/outer segment junction, is crucial for photoreceptor integrity. The Retinal Pigment Epithelium (RPE) helps nourish retinal visual cells and is involved in the phagocytosis of photoreceptor outer segments, while Bruch's Membrane (BM) is a thin, multi-layered structure that supports the RPE and serves as a barrier between the retina and the choroid [21]. Although more layers exist, these were the ones that were classified as crucial by ophthalmologists. A segmentation

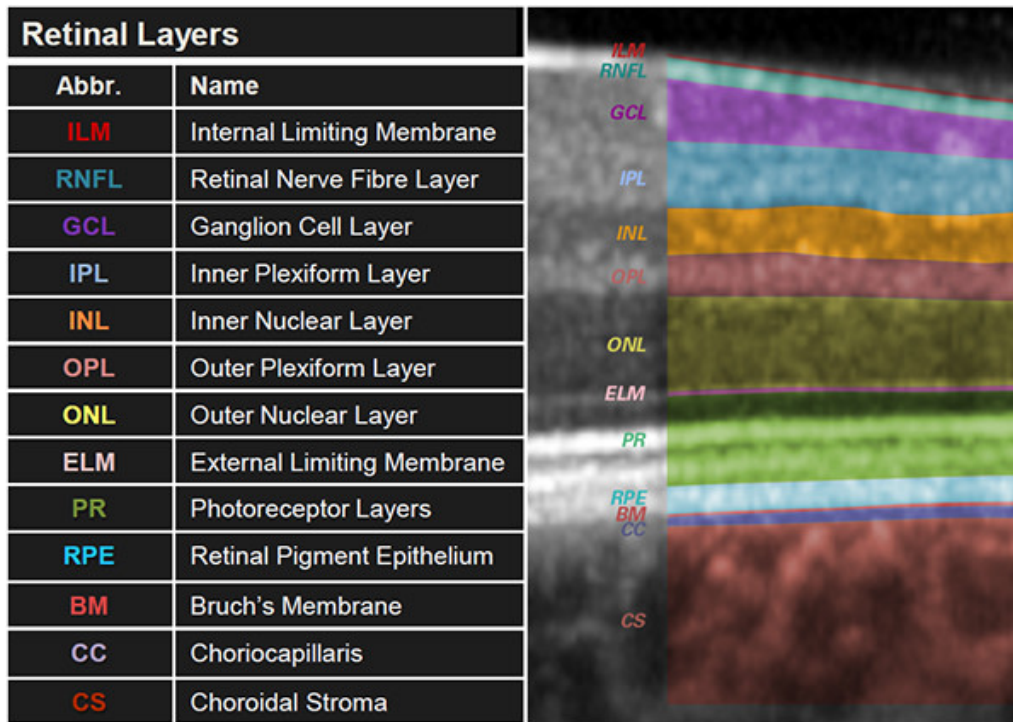


Figure 1.4: Overview of all retinal layers visible on an OCT. Taken from [31].

of these layers can enable quick and easy assessment of the structural integrity of the retina facilitating ophthalmologists' analysis tasks.

## 1.5 Eye Diseases

Diabetic Retinopathy (DR) and Age-related Macular Degeneration (AMD) are two prevalent eye diseases that significantly impact individuals' vision and quality of life. Both conditions require careful management and early detection to prevent irreversible vision loss. Their treatment, however, is still dependent on subjective evaluation of medical experts. The surge of segmentation architectures in the Optical Coherence Tomography (OCT) domain and corresponding automatic quantification algorithms could revolutionize this process as treatment can now be made dependent on quantifiable measurements. As most clinics still use rather old software AI could be used to improve hospital systems and further support ophthalmologists during their work. This section provides a brief overview of DR, AMD and their prevalence.

### 1.5.1 Diabetic Retinopathy

Diabetic retinopathy (DR) represents a serious complication of diabetes, affecting a substantial portion of individuals diagnosed with this metabolic disorder. Approximately 35% of diabetes patients grapple with the risks associated with DR [91]. The global scale of this issue is exemplified by the alarming statistics from 2021, estimating that 536.6

million people between the ages of 20 and 79 were living with diabetes. Forecasts predict that this number will surge to a staggering 783.2 million by the year 2045 [1]. This data suggests an alarmingly large population of roughly 187.8 million DR patients worldwide. DR is characterized by damage to the blood vessels in the retina, the light-sensitive tissue at the back of the eye. The condition is differentiated into two main stages: Proliferative DR (PDR) and Non-proliferative DR (NPDR). PDR is the first stage of DR, which can cause some vision impairment but generally goes unnoticed [17]. However, this stage is a crucial point for the patient as it is treatable and treatment avoids progression to the second final stage. Conversely, NPDR is not treatable and yields irreversible damages to the retina, which ultimately end in vision loss [17]. One symptom that majorly contributes is Diabetic Macular Edema (DME). DMEs are an accumulation of fluids inside the retina around the macula (The point of central vision), which is believed to be caused by hyperpermeability of blood vessels with abnormal blood sugar levels [58]. This hyperpermeability is caused by the upregulation of Vascular Endothelial Growth Factor (VEGF), a hormone that promotes the growth of new blood vessels [58]. Through treatment one can downregulate VEGF and consequently save the patient's vision. Hence, treatment is essential in managing the alarming numbers of DR patients effectively.

### 1.5.2 Age-related Macular Degeneration

Age-related Macular Degeneration (AMD) is a similar disease to DR. However, it is not related to any metabolic disorder but rather a symptom of old age. Similar to DR, it affects a large number of the population and is a leading cause for loss of central vision among the elderly [2]. In Europe alone, it is estimated to affect 67 million people with an expected increase of 15% by 2050 [48]. Globally, we are looking at numbers of 190 million people affected by AMD [6].

AMD primarily affects individuals aged 50 and older. It involves the progressive deterioration of the macula, which is responsible for sharp, central vision [2]. Like DR, early detection and timely intervention are critical in addressing AMD. AMD is divided into three stages: Early, intermediate and late stage. Additionally, one differentiates dry and wet AMD by the symptoms of AMD. Dry AMD is characterized by the development of drusen (Small accumulations of extracellular material), whereas wet AMD is caused by the growth of new blood vessels that can leak blood and fluids into the retina [2, 3]. Wet AMD leads to the loss of central vision, which causes a severe drop in life quality for patients [2]. Dry AMD is not treatable. However, similar to DR the growth of new blood vessels in wet AMD can be treated. The large population affected by this disease shows the importance of treatment of wet AMD to ensure the patients' vision.

### 1.5.3 Treatment & Guidelines

As mentioned before, both DR in its early stage and wet AMD are treatable [28, 29]. This treatment involves using a VEGF inhibitor like Aflibercept [67] or Brolucizumab [59]. These drugs are injected directly into the eye in a procedure called IVOM<sup>10</sup> and inhibit the growth of new blood vessels. In most cases, they cannot restore vision but prolong the further deterioration [2]. Currently, there exist two guidelines or schemes on when to upload IVOMs: Treat-and-Extend (TAE) [32] and Pro Re Nata<sup>11</sup> (PRN).

<sup>10</sup>From the German name for the procedure: "IntraVitreale Operative Medikamentenapplikation"

<sup>11</sup>Translation from Latin: As things stand

TAE starts with a series of three IVOMs for AMD treatment and six IVOMS for DME treatment with four week intervals inbetween. After each series there is a control examination usually involving an OCT Scan. If the findings do not show a worsening condition, also called stable or inactive findings, the interval between each IVOM will be increased by 2 weeks. Now both AMD and DME will be uploaded in series of three IVOMs until the next control examination. Intervals can increase to a total of 12 weeks between two IVOMs. If at any control examination, the findings got worse, called active or instable findings, then the intervals will be decreased by 2 weeks to a minimum of 4 weeks for the next series. If the series with maximum intervals has been uploaded, there will be a series of control examinations with 12 week intervals without IVOMs. Afterwards there will be one year of control examinations every 8 weeks to ensure that the OCT stays inactive.

The PRN treatment is much simpler. If the initial treatment conditions are met, there will be a series of three IVOMs with 4 week intervals inbetween. Afterwards there will only be control examinations with 4 weeks inbetween until the findings are active again. If the findings stay inactive, the patient just visits every 4 weeks to ensure no worsening of the condition. Otherwise there will be another series of three IVOMs.

#### 1.5.4 Indication & Biomarkers

Early diagnosis of AMD and DR in general is outside of the scope of this thesis. However, it is important how to find indicators or biomarkers for treatment, which are retinal fluids. In general, any type of fluid can be caused by either Diabetic Macular Edema (DME) or Choroidal Neovascularization (CNV) and in all cases an eye with retinal fluids should be treated [28, 29, 67, 59]. Figure 1.5 shows a variety of retinal lesions on OCT images taken from Fu et al. [26]. Lesions that indicate treatment are sub-retinal and intra-retinal fluids (SRF and IRF; seen as blue and red stars). Other lesions such as drusen (seen as black arrows), pigment epithelial detachment (PED; seen as yellow arrows) and scarring (seen as a green triangle) are also symptoms of AMD or DR. However, they are not relevant to the treatment indication. SRF and IRF can be seen as black holes inside the retinal layers.

Other biomarkers are visual acuity and retinal thickness. Visual acuity refers to the clarity or sharpness of vision, commonly measured using the decimal and Logarithm of the Minimum Angle of Resolution (LogMAR) scales. The decimal scale expresses acuity as a simple ratio, such as 1.0 for normal vision, 0.5 for half-normal vision, and 2.0 for double-normal vision. The LogMAR scale quantifies visual acuity based on the logarithmic scale of the angle of resolution. A LogMAR value of 0 corresponds to normal vision, with positive values indicating worse vision and negative values indicating better-than-normal vision [34]. In this work, decimal scale was used, because that was the most common measurement in the OphthalmoAI data. However, usually LogMAR is preferred in research. A conversion exists, but it is not reliable [51]. Hence, in the following visual acuity always refers to the decimal scale. Retinal thickness reduces with progressing AMD and is an early biomarker for the disease [93].

## 1.6 Research Goals

In the scope of this thesis, I endeavor to develop and evaluate a Clinical Decision Support System (CDSS) in the form of a dashboard tailored for ophthalmologists specializing in the domains of Diabetic Retinopathy (DR) and Age-related Macular Degeneration



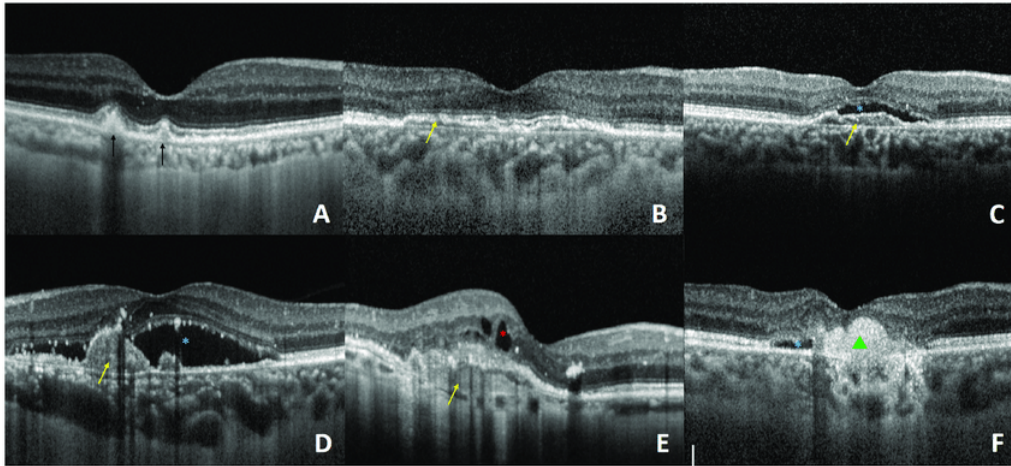


Figure 1.5: Examples of lesions shown on OCTs from different forms of AMD. A = Dry AMD; B = Non-active wet AMD; C-F = Active AMD. Dark and yellow arrows indicate drusen and pigment epithelial detachment (PED), respectively. Blue and red dots indicate sub-retinal fluids (SRF) and intra-retinal fluids (IRF). The green triangle indicates scarring. Image taken from Fu et al. [26]

(AMD), both of which are prevalent eye diseases. The objective of this work is to improve ophthalmologists' efficiency, informedness, and overall user experience by improving the therapy process. Efficiency, herein, refers to the reduction of medical professionals' workloads in the diagnosis and treatment of these diseases, consequently affording them more time to allocate to other aspects of patient care. Informedness pertains to the enhancement of an ophthalmologist's capacity to make well-informed treatment decisions. This involves the presentation of important data features crucial to decision-making and the provision of treatment recommendations given by established treatment guidelines. In pursuit of this goal, I aim to incorporate two AI methods: A segmentation model and a forecast model. An implementation of the segmentation model from the OphthalmoAI project will be used, while the forecast model will be developed and evaluated in this work. As these models may lack interpretability, it's essential to consider whether doctors comprehend and have confidence in the system. Consequently, this aspect of trust will also be examined. Moreover, the model's predictions will be evaluated using Explainable AI (XAI) methods to ensure its clinical relevance. Additionally, an investigation towards user experience will be done. Since suboptimal user experiences not only increases frustration but also the workload, it ultimately leads to reduced patient care. The goal of this thesis is twofold: First, create and evaluate a time series forecast model for ophthalmologic health metrics. Second, develop an holistic, AI supported Clinical Decision Support System in ophthalmology and assess its effects on efficiency, informedness, user experience and trust.

---

## Chapter 2

### Related Work

#### 2.1 Explainable Artificial Intelligence

In an era marked by the proliferation of AI systems and their increasing involvement in critical decision-making processes such as medical diagnostics, the demand for transparency and comprehensibility in AI algorithms has never been greater. Explainable AI (XAI) is an interdisciplinary field that aims to address this by providing a clear, intelligible, and interpretable framework for understanding the decisions and predictions made by artificial intelligence systems [56]. While some AI systems are inherently understandable like the coefficients of linear regression or the if-else-trees of decision trees, others such as Deep Learning (DL) approaches do not offer direct insights into their internal workings in an understandable way [11]. XAI encompasses many approaches to delivering explanations for such AI models: global and local, model-specific and model-agnostic, interactive and static, and data-centric and model-centric explanations [56]. Global XAI methods try to explain the general behaviour of models, while local methods explain the prediction for a certain input. Model-specific explanations are tailored towards a certain type of ML algorithm, while model-agnostic XAI is applicable to a wide range of AI models. Interactive XAI includes some sort of interaction between the model and the user. Literature suggests that this type increases understandability of model decisions more than static explanations [18, 12, 10]. The XAINES project even implies that XAI models should have interactive and incremental narratives, such that the user can have a conversation with the model in case of uncertainty [35]. XAI's interpretable and comprehensible explanations allow to compare an AI's decision making to clinical guidelines. Hence, XAI methods can be used to clinically validate a model's reasoning [5].

##### 2.1.1 Model-centric Explanations

Model-centric explanations delve into the internal workings of the AI model. They reveal the model's architecture, its decision-making processes, and the importance assigned to

different features during inference [56]. LIME trains an inherently interpretable model such as linear regression or decision trees to explain the original models predictions. RISE computes importance maps for visual input by manipulating said input and inferring each pixels importance towards the classification [62]. Grad-CAM looks at the gradients inside the model and infers an importance map [69]. In general, literature suggests that using model-centric explanations increases understanding of non-expert users [18, 12, 56]. However, Cheng et al. have found that the explanations do not necessarily increase trust in the system [18].

### Shapley Additive Explanations

Shapley Additive exPlanation (SHAP) values are an adaptation of the Shapley values, introduced by Lloyd Shapley in 1953 [71, 50]. Shapley values, originating from cooperative game theory, are used to fairly distribute both gains and costs among players based on their contributions. They assign each player an amount reflecting their contribution to the overall success of the coalition, ensuring fairness. Similarly, SHAP values measure the contribution of a feature to a model’s prediction, rather than the contribution of players to a game. Features, hereby, describe measurable characteristics such as blood pressure, which are fed as input to the model. However, SHAP is also applicable to image data, where the importance of a pixel is measured. The calculation of a SHAP value involves considering all possible subsets of features and determining each feature’s marginal contribution to these subsets. Specifically, for a feature in a subset, the SHAP value is the average of the marginal contributions of that feature across all possible subsets of the other features. Formula 2.1 shows how the SHAP value  $\phi_i$  for feature  $i$  is computed, where  $N$  is the set of all features,  $S$  is a subset of  $N$  that does not include feature  $i$ ,  $|S|$  is the number of features in subset  $S$ , and  $v(S)$  is the prediction for the subset  $S$ :

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \quad (2.1)$$

SHAP values have several desirable properties. They provide both local interpretability, allowing analysis of each feature’s impact on individual predictions, and global interpretability, enabling the assessment of the total impact of a feature by averaging its SHAP values over many predictions. Additionally, SHAP values are model-agnostic, meaning they can be applied to any machine learning model. Despite the computational challenges inherent in calculating exact SHAP values due to the need to evaluate all possible feature subsets, approximation methods and efficient algorithms, such as those implemented in the SHAP library<sup>12</sup>, make their practical use feasible.

### DeepSHAP

DeepSHAP is an extension of SHAP specifically designed for interpreting deep learning models, which was also introduced in the original paper [50]. Combining the principles of SHAP values with the DeepLIFT (Deep Learning Important FeaTures) algorithm, DeepSHAP provides a robust and efficient method for understanding the contributions of each feature in deep neural networks. DeepLIFT decomposes the output of a neural network by comparing the activation of each neuron to a reference activation, attributing contributions based on differences from this reference [73]. DeepSHAP uses this

<sup>12</sup><https://github.com/shap/shap>

decomposition to approximate SHAP values, making the computation feasible for large and complex neural networks. By combining these techniques, DeepSHAP retains the desirable properties of SHAP values, such as local accuracy and consistency, while being computationally efficient. This approach allows for insightful interpretations of deep learning models, enabling a deeper understanding of how features influence model predictions.

### 2.1.2 Data-centric Explanations

Another way of explaining a model's decision is to look at the data it was trained on. Data-centric explanations focus on illustrating the relationships between the input data and the model's output, the distribution of input data and potential biases [4]. Explanations can answer multiple questions like how was the data collected, how do the demographics behind the data look like, what is the recommended usage or are there any known issues with the dataset [4]. Data-centric methods have been found to increase trust in the model's decisions, as the user gets a better understanding of which and how much data has been used to create the model [4].

## 2.2 Clinical Decision Support Systems

Clinical Decision Support Systems (CDSS) are software systems that support medical experts in several tasks such as diagnostics, visualisation, data collection and decision-making [9]. Hence, CDSS can range from simple data viewer or data base programs to very complex AI applications. In today's medical field, as big data becomes more common, this latter type of CDSS is becoming more important. This comes to no surprise as AI especially DL works best on large amounts of data. AI can even outperform traditional systems or medical experts. Barnett et al. developed a system that detects Multiple Sclerosis lesions with a much higher sensitivity than traditional radiology reports [8]. Furthermore, researchers at Google created Med-PaLM and Med-PaLM 2 [74, 75], a large language model that can answer medical questions. It has been found that Med-PaLM's answers were preferred by study participants over physicians' answers. Moreover, DL architectures have been shown to have wrist fracture detection sensitivity of 81 to 92 %, which was significantly better than that of radiologists [49]. Topol identified a variety of AI models in medicine that can exceed humans through the analysis of big data in his review from 2019 [83]. He also warns of the limitations of AI. They need to be properly validated such that errors cannot happen, as they directly impact patient safety.

While the adoption of CDSS has been substantial in several medical domains [61], the application of such systems in the field of ophthalmology, particularly concerning Diabetic Retinopathy (DR) and Age-Related Macular Degeneration (AMD), has remained relatively underexplored [22]. Models that outperform human experts in this domain exist [45, 46, 53, 47, 82]. However, they usually stand as a sole demonstration of the model's performance versus the experts' performance without being implemented into a usable CDSS. Hence, the effects of the usage of such systems by experts is unknown. A review from 2011 screened 91 studies related to CDSS and found that approximately 57 % of them showed a significant improvement in practitioners' performance [41]. While these results are already promising, they were acquired before DL became popular. A recent review from Susanto et al. on ML based CDSS found that especially in image recognition ML methods can outperform medical experts and lead to improved patient

care [79]. Moreover, they showed that CDSS can reduce the time needed per diagnosis. In general, the CDSS' functions are designed to be beneficial for medical experts. However, in 2020 Sutton et al. [80] identified many risks associated with some functions of CDSS' that one should keep in mind in the creation process. A decision support should improve patient safety through alerts or treatment recommendation. But, if the system sends too many of these, the expert might grow numb to it and dismiss it regardless of the importance. This issue should be addressed by evaluating the CDSS thoroughly and asserting that treatment recommendations or alerts are given only when absolutely necessary [80]. Low importance alerts should be labelled as such. The system should also support diagnostics such that the expert can make use of the available data. However, this affords the expert to trust the system, which is seldom the case [61]. To leverage this problem, one should always reference expert knowledge such as guidelines or make the model explainable [80].

### 2.2.1 Usability Guidelines

When creating CDSS one needs to consider a few requirements, that the system must fulfill in order to be of use for the target group. Like any piece of software CDSS have to be usable. Usability includes many factors. According to the Usability Guidelines for Use Case Applications of the THESEUS Programme [77], a usable product is easy to learn, efficient to use, can quickly resolve errors, has easily memorizable control, provides enjoyment while usage and is pleasant to look at. These usability guidelines provide a variety of usability engineering methods. For this work specifically interesting are: User and task observations, interviews and questionnaires, benchmarking, and usability testing. User and task observations can be used to identify common tasks and problems for ophthalmologists in the DR and AMD domain. This is especially useful in the preliminary studies. Interviews and questionnaires can be used to evaluate a prototype. Benchmarking provides a way to assess the efficiency of the prototype. Additionally, usability testing provides a holistic analysis of the CDSS' usability in real scenarios. Moreover, according to the Clinical Data Intelligence (KDI) project [84] the integrated decision support must answer two questions: How can we get from medical guidelines like PRN or TAE to a decision support and how can we show the reason for that decision. While the first question is rather a technical question on how we get from our data to a treatment decision, the second question refers to XAI methods. The KDI project, therefore, also suggests that these decisions must be understandable. A review by He et al. from 2023 also suggests explanations to decisions made by the CDSS among 25 guidelines on how and what to design [90]. Additionally, they suggest that one should provide multiple options and alternatives for the treatment decision together with values indicating the certainty or uncertainty of these options. Moreover, one should include a tutorial to provide instructions on how to use the dashboard. The review found that the design in general should be simplified as much as possible to reduce the workload of the experts.

### 2.2.2 Examples

#### The Skincare Project

The Skincare Project by Sonntag et al. [76] is an example of a CDSS with a DL segmentation component. It features the analysis of pictures of moles, birthmarks and other

potentially carcinogenic skin conditions. The analysis follows a two step approach: First the image will be segmented using a DL architecture, then the classification is done using computer vision approaches and a diagnosis guideline called the ABCD rule. Similarly, one could design an approach for our use case: First segment the OCT, second compute relevant bio-markers such as the existence of retinal fluids and then decide according to the TAE or PRN guidelines, how to proceed. The CDSS also features an "Explain" button with several different callable methods such as RISE [62] and Grad-CAM [69] highlighting the need for XAI in the medical domain and showcasing the relevancy of the given classification.

### Monitoring Diabetes Onset

Bhattacharya et al. (2023) [10] have designed a CDSS for monitoring diabetes onset. They created their dashboard in a two stage process. First they developed a low fidelity (LoFi) click-through prototype. After evaluation with health care experts, they created a high-fidelity prototype with actual functionality and evaluated again including both experts and non-experts. Their dashboard involved an ML algorithm, which predicted the risk score of getting diabetes from certain health metrics and how this score would change, when the patient lost weight or was physically more active. A key question they strived to answer was how to explain the models decision making to the users. The study found that data-centric approaches were most used by the study participants in order to explain the models decisions. Model-centric explanations were less understandable to users and, hence, were not consulted much. They also claim that the interactivity with the model's predicted risk score when selecting different actions helped alleviate trust of the users in the system.

## 2.3 OCT Segmentation

Machine Learning methods have become a useful tool for the analysis of medical images [81, 30]. Especially, DL architectures such as Convolutional Neural Networks (CNNs) for classification or UNet architectures [65, 94, 38] for segmentation have been very successful. OCT segmentation, along with medical segmentation more broadly, involves the classification of each pixel in a medical image into its corresponding biological functional unit or abnormality. In the OCT domain, pixels will be assigned either to retinal layers or to lesions such as fluids or PED. UNet architectures feature a downsampling and upsampling branch, which each consist of CNN blocks forming the name giving "U" structure. Each block of one branch has a skip connection to a corresponding block in the other branch, which allows for pixel wise segmentations. Based on this architecture Farshad et al. created the YNet architecture [24]. YNet expands the UNet architecture by adding another encoder branch, which uses fast Fourier transform blocks to analyze the spectral domain instead of CNN blocks, which represent the spatial domain. Through this expansion the network improved upon the UNet architecture in segmenting fluids on the OCT dataset DUKE [19, 24].

EdgeAL introduced by Kadir et al. [43] is another expansion that builds on the YNet architecture. This version makes use of Active Learning (AL) methods to reduce the annotation cost of OCT datasets. AL methods incorporate human experts into the training process by querying the user to label uncertain inputs [70]. Through including edge entropy and edge divergence information into the training EdgeAL can outperform other models on very few data, while still being equally performant on large amounts of data.

EdgeAL will be trained on the same dataset as the one that is being used for this work as both are part of the OphthalmoAI project. An AL model in the ophthalmology domain can be used to query medical experts to annotate uncertain OCT segmentations. Consequently, the model can continuously learn and improve, while doctors can supervise the model.

## 2.4 Time Series Forecasting in Healthcare

As Bhattacharya et al. [10] demonstrated, an interactive prognosis tool helps users trust and understand the decision support system. Moreover, several studies show that treatment response can be predicted using ML methods. Jin et al. have developed a DL model that predicts the treatment response of chemoradiotherapy for rectal cancer patients from magnetic resonance imaging scans and blood-based biomarkers with an AUC of 0.95 [42]. Furthermore, many studies use ML methods to predict drug response of depression patients [87, 78, 88]. Additionally, as the aforementioned guidelines suggest one should implement different options and treatment alternatives together with values indicating their certainty or uncertainty [90]. The health development prognosis can be seen as such a value. Therefore, in this thesis I will implement a ML model, that predicts the development of certain health metrics such as fluid quantifications and visual acuity depending on treatment factors. I expect this model to help the experts see why the model made a decision as they can see how the model expects the patients health to change. Moreover, I want to include the option for the expert to select any other treatment decision and see the expected changes, as literature suggests that interaction with the model improves trust [18, 12, 10, 35]. However, one needs a precise model for this health metric prediction, because bad performance negatively impacts trust [63].

A meta analysis of time series prediction models in healthcare by Morid, Sheng and Dubar from 2022 has shown that this field is growing rapidly [57]. The most commonly used DL architectures were Recurrent Neural Network (RNNs) architectures. Among them base RNNs, Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks.

### 2.4.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of artificial neural networks designed to process sequential data by maintaining a hidden state that captures information about previous elements in the sequence [52]. Unlike traditional feedforward neural networks, RNNs have connections that form directed cycles, enabling them to retain information across time steps. This makes RNNs particularly well-suited for tasks where the order and context of data points are crucial, such as time series forecasting. In practice, RNNs are trained using backpropagation through time (BPTT), which adjusts the network's weights based on the error gradients calculated over multiple time steps. However, standard RNNs often face challenges with long-term dependencies due to issues like vanishing and exploding gradients, which can hinder learning [60].

### 2.4.2 Bidirectional LSTMs

Long Short-Term Memory (LSTM) networks are designed to effectively capture long-term dependencies in sequential data. Introduced by Hochreiter and Schmidhuber in

1997, LSTMs address the limitations of traditional RNNs [36]. The LSTM architecture includes special units called memory cells, each containing three gates: the input gate, the forget gate, and the output gate. These gates regulate the flow of information, allowing the network to maintain and update the cell state over time selectively. By doing so, LSTMs can remember important information for extended periods and are particularly well-suited for tasks involving time series data, natural language processing, and speech recognition. Their ability to handle long-range dependencies makes LSTMs a powerful tool for modeling complex temporal dynamics and sequential patterns in various applications.

Bidirectional Long Short-Term Memory (BiLSTM) networks introduced by Graves and Schmidhuber extend the capabilities of standard LSTM networks by processing data in both forward and backward directions, thereby capturing information from past and future contexts simultaneously [33]. In a typical LSTM, information flows in a single direction (from past to future), which can limit the network's ability to understand dependencies that rely on future context. BiLSTMs address this limitation by combining two LSTM layers: one that processes the input sequence from the start to the end (forward LSTM) and another that processes it from the end to the start (backward LSTM).

The outputs of these two LSTM layers are then concatenated at each time step, providing the network with a richer representation of the input data. This bidirectional approach allows BiLSTMs to leverage context from both directions, making them particularly effective for tasks where the meaning of a given element in the sequence depends on both preceding and succeeding elements, such as in natural language processing (e.g., part-of-speech tagging, named entity recognition, and machine translation) and speech recognition. By considering the entire sequence context, BiLSTMs enhance the network's ability to model complex dependencies and improve performance. The aforementioned study found that bidirectional LSTMs and GRUs outperformed their unidirectional counterparts and the basic RNN architecture in the medical domain [57]. Therefore, in this work a BiLSTM network will be developed and evaluated for the time series forecast of several ophthalmologic metrics.

### 2.4.3 Predicting Visual Acuity

A study by Rohm et al. from 2018 [64] deployed several traditional ML algorithms in order to predict visual acuity in AMD patients for three and twelve month time frames. They report a LogMAR Mean Absolute Error (MAE) of 0.11 and 0.16 for their best three and twelve month prediction model, which was a linear model trained with L1 regularization (Lasso). However, in this study the input to the model consisted of features from the current visit combined with aggregated values from past visits. Hence, it fails to capture more meaningful sequential relations from past visits. The authors also reported that missing data was a problem in the creation of the dataset. To mitigate this problem they used different aggregation techniques such as mean, min and max operations to cluster the data from multiple visits together.

Another study from Shi et al. published in Scientific Reports in 2023 [72] investigates accuracy of several traditional ML regression models to predict the short term efficacy after Anti-VEGF treatment. The models were trained on real-world data and predicted several clinical indicators including LogMAR visual acuity. The best performing model was a regression tree, which achieved 0.03 MAE for the LogMAR visual acuity and a determination coefficient ( $R^2$ ) of one implying an extremely good model. However, their data set contains the data of 280 patients only, because of strict inclusion and exclusion



criteria. Therefore, variation among the patients might be low. Moreover, the input data only considers two visits, while in a real life setting much more data would be available. A recent study from 2024 by Schlosser et al. [68] have applied several ML methods including RNNs and LSTMs to classify patients into a Winner, Stabilizer and Loser scheme (WSL). Winners are visits, where the LogMAR visual acuity increased by at least 0.1, while losers are visits where it decreased by the same amount. Stabilizers lie within this range. The models were trained on the data of several window sizes from four up to ten included visits. A Multi-Layer Perceptron (MLP) performed best and even outperformed ophthalmologists with an F1-Score of 69%. However, while the model can show a tendency, the range of its classes leaves some uncertainty restricting its usefulness in a clinical setting.

## 2.5 User Study Evaluation

User studies are an important tool to evaluate the usability of systems. In this thesis, a qualitative user study with experts will be conducted, hence, qualitative analysis tools are needed. Moreover, the usability of the system must be quantified in a comprehensible manner.

### 2.5.1 System Usability Scale

The System Usability Scale (SUS) is a widely used tool for evaluating the usability of a system, product, or service developed by John Brooke in 1986 [16]. SUS consists of a ten-item questionnaire that participants respond to using a five-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree." The items alternate between positive and negative statements to reduce bias. The final score, ranging from 0 to 100, provides a quantitative measure of the system's overall usability, with higher scores indicating better usability. Formula 2.2 shows how the SUS score is calculated, where PQ and NQ represent positive and negative questions, respectively. SUS delivers a numerical value on the usability of a software system, and, hence, gives a comparable metric.

$$SUS = 2.5 \left( 20 + \sum_{i=0}^5 PQ_i - \sum_{j=0}^5 NQ_j \right) \quad (2.2)$$

### 2.5.2 Thematic Analysis

Thematic analysis is a qualitative research method employed to identify, analyze, and report patterns or themes within data. It was introduced by Clarke and Braun in 2006, who continually refined and improved their method [13, 14, 15]. This approach is particularly effective for interpreting complex textual data, such as interview transcripts, survey responses, or field notes, by systematically categorizing and organizing the data to uncover recurring themes that address the research questions. The process involves several steps: familiarization with the data, generating initial codes, searching for themes among these codes, reviewing and refining the themes, and producing the final report. Thematic analysis is valued for its flexibility, as it can be adapted to various theoretical frameworks and research objectives. This method's capacity to provide deep, nuanced

insights makes it a vital tool in qualitative research. For time reasons and lack of study participants, this approach will be used to analyze the underlying themes emerging from the usage of the dashboard instead of a quantitative approach.

### 2.5.3 Cohens Kappa

Since the thematic analysis is inherently subjective, a measurement is needed that ensures the validity and reliability of the reported themes and codes. Cohen's Kappa ( $\kappa$ ) is a statistical measure used to evaluate the level of agreement between two raters who each classify items into mutually exclusive categories [20]. Unlike simple percent agreement, Cohen's Kappa accounts for the agreement occurring by chance, thus providing a more accurate assessment of interrater reliability. The value of Cohen's Kappa ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement better than chance, and negative values indicate disagreement not by chance. The calculation of Cohen's Kappa involves constructing a contingency table that shows the frequency of each rater's classifications and then applying formula 2.3, where  $p_o$  is the observed agreement (the proportion of times the raters agree) and  $p_e$  is the expected agreement by chance, calculated from the marginal totals of the table. By adjusting for chance agreement, Cohen's Kappa provides a more robust measure of the consistency between raters than a simple agreement percentage.

$$\text{Cohen's } \kappa = \frac{p_o - p_e}{1 - p_e} \quad (2.3)$$

---

## Chapter 3

# Implementation

In this chapter I will go over the implementation details of the high fidelity dashboard prototype, that was used to answer the research goals. First, general information about the technology stack and the selected design approach will be given. Then, I will explain several backend aspects such as the data and used algorithms. Finally, i will describe the visual components of the high fidelity prototype. The general structure of this prototype was developed from a low fidelity prototype, which was evaluated by one ophthalmologist (for more information see section 4.1.2). Figure 1.1 shows an overview of the high fidelity prototype. The dashboard consists of a sidebar and a main page. The sidebar can be collapsed and mainly holds functionality for patient selection. The main page features an info bar on top, a medical image visualization tool on the left and other data visualization methods on the left. Not all features can be seen on this figure. Please refer to section 3.9 for complete and detailed descriptions of them. The high-fidelity CDSS prototype was a web application hosted by the DFKI.

### 3.1 Technology Stack

The front- and backend of the dashboard is built using Python. Python offers a simple syntax as well as many libraries for data analysis, ML and visualizations. You can find all used programming languages and libraries in table 3.1 along with their version, usage and URL to their respective homepage. For the storage and retrieval of the data, SQLite was used. For the creation of the dashboard, Streamlit was utilized because of its simplicity and its free and easy-to-use addons from the community. Streamlit is a python package that translates python code into javascript in order to create interactive, fast and good looking websites. Its builtin features for data visualizations made it especially useful. For tasks related to computer vision, OpenCV's python version was used. PyTorch and SciPy helped with machine learning and deep learning tasks, while Pandas and NumPy were used for data processing. To understand model explanations, I used the SHAP Python package. Finally, for creating visualizations, I turned to the user-friendly Plotly package.

Name	Version	Use	URL
Python	3.11.1	Core language	<a href="https://www.python.org/">https://www.python.org/</a>
SQLite	3.46	Database	<a href="https://www.sqlite.org/">https://www.sqlite.org/</a>
Streamlit	1.31.0	Frontend	<a href="https://streamlit.io/">https://streamlit.io/</a>
OpenCV Python	4.8.1.78	Computer Vision	<a href="https://opencv.org/">https://opencv.org/</a>
PyTorch	2.1.0	Deep Learning	<a href="https://pytorch.org/">https://pytorch.org/</a>
SciPy	1.11.3	Machine learning	<a href="https://scipy.org/">https://scipy.org/</a>
Scikit-Learn	1.3.2	Data processing	<a href="https://scikit-learn.org/stable">https://scikit-learn.org/stable</a>
Pandas	2.2.1	Data analysis	<a href="https://pandas.pydata.org/">https://pandas.pydata.org/</a>
NumPy	1.24.3	Numerical computations	<a href="https://numpy.org/">https://numpy.org/</a>
SHAP	0.44.1	SHAP computations	<a href="https://github.com/shap/shap">https://github.com/shap/shap</a>
Plotly Python	5.15.0	Plotting	<a href="https://plotly.com/python/">https://plotly.com/python/</a>
YNet model	-	Segmentation Model	<a href="https://github.com/azadef/ynet">https://github.com/azadef/ynet</a>

Table 3.1: Main programming language and libraries used for this thesis. This list might not be complete.

## 3.2 User Centered Design

This dashboard was developed following a User Centered Design (UCD) process. User-Centered Design (UCD) is an iterative design process that prioritizes the needs, wants, and limitations of the end-users at every stage of the design process. This approach ensures that the final product is highly usable and provides a positive user experience. Key stages in UCD include user research, ideation, prototyping, and usability testing. In a first iteration, a low-fidelity prototype was developed following a preliminary workflow assessment interview. This prototype was evaluated by one medical expert. The feedback was used to create the high-fidelity prototype described in this chapter. Section 4.3 reports the feedback to this prototype.

## 3.3 Data

The data used in this thesis was collected in the course of the OphthalmoAI project [25]. It originates from two different eye clinics in Sulzbach and Münster, which are specialised in the treatment of DR and AMD. The clinics provide OCT and electronic historical records (EHR) to a data warehouse called Xpl0it [27]. The EHR was anonymized and text annotations were created from Systematized Nomenclature of Medicine (SNOMED) data from Xpl0it using Natural Language Processing (NLP) techniques by LangTec. OCT and EHR data was then saved in a data warehouse. Through an annotation tool developed by the DFKI, which is based on the Hierarchical Universal Modular ANnotator (HUMAN) [89], medical experts from the clinics could provide pixel-level segmentation masks to the OCTs.

A data base was constructed, which contained six tables, that can be seen in figure 3.1: visits, patients, patient\_labels, oct\_files, xploit\_parameters and fluids. The first five tables could be extracted from the data warehouse Xpl0it, while the last table was computed using a quantification algorithm (see section 3.6). The visits table contained an identifier key, a visit date and a patient identifier connecting it to the patients table. The patients table contained a patient's identifier key, pseudonymized birthday (birthday are replaced by some random offset without changing the age significantly), gender, diagnosis (AMD or DR) and, from which eye clinic the data comes. The xploit\_parameters table contains a parameter identification key, a german and english description of the parameter, its python and xploit type and a unit. The patient\_labels table combines the latter two tables containing identifiers for the patient and a parameter. Moreover, it contains the parameter's value, which eye side this parameter was annotated for, an OCT URI, if one is associated, and a date for, when this parameter was annotated. In total, 3,192 different parameter types were presented in the database with 1,656,611 total annotations. The oct\_files table contained a file URI, which function both as a key and a file location. Moreover, it contained some information about the OCT like number of slices, resolution parameters, patient and visit identifiers, eye side and a preparation date, which indicates when this OCT was added to the database. The fluids table was saved quantification data, which will be explained in section 3.6. This table used the file URI as a key identifier, as it contained the quantification results of that associated OCT. It contained information about the volume and number of fluids, drusen and PEDs. Moreover, it contained the mean thickness values of all seven annotated layers as well as the total retinal thickness.

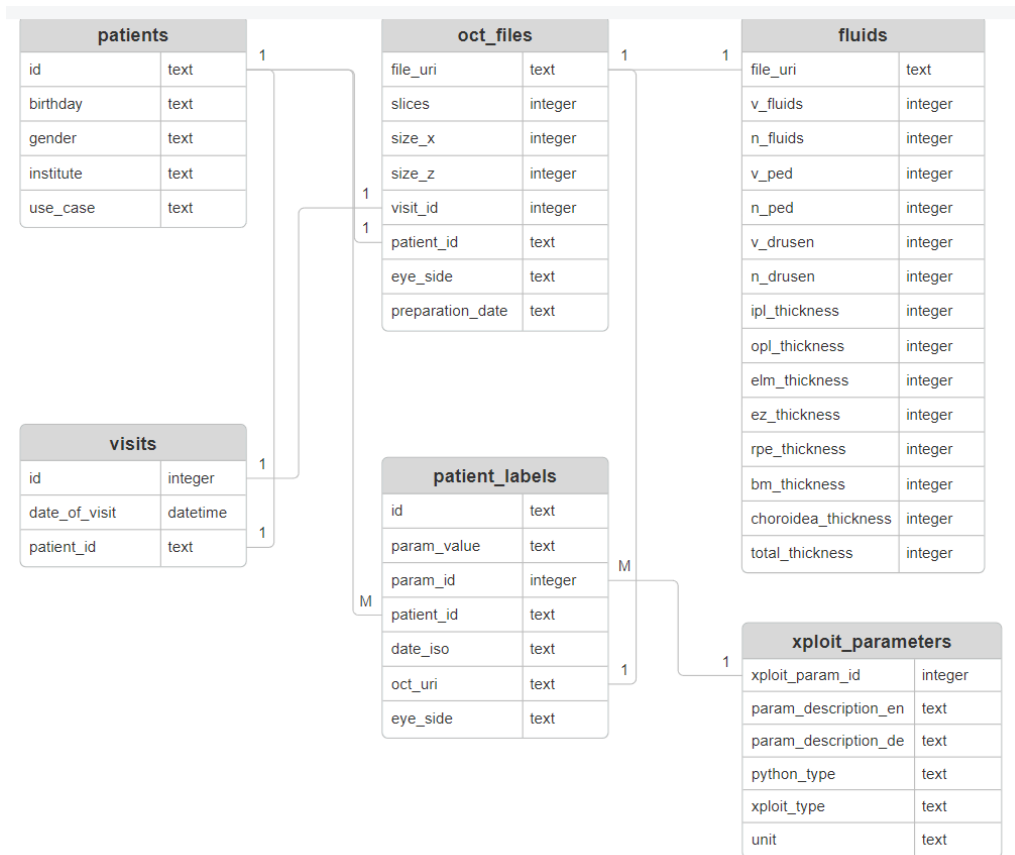


Figure 3.1: Entity Relationship Diagram of the used data base.

### 3.4 Alignment

During development, a significant problem emerged, complicating visual comparisons: OCT slices and IRSLO images were rarely correctly aligned with data from previous visits. Despite follow-ups being intended for alignment, this consistency was lost during transmission to the data warehouse. Consequently, an algorithm was needed to rectify this misalignment. The algorithm for aligning two images can be seen in listing 3.1. Alignment is very important to identify local changes in scans. An example of two misaligned OCTs and the computed alignment can be seen in figure 3.2.

```

1  def align(image1, image2):
2      keypoints1, descriptors1 = ORB(image1)
3      keypoints2, descriptors2 = ORB(image2)
4
5      matches = bruteforceMatch(descriptors1, descriptors2)
6      matches = sortByDistanceAscending(matches)
7      best_matches = matches[:100]
8
9      points1 = getKeypoints(keypoints1, best_matches)
10     points2 = getKeypoints(keypoints2, best_matches)
11
12     transformation_matrix = computeAffineMapping(points1, points2)
13     return warp(image1, transformation_matrix)

```

Listing 3.1: Pseudocode of the alignment algorithm

First, the Oriented FAST and Rotated BRIEF (ORB) algorithm was used. The ORB algorithm is an efficient and robust method for image alignment that combines keypoint detection and descriptor computation [66]. Initially, ORB uses the Features from Accelerated Segment Test (FAST) algorithm to quickly detect keypoints in an image by examining the intensity of a circular neighborhood around each pixel. To ensure the selected keypoints are stable and well-distributed, it refines them using the Harris corner measure. To achieve rotation invariance, ORB orients the keypoints based on the direction of the intensity centroid within the local neighborhood. For descriptor computation, ORB employs the BRIEF (Binary Robust Independent Elementary Features) descriptor, which is a binary string representing the local image patch. ORB modifies BRIEF to be rotationally invariant by aligning the patch orientation with the keypoint's orientation.

First, the descriptors are matched using a bruteforce method, which compares each descriptor from the first image with each descriptor from the second image. Hence, giving the most accurate matches possible. Next, these matches are sorted by their distance. From these the best 100 matches are being selected for keypoint extractions. A set of 100 points from each image will be taken and an affine mapping between these points, the transformation matrix, will be computed. An affine mapping includes translation, rotation, shearing and scaling. Finally, the transformation matrix is applied to the first image. Now both images should be aligned. The quality of the alignment depends on how good the descriptors could be matched. Sometimes, the alignment lead to bad results. An algebraic way of assessing alignment quality was looking at the transformation coefficients<sup>13</sup>. Formula 3.1 shows a possible two dimensional transformation matrix, where  $c_i$  are the rotation, shearing and scaling influencing coefficients and  $t_x$  and  $t_y$  are the translation coefficients in x and y direction. If the sum of  $c_i$  becomes large, one can assume that alignment is incorrect, as usually slices must only be rescaled, sheared and rotated a little bit. Similarly, large translations can be assumed to be incorrect. Therefore,

<sup>13</sup>For more information see [https://en.wikipedia.org/wiki/Transformation\\_matrix](https://en.wikipedia.org/wiki/Transformation_matrix)

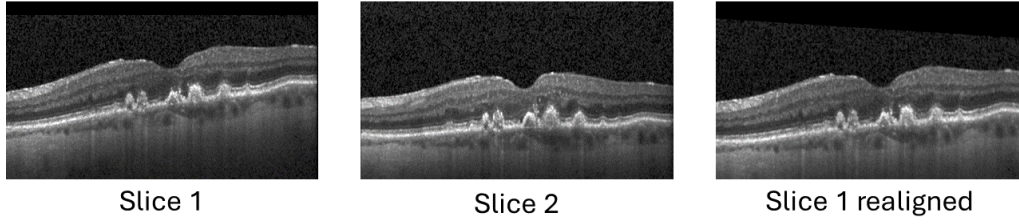


Figure 3.2: Example of a slice from an older OCT (Slice 1; left) and a newer OCT (Slice 2; middle). One can see that the two images are positioned differently. Aligning slice 1 (right) gets rid of those differences and the images can be easily compared.

the dashboard shows a warning (see section 3.9.2), when the sum of  $c_i$  exceeded 2.5. This threshold was selected by examining bad alignments and their corresponding sum of  $c_i$ .

$$\text{Transformation Matrix} = \begin{pmatrix} c_{11} & c_{12} & t_x \\ c_{21} & c_{22} & t_y \\ 0 & 0 & 1 \end{pmatrix} \quad (3.1)$$

### 3.5 Segmentation model

DFKI researchers trained a segmentation model on the acquired OCTs and annotations. Its architecture was a YNet architecture trained specifically on the OCT data annotated in the OphthalmoAI project. The code was taken from the github repository<sup>14</sup> of the original YNet paper [24]. The training data included 1233 OCTs which were labelled with eleven classes, which can be seen in table 3.2. The average mean dice score was 0.762. The model was trained on the sum of the dice loss and a cross entropy loss for 100 epochs with a batch size of 32. The cross entropy loss was used to classify pixels into whether they are correctly classified or not. This is especially useful for lesions as the number of samples for these classes is rather low. The dice loss is used to evaluate how good a predicted segmentation is. It is similar to the Intersection-over-Union (IoU) score. However, instead of dividing the intersection by the union, it divides two times the intersection by the sum of the areas of the prediction and the groundtruth [95]. This model was trained as part of the OphthalmoAI project and not as part of this work. The weights were shared for the creation of the OphthalmoDashboard [25].

The model takes as input one OCT slice and computes a semantic segmentation mask. Semantic segmentation means that every pixel will be classified into one of the eleven classes. Figure 3.3 shows an example of a segmented OCT slice.

### 3.6 3D Reconstruction and Quantification of fluids

To be able to quantify fluids and other lesions, it was necessary to first reconstruct the OCT in a three-dimensional space. Usually, assessing the presence and severity of lesions

<sup>14</sup><https://github.com/azadef/ynet>



Class	Number of samples	Annotation Type	Class dependent dice score
ILM	1233	Layer	0.98
IPL	1218	Layer	0.91
OPL	1210	Layer	0.8
ELM	1190	Layer	0.58
EZ	1190	Layer	0.46
RPE	1214	Layer	0.59
BM	1219	Layer	0.52
Choroidea	1211	Layer	0.85
Drusen	937	Lesion	0.79
PED	523	Lesion	0.9
Fluids	577	Lesion	0.88

Table 3.2: Classes, their number of samples in the dataset, what they annotate and their class dependent test dice scores.

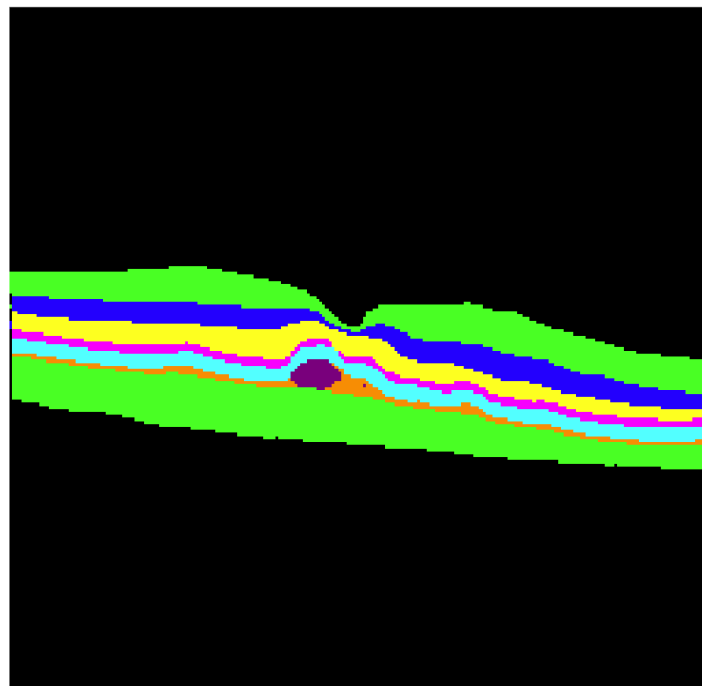


Figure 3.3: Example of a segmentation mask.

depends on a subjective measurement of the doctors. The reconstruction followed a scanning approach, whereas the slices were scanned from top to bottom and lesion objects were being tracked. The 3D reconstruction was only done on lesions and not on layers, as the algorithm is quite performance heavy and the lesions are clinically more relevant.

### 3.6.1 Algorithm

The 3D reconstruction works by iterating over the segmented masks produced by the segmentation model and creating reconstruction objects. Figure 3.4 shows an overview on how the algorithm works visually. More detailed information on the reconstruction algorithm can be found in appendix A.

### 3.6.2 Quantification

The quantification of these reconstructions can easily be done by computing the convex hull of all points in the reconstruction. In my research, I utilized subpackage "spatial" of the SciPy Python package [85]. This subpackage makes use of the Qhull library to compute the convex hull of a point cloud [7]. The convex hull of a set of 3D points is the smallest convex polyhedron that encloses all the points within its volume. Figure 3.5 shows a random set of points generated in a circle and their convex hull. Moreover, convex means that you cannot draw any line between any two points, which would not also lie within the hull. This means that the hull does not have any dents or indentations. This might lead to inaccurate reconstructions, where the volume is larger than it should be. However, convex hulls can be efficiently computed. At its core, Qhull employs the QuickHull algorithm, a robust and fast method for computing convex hulls in multidimensional spaces. This algorithm iteratively constructs the convex hull by partitioning the point set into subsets and recursively identifying the facets of the convex hull. Other algorithms such as Delaunay Tesselation [86], which offer more accurate objects, are more performance heavy. Since the efficiency inside the dashboard was critical and the differences between these reconstruction methods was marginal, I decided to stick with the convex hull reconstruction. From the convex hull one could easily compute the volume of a reconstructed object.

## 3.7 Prognosis model

In order to be able to identify and quantify patterns in the patient data and make them available for a treatment recommendation, a prognosis model was developed and trained. The prognosis model was trained to predict certain health metrics such as visual acuity and number and total volume of fluids. Visual acuity is a measurement that describes a patient's vision. Different scales exist for acuity, most notably the decimal scale and the Logarithm of the Minimum Angle of Resolution (LogMAR) scale. In the following, visual acuity will always refer to the decimal scale. Through this forecast I made sure to include most of the patient's history data and make an informed decision about the effectiveness of the treatment. Hence, the recommendation of my system is not only based on the presence of fluids as the guidelines would suggest, but it also includes other features like patient age, smoking behaviour, BMI and other risk factors, which a medical expert would use in their decision.

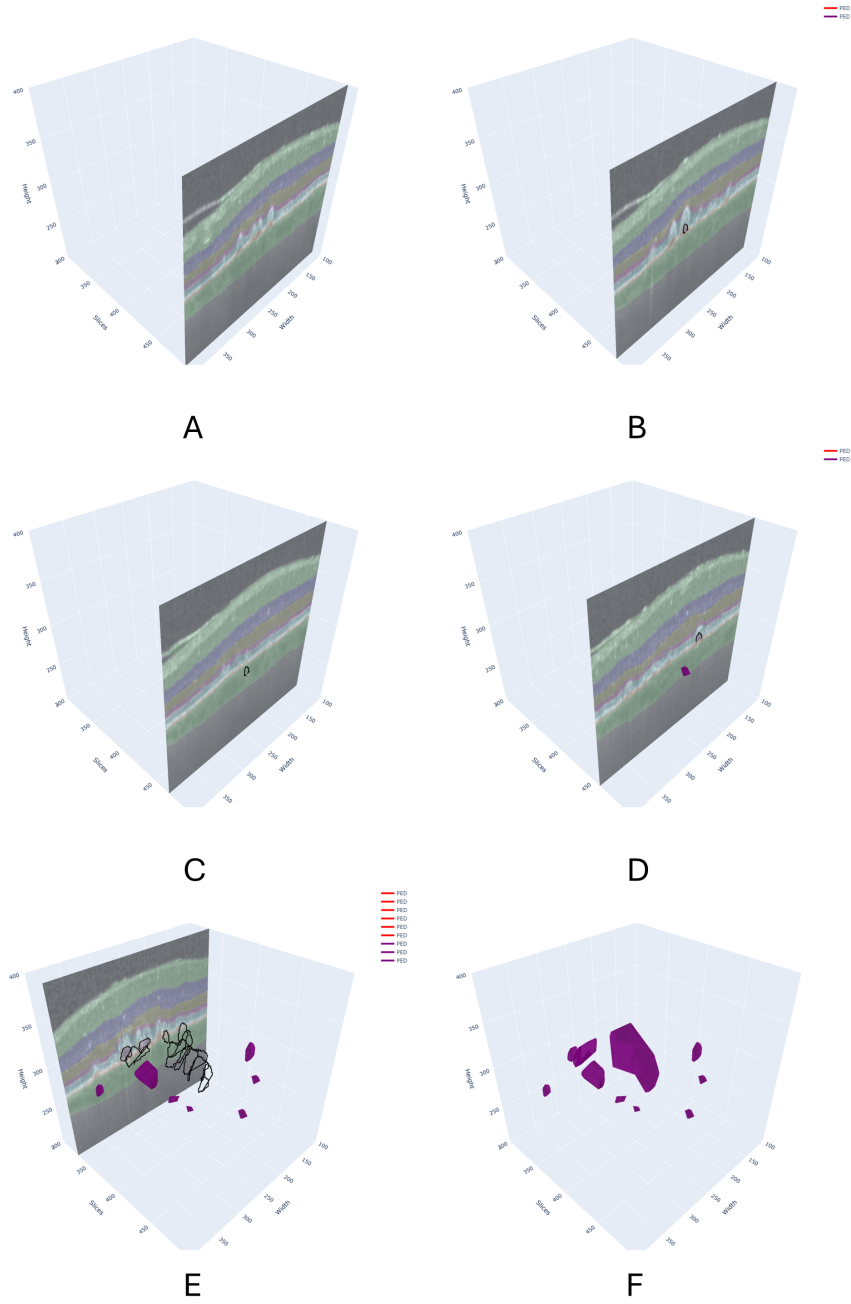


Figure 3.4: Visualization of the 3D reconstruction algorithm through steps A to F. Pigment epithelial detachments (PEDs) are reconstructed in this example. In step A the algorithm starts with iterating through the masks. In step B the algorithm finds the first annotation of PED. In the next step, step C, the algorithm cannot find another annotation. Hence, in the following step D it marks this as one PED object, which is completely reconstructed. Step E shows a multitude of completely reconstructed PEDs shown as volumes as well as two large WIP reconstructions shown as lines through their contours. Step F is the finalized reconstruction of PEDs.

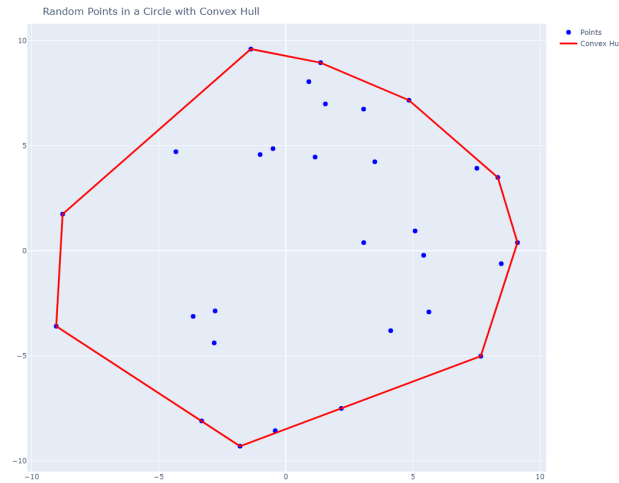


Figure 3.5: Convex hull of a random set of points in a circle.

### 3.7.1 Data set

In this section, the process of the creation of the data set will be discussed. Firstly the feature selection process will be highlighted. Second, the preprocessing steps will be discussed. Then, I will explain the data split and, finally, the data point generation process will be explained.

#### Feature Selection

The dataset was generated from the data of the OphthalmologyAI project (see section 3.3 and the analysis of the OCT files. Eight features were directly accessible from the data annotations: age, gender, diagnosis (AMD or DR), eye side, Body Mass Index (BMI), visual acuity, IOP, treatment type and smoking behaviour. Additionally, some features were computed from the visit dates and the treatment type such as the number of days since first/ last visit/ treatment or the total number of given IVOMs. Note that for the features BMI and smoking behaviour data was not always available. Hence, statistics from the year 2021 of several body measurements including BMI of the german population was consulted to supplement a value for the BMI [23]. If no smoking behaviour was specified, the patient was treated as a non smoker.

Other features had to be generated from multiple annotations, since using all annotations would lead to very sparse and unnecessarily large input data. Table 3.3 shows the new features and the number of annotations that were used in the generation of them. All new features are binary features that show whether or not the feature is present currently. These features are generated by searching for certain keywords in the annotations and merging them together. For example the "Bleeding" feature was generated by searching for all annotations that contained either the word "bleeding" or "hemorrhage". Whenever a bleeding or hemorrhage was present and annotated, then the new feature "Bleeding" will also be true. If none was annotated or the absence of such was annotated, then that feature will be set to false. In total, I could reduce the number of features from 2202 annotations to 17 binary input features.

Moreover, the analysis of the OCT was used to add more features. Through the 3D

New feature	Number annotations used for generation
Edema	130
Bleeding	501
Aneurysm	62
Inflammation	15
Scars	351
Ischemia	3
Degeneration	56
Neovascularization	269
Exsudate	77
Cysts	35
Tumors	8
Glaucoma	5
Hypertrophy	20
Atrophy	203
Detachments	69
Drusen	244
Hyaline Bodies	156

Table 3.3: Features that are generated from multiple annotations and the number of annotations they are generated from. For example the new feature "Bleeding" was generated by merging all features that contained "bleeding" or "hemorrhage" in their name, which accounts for a total of 501 features.

reconstruction (see section 3.6) the quantification in number and volume of lesions such as Fluids, PED and Drusen could be added. These values were computed once via a script and then added to the data set. Additionally, the average thickness of each layer could have been added. However, I decided to only add the total thickness of the retina as it was clinically more relevant and it was a much more robust measurement than the thickness of the individual layers.

## Preprocessing

Since not all annotations are always given and not every patient visit included an OCT, which could be analyzed, the data was left with several gaps for several features, which had to be addressed. In order to have data points for every feature for every patient visit, the data was interpolated using different techniques. Qualitative features such as smoking behaviour or diagnosis were interpolated using a forward and a backward fill. The merged features were not interpolated as they were already given for each visit. Numerical features such as visual acuity, volumes, IOP or BMI were interpolated using smoothed linear interpolation. Smoothed linear interpolation was computed by taking a weighted mean of a window of 90 days, whereas the computed point is in the middle of the window and the middle has the highest weight with the weights shrinking towards the edges of the window. Figure 3.6 shows the actual data points for a patient's visual acuity versus the interpolated visual acuity. By using this type of interpolation one can mitigate noisy data. As one can see in figure 3.6 part A the visual acuity data can have large amplitudes in a very short time frame. The smoothed interpolation takes the mean between multiple values and returns a value that is more coherent with the rest of the curve. Visual acuity usually does not change much in short time frames, which is why this method helps remove potential measurement mistakes. However, this

type of interpolation also damps extrema in the curve and, hence, artificially reduces the amplitude of the curve, which might introduce some errors. The interpolation was computed using pandas rolling window average.

Additionally, each feature was standardized to have zero mean and unit standard variance. No other preprocessing was applied to the data.

## Windowing

The datapoints were created by sliding a window that covers twelve visits over the patient's history. This selection of twelve visits was arbitrary and an evaluation of other window lengths could be conducted, but was not possible in the given time. One data point does not cover a fixed timeframe but a fixed number of visits, which can have varying time intervals inbetween them. However, to account for that the time since the last visit/ treatment is also a feature. Getting the target variable at exactly one, three, six, nine and twelve months was rarely possible, because visit dates might differ. Therefore, the interpolated values were sampled at those time points and used as the target variable, although there might not be an actual measurement. However, only values that are inbetween actual recorded values were used as targets. Data windows whose target would be outside of any real visit would require extrapolation from the given data which would in turn add bias to the data, which is why these windows were not added to the data set. This led to different data set sizes depending on how far in the future the target was. Naturally, this means that the data sets for targets farther in the future are smaller than those who are only a few months in the future. Table 3.4 shows the amount of data windows for each target time frame.

## Data split

The data was split into 80% training data and 20% test data, where the split was done on a patient scale and not on a data point scale. Consequently, data sets and splits vary in number of data points depending on the split and the targeted time. Furthermore, it also means that the test set does not include any data from a training patient and the model needs to generalize to unseen patterns. Altogether, the data set included the data of 1653 patients. Therefore, we have the data of 1322 patients in the training set and that of 331 patients in the test set.

### 3.7.2 Bias

The data is heavily biased since all data stems from unhealthy patients. Figure 3.7 shows the distribution of the target metrics as violin plots. One can see that for the visual acuity we have a bias towards smaller values. However, against the intuition of observing more unhealthy values the volume and number of fluids values are heavily biased towards zero. This could be an artifact of the unbalanced training data of the segmentation model, which is then propagated onto this data.

Moreover, the data is biased because of the annotations of the medical experts. The information content of all features, specifically the binary features, depends completely on the accuracy of the annotations of symptoms from the doctors. In this sense the model also learns the possible mistakes of the attending physicians. To mitigate this problem, annotators were trained to use the annotation tool and the quality of annotations was assured using inter-rater metrics.

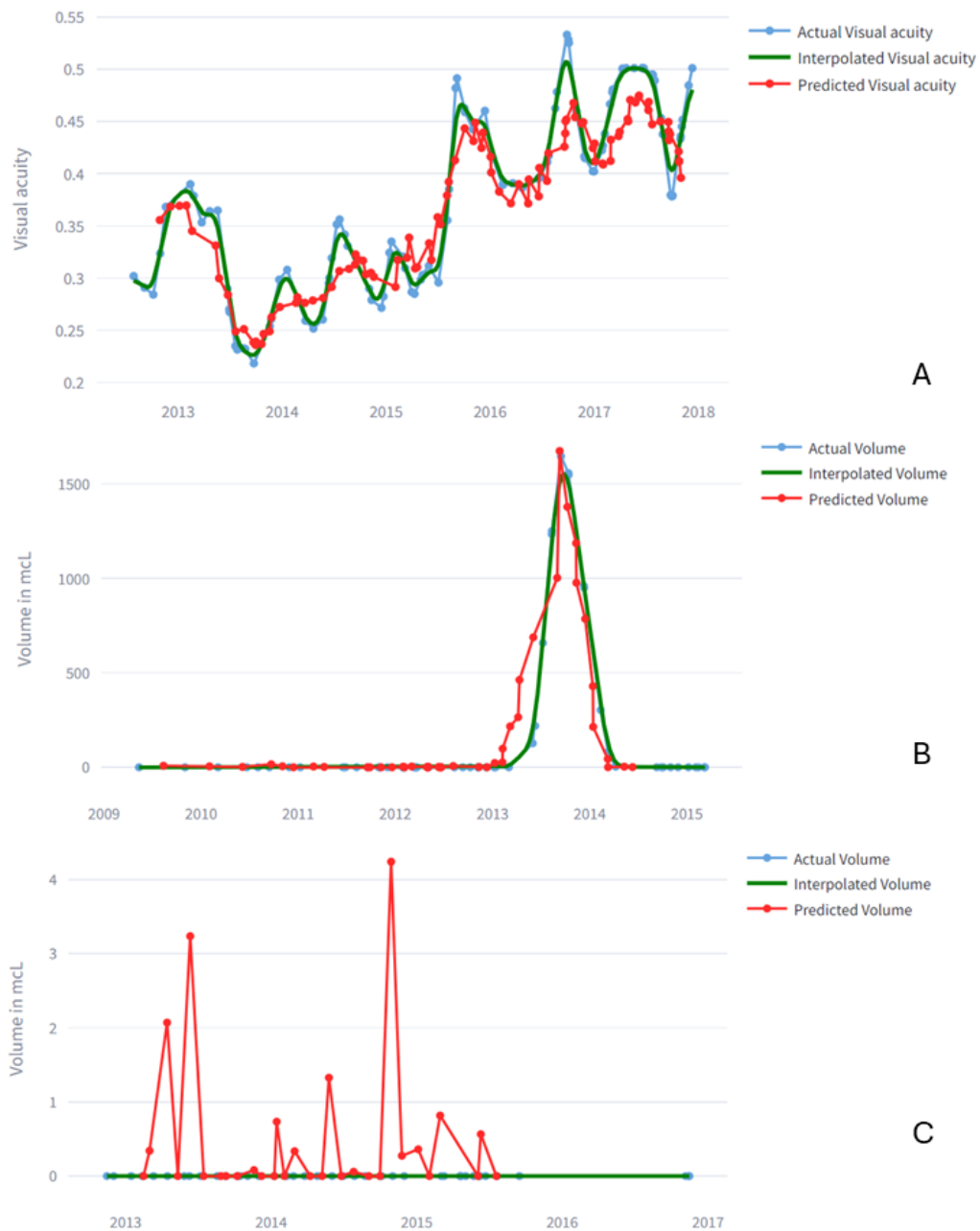


Figure 3.6: Linegraph of actual datapoints (blue) of visual acuity and volumes versus the interpolated data created by smoothed interpolation (green) versus the data predicted by the three months forecast model of the specific metric (red). A) shows the visual acuity. B) and C) show the volume of fluids. One can see that A) and B) accurately predict the development. In C) we can see that the model keeps predicting the presence of fluids, while none can actually be observed, although notably on a very small scale.

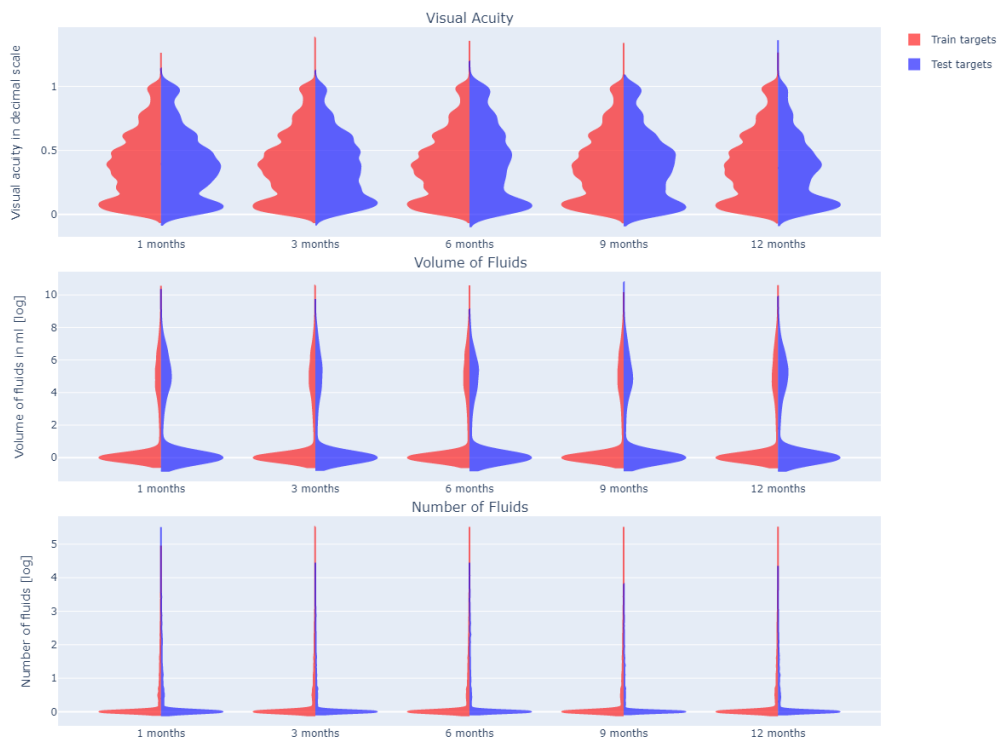


Figure 3.7: Violin plots of visual acuity (top), volume of fluids (middle) and number of fluids (bottom) values. Note the logarithmic scale for volumes and number of fluids. The violins show the distribution of data along the metric's possible values. All metrics are lower bound by zero.



Metric	Time target [months]	Number of training samples	Number of test samples
Number of fluids	1	58767	16290
	3	57594	15141
	6	55814	13327
	9	53204	12220
	12	51098	10778
Volume of fluids	1	61887	13170
	3	57922	14813
	6	56888	12253
	9	51511	13913
	12	48212	13664
Visual Acuity	1	60214	15654
	3	58706	14801
	6	56620	13275
	9	52753	13401
	12	50211	12363

Table 3.4: Target timeframes and their respective number of datapoints in the dataset.

Parameter	Type	Possible values
Learning rate	Hyperparameter	[0.01, 0.001]
Learning rate reduction factor	Hyperparameter	[0.9, 0.5]
Batch size	Hyperparameter	[64, 128, 256]
Number of LSTM layers	Architecture	[1, 2, 4]
Hidden size of LSTM layers	Architecture	[8, 16, 32]
Dropout rate inside LSTM	Architecture	[0.0, 0.1]

Table 3.5: All tuned parameters included in the hyperparameter tuning and their respective possible values.

### 3.7.3 Architecture and Hyperparameter Tuning

The architecture of the prognosis model is a simple neural network consisting of a varying number of LSTM layers depending on the metric and timeframe to predict, a normalization layer and a fully connected layer. Table 3.5 shows the hyperparameters for each combination of health metric and targeted timeframe.

Hyperparameter Tuning was done for the following health metrics: visual acuity, number and total volume of fluids. Moreover, hyperparameter tuning was only done for the timeframes of three and six months. This was unfortunately necessary in order to keep the number of models to train in practical range. Three and six months were chosen for hyperparameter tuning as it is clinically the most relevant timeframe. The models were trained for 500 epochs on the Mean Absolute Error loss. Training could stop early if the validation loss did not improve for 100 epochs. The learning rate was schedule using a reduction factor, that was multiplied to the learning rate whenever the validation loss hit a plateau. Patience before the optimiser found a plateau was five epochs. A total of 144 models was trained with a total training time of approximately 136 hours.

Metric	Time in months	Mean AE	Median AE
Number of fluids	1	0.63	0.07
	<b>3</b>	<b>0.48</b>	<b>0.03</b>
	<b>6</b>	<b>1.05</b>	<b>0.10</b>
	9	0.57	0.02
	12	0.68	0.04
Volume of fluids	1	4.49E-5	5.59E-6
	<b>3</b>	<b>5.56E-5</b>	<b>1.90E-6</b>
	<b>6</b>	<b>6.32E-5</b>	<b>1.87E-6</b>
	9	1.14E-4	4.94E-6
	12	1.12E-4	3.14E-6
Visual Acuity	1	0.02	0.02
	<b>3</b>	<b>0.04</b>	<b>0.03</b>
	<b>6</b>	<b>0.05</b>	<b>0.03</b>
	9	0.05	0.04
	12	0.05	0.05

Table 3.6: Performance of prognosis models in mean and median absolute errors. Bold numbers indicate that the model was finetuned. AE = Absolute Error

### 3.7.4 Results

The best parameters from hyperparameter tuning were used to do a smaller scale hyperparameter tuning on the remaining models. Figure 3.8 shows boxplots of the absolute error of the different predicted metrics for their various timeframes. The mean and median absolute errors can be found in table 3.6. For the visual acuity, one can see that the absolute error gets worse with increased time reaching errors of up to 0.8. However, for the shorter timeframes the error is relatively small. This increase can also be observed in the mean and median absolute errors. However, these errors are smaller or equal to 0.05 in decimal scale, which indicates a good average prediction for visual acuity. For the volumes, similar observations can be made. The errors' range increases with time. However, the mean and medians are roughly the same. In general, they seem to be extremely low, although the errors are in milliliter scale, such that the differences might still be significant on tissue scale. If one takes the data distribution bias as shown in figure 3.7 into account, then the errors can be explained by the large amount of near zero targets, which, when predicted correctly, contribute massively to a small mean error. The same can be said for the predictions of the number of fluids. The mean and median are very small, however, the distribution of data indicates that a large portion of target values lies near zero. Hence, the models might be good in predicting small, near zero values but might perform suboptimally on data, that actually contains fluids. This becomes especially clear when looking at the large amount of outliers in both metrics, which are even more extreme considering the logarithmic scale.

Three examples of the three months forecast models can be found on figure 3.6. On part A) of the figure, one can observe that the model reliably predicts the overall trend in visual acuity, although it misses the amplitudes from 2014 to 2015. The predictions of the total volume of fluids in part B) show similar results. This has high clinical relevance, as it would theoretically give a treatment indication before the patient suffers the first symptoms allowing for preventive medication. Looking at C) we can see that the predictions are not always as fitting. The patient does not have any fluids, while the model keeps predicting a rise in fluids although in almost negligible amplitude.

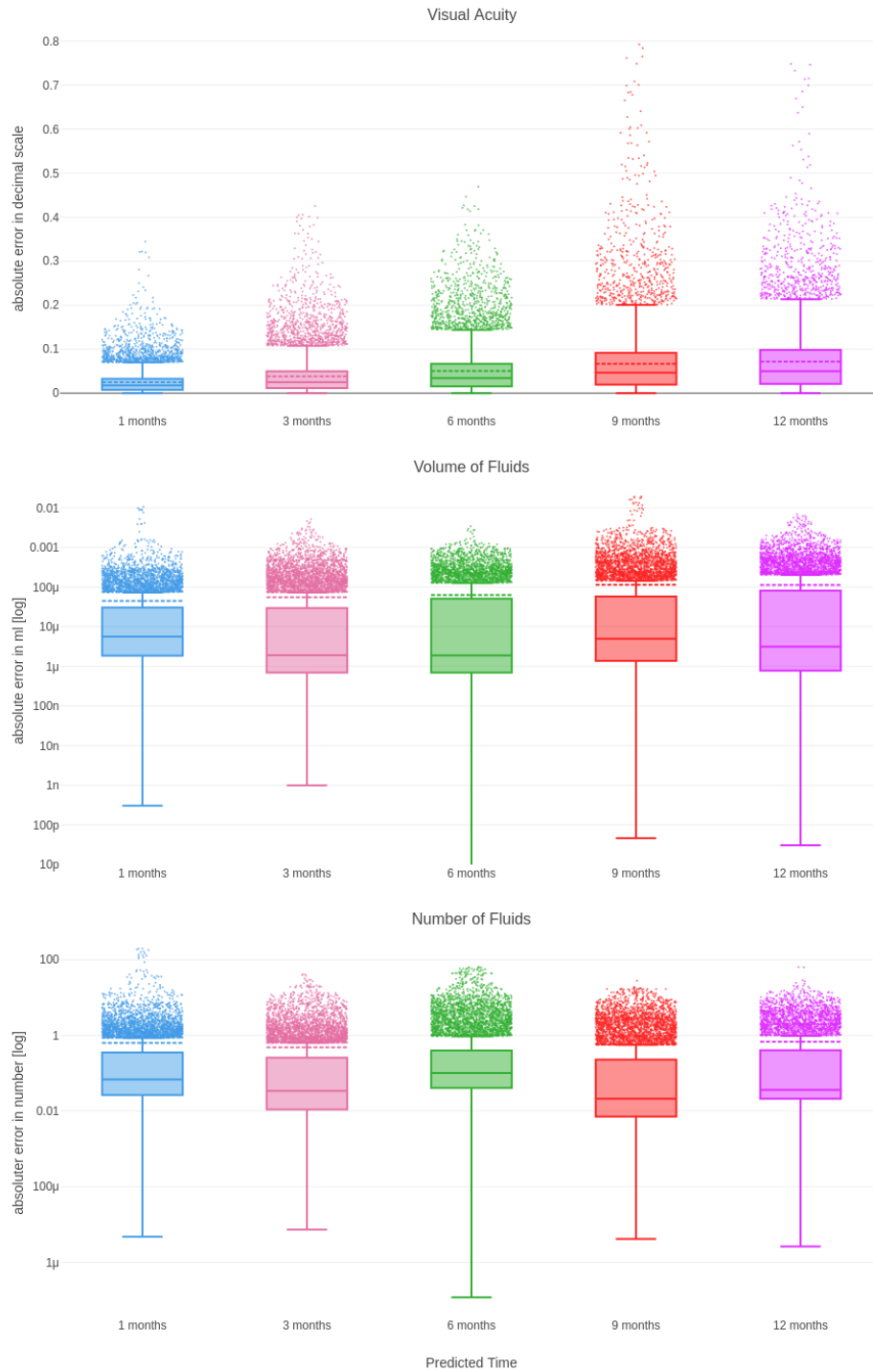


Figure 3.8: Boxplots of the absolute error of test set predictions for visual acuity (top), total volume (middle) and number of fluids (bottom). Please note the logarithmic scale on the volume and number of fluids plot. The dashed line represents the mean and the continuous line the median of absolute errors. Dots above the boxplot represent outliers. Note that the data is lower bound by 0, hence outliers will only be above the boxplot.

### 3.7.5 Explaining model predictions with SHAP values

To understand the models' predictions, SHAP values were consulted. The SHAP values were computed using the shap python package <sup>15</sup>.

#### Computation of SHAP values

An explainer model was fit to 5000 samples from the training data of each metric and the time target of three months, as this time range yielded the best performing models. The explainer model was then used to compute the SHAP values of 1000 samples from the test data. In order to cover the whole range of possible target values, target values were sorted and features were uniformly sampled using the sorted indices for both fitting the explainer and computing SHAP values. Traditionally, one uses the mean of the absolute SHAP values across the features to compute the feature importance. This means that important features on average had large absolute values across the number of points and the sequence length. However, this measure is not perfect as features, whose impact is only large for a certain part of the sequence, will have lower impact on the prediction than features, whose impact on average is intermediate. Hypothetically, medication for example might only have an impact on the prediction of fluids, if it was given in the last three visits. In these cases its impact would be high. However, its impact on the prediction from visits that are longer ago might be near zero. The average of its impact might be skewed towards a lower value because of this, while another feature has a low average impact across all points in the input sequence. This other feature will be ranked higher, although clinically the impact of the medication is more interesting. To mitigate this problem, the maximum across the means of absolute values for each sequence point was taken as a measurement of the importance of a feature. Formula 3.2 shows how the feature importance was computed, where  $S_{ijk}$  is the SHAP value for feature  $i$ , sequence point  $j$  and sample  $k$ .

$$\text{FeatureImportance}(f_i) = \max \left\{ \frac{\sum_{k=0}^{n_{\text{Samples}}} |S_{ijk}|}{n_{\text{Samples}}} \mid \forall j : 0 \leq j < n_{\text{Sequences}} \right\} \quad (3.2)$$

#### Feature importance for visual acuity prediction

Figure 3.9 shows the computed SHAP values of the visual acuity model for the three months forecast. One can see that to predict visual acuity, the past visual acuity values are most impactful. Specifically, the visual acuity of the second most recent visit has high impact on the prediction, while influence becomes less large the longer ago the visits were. Surprisingly, the most recent visit does not have the largest impact on the model. Large visual acuity values positively influence the prediction, while low values negatively influence it. This means that high visual acuity values also lead to high visual acuity predictions, when looking at this feature alone. The second most impactful feature is the number of days since first treatment. Looking at the SHAP values of the most recent visit one can observe that the longer the first treatment is in the past, the more it negatively affects the visual acuity prediction. This pattern can be observed for the most recent three visits. However, then the pattern switches meaning that visits that lie longer in the past. In figure 3.9 blue color in the medication variable is associated with

<sup>15</sup><https://github.com/shap/shap>

no medication, whereas all other colors can be associated with some active ingredient. As expected, medication has a positive SHAP value meaning, that, if there was treatment before, the model predicts higher visual acuity values, while no medication is associated with negative to no impact on the prediction. Again, one can observe that the absolute influence of the second most recent visit has the highest contribution to the prediction, while the first and latter ones have less impact. Moreover, the absolute SHAP values of the medication feature and the other features are much smaller than those of the visual acuity feature speaking for a smaller relevance to the prediction. Aneurysms as well as PEDs seem to have a positive impact meaning that their existence influences the model to make a better prediction for the visual acuity. This seems counterintuitive at first, however, it could be correlated to extensive treatment or surgical procedure such as photocoagulation, which in turn improve visual acuity. Atrophy generally has negative impact when present, which naturally makes sense as a thinning of the retina comes with a worsening of the visus.

### Feature importance for fluid prediction

Figure 3.10 shows the SHAP values of the model that predicts total volumes of fluids. Again, one can see that the metric to be predicted is also the most important input feature and its impact on the prediction scales proportionally with its values. However, contrary to the visual acuity prediction very low values do not majorly contribute negatively but rather have very few influence, while very high values still have large positive impact. The same holds for the number of fluids' and the number of PED's impact on the model's prediction. Studies have found that PEDs are associated with CNV, which again is associated with the development of fluids [92]. Hence, these SHAP values clinically make sense. Furthermore, visual acuity has the same effect on the volume of fluids prediction as it has on the visual acuity prediction, although notably smaller. Naturally, this does not make sense, as high visual acuity values should not correlate with high fluid volumes. If one takes into consideration, that patients usually only seek medical advice, once they develop first symptoms, it becomes apparant, where this correlation is rooted. Patients that just started treatment, generally still have relatively high visual acuity, but are in therapy, because of early symptomes such as fluids. Additionally, towards the later stages of AMD and DR, the retina is left with scar tissue and fluids are less present. Therefore, longlasting patients with low visual acuity values also develop less fluids. Finally, high age positively impacts the development of fluids, which is also in line with clinical knowledge. This shows that the model's forecast is clinically comprehensible and, hence, learned to extract meaningful information for the fluid prediction.

## 3.8 Treatment recommendation

The CDSS aids medical professionals by providing visual analytics and presenting the most crucial information. However, it could be further enhanced by supporting the treatment decision process, given the clear guidelines. Specifically, if the OCT is active (meaning fluids are present), treatment is required; if not, it is unnecessary. Additionally, treatments should follow a series of three IVOMs at monthly intervals before a break. Furthermore, Guidelines determine the length of breaks and when to resume treatment.

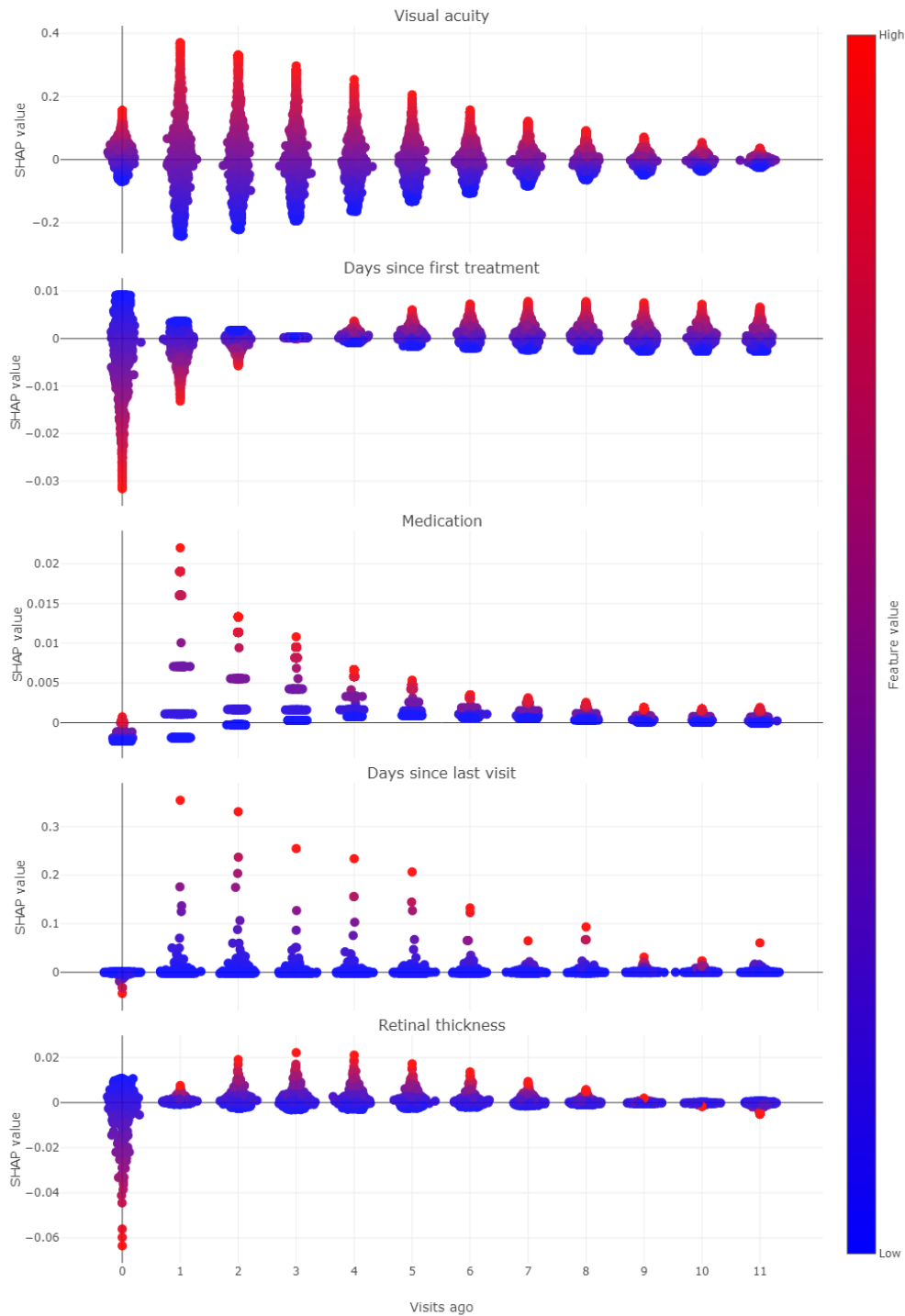


Figure 3.9: SHAP values of the impact of the five most impactful features on the three month prediction model of visual acuity. Each datapoint shows the SHAP value of one single prediction. Positive SHAP values impacted the predicted value positively, while negative values impacted it negatively. The x axis shows how many visits ago that datapoint lies in the sequence. Red dots indicate high feature values, while blue dots indicate low feature values. For example: High visual acuity values on the second most recent visits indicate a positive impact on the prediction, whereas low values indicate a negative impact. The more visits ago, the less influence they have on the prediction except for the first visit, which has smaller impact.

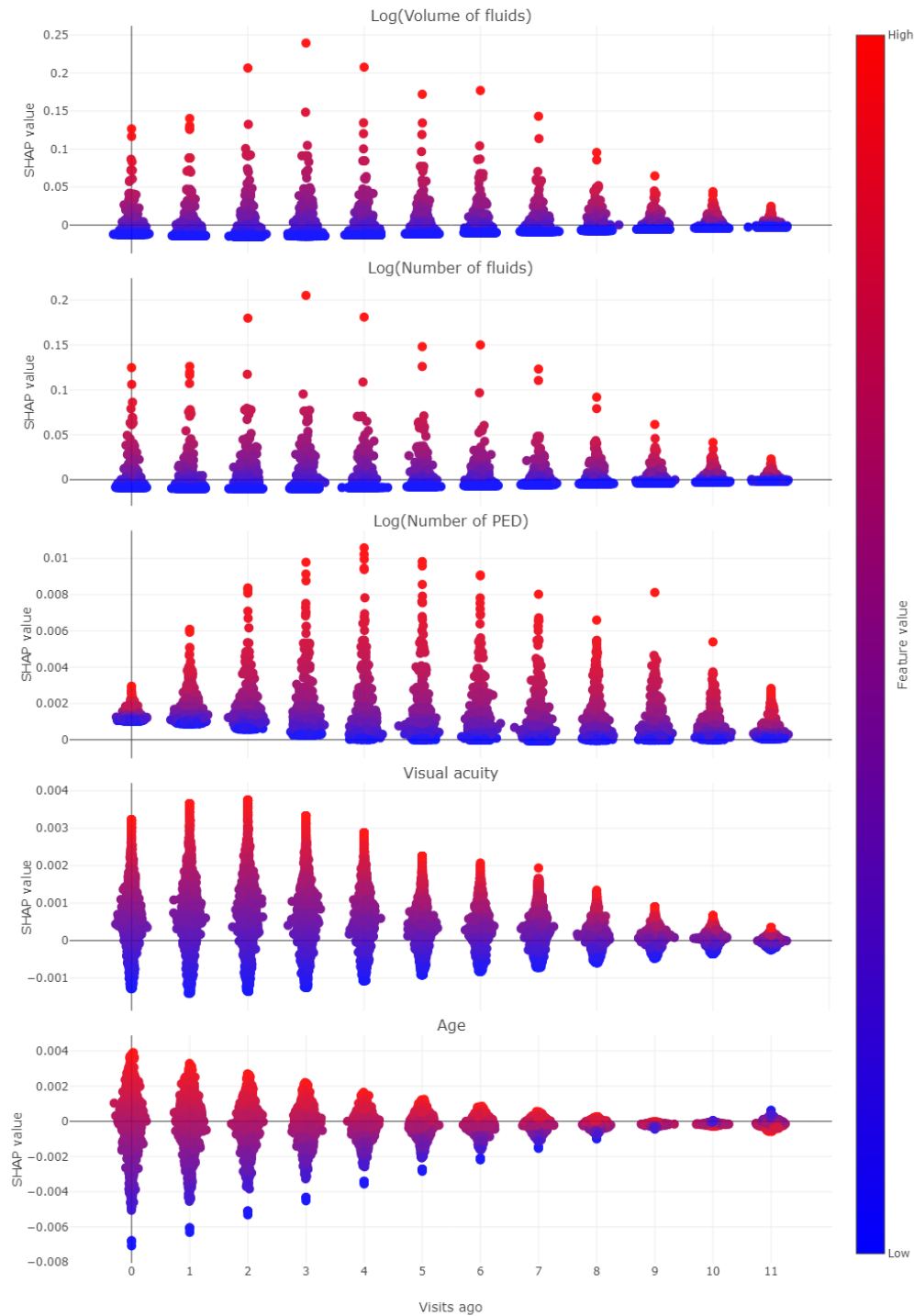


Figure 3.10: SHAP values of the impact of the five most impactful features on the three month prediction model of total volume of fluids. Each datapoint shows the SHAP value of one single prediction. Positive SHAP values impacted the predicted value positively, while negative values impacted it negatively. The x axis shows how many visits ago that datapoint lies in the sequence. Red dots indicate high feature values, while blue dots indicate low feature values. For example: High volume values on the most recent visits indicate a positive impact on the prediction, whereas low values indicate a negative impact. The more visits ago, the less influence they have on the prediction.

### 3.8.1 Algorithm

A treatment recommendation system was developed, which can be seen in figure 3.11. First, abort conditions will be checked such as extreme IOP or extremely low visual acuity values. These values indicate that treatment is too dangerous for its potential benefit. Hence, if they are fulfilled, the medication should be aborted. Otherwise, treatment status will be checked. If an IVOM series has already been started but not finished, meaning we are on the second or third injection for this cycle, we move on to the decision for which medication to use. If the patient is not in an unfinished IVOM series, then the segmentation model is used to identify, whether the patient's recent OCT is active. If that is not the case, we do not start another treatment round. However, if it is still active the prognosis model is consulted to check for the best medication for this specific patient.

This tool predicts three different metrics for a total timeframe of twelve months. For the recommendation, I only consulted the visual acuity metric as one wants to improve this for the patient as it directly measures their vision. The prognosis model will be fed the data of the current visit and the last eleven visits. It will give a prediction for each medication by switching this feature in the current visit for every possible Anti-VEGF drug in our database. Hence, we get a prediction for the development of visual acuity for every possible treatment on that visit date. With these predictions we can decide, whether we want to continue treating with the same medication as before or if there might be a more effective option. However, since the model's accuracy deteriorates with increased time to predict, we cannot weigh every prediction equally for comparison. Hence, the algorithm looks at a weighted average of the prognosis of one, three, six, nine and twelve months, whereas the weights are 20%, 35%, 15%, 10% and 5%, respectively. A switch to new medication will only be recommended, when the new medication promises at least 10% improvement over the old medication. Otherwise, the system recommends to continue using the already given medication.

The treatment recommendation system follows the PRN scheme, as this was predominantly used in our database. Moreover, identifying the TAE scheme proved to be a hard task, because series are not finished, have unregular breaks or other deviations from the treatment plan. Hence, the system always recommends a cycle when one is needed without analyzing previous cycles, their intervals and break durations, although this information is contained inside the prognosis data to some extent.

### 3.8.2 Evaluation & Results

The recommendation system was evaluated by comparing its treatment suggestions with historical decisions. A random subset of test samples from the preprocessed prognosis data was used, and the recommendation model was fed these samples. For each active ingredient, the goal was to obtain 500 sample points where the drug had been administered. However, due to varying frequencies in the data, only Aflibercept, Ranibizumab, and Bevacizumab met this criterion, resulting in a sample size of 2,179 visits where IVOM medication was administered. To balance the test data, an equal number of visits without Anti-VEGF therapy were included summing up to a total of 4,358 data points. The data was then classified into two ways: treatment vs. no treatment and different medications including no medication.

Figure 3.12 shows the confusion matrix for the treatment classification. The system's recommendations matched the historical data in only 60% of cases. It almost equally recommended medication where historically none was given, and no medication where



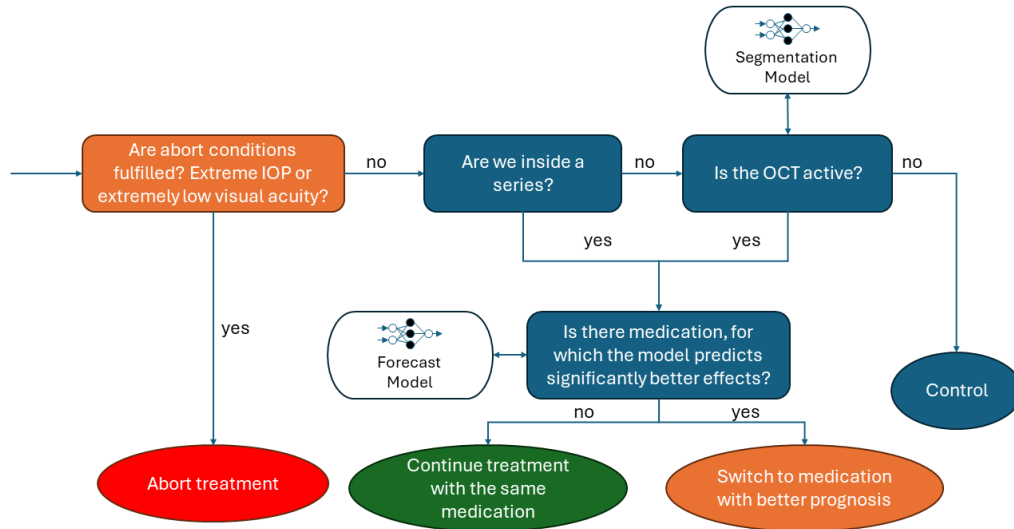


Figure 3.11: Scheme of the recommendation algorithm.

historically IVOMs were administered. Several factors could contribute to these errors. Firstly, the data is incomplete, lacking critical information such as patient needs or surgical procedures. Secondly, the data is biased due to quantification algorithms and ophthalmologists' annotations. Incomplete or outdated annotations can mislead the system. Unlike DL and ML algorithms that can handle data outliers by learning patterns, the recommendation system struggles with outliers and cannot identify these situations accurately. For instance, if a doctor previously annotated neovascularization as an indication for treatment but didn't update this as resolved in subsequent visits, the system would still recommend medication. Finally, the historical data itself may not always reflect the correct indication due to deviations from guidelines, human error, non-clinical reasons for stopping treatment, or other concurrent diseases. Thus, the historical data is not a perfect standard for evaluating the model's performance. In this sense, the model does not attempt to reproduce historical data as a DL or ML model would. Instead, it theoretically adheres to standard guidelines, which may differ from real-world practices. Therefore, despite these discrepancies, its recommendations remain clinically relevant to some extent.

When the system recommends therapy, it suggests the same drug as historically administered in 85% of cases. This is likely because medication switches are rare and IVOMs are typically given in series, making drug recommendations straightforward. Once a series is initiated, it is usually continued until completion, and new series are often started with the same medication, leading to a high consistency rate in drug recommendations.

### 3.9 Visual Components

The dashboard is separated into six visual components (VC), which can be seen in figure 3.13: The top bar, the OCT viewer, the linegraphs, the metrics, the recommendation and the infobox. The dashboard was implemented using streamlit [40]. Streamlit is a

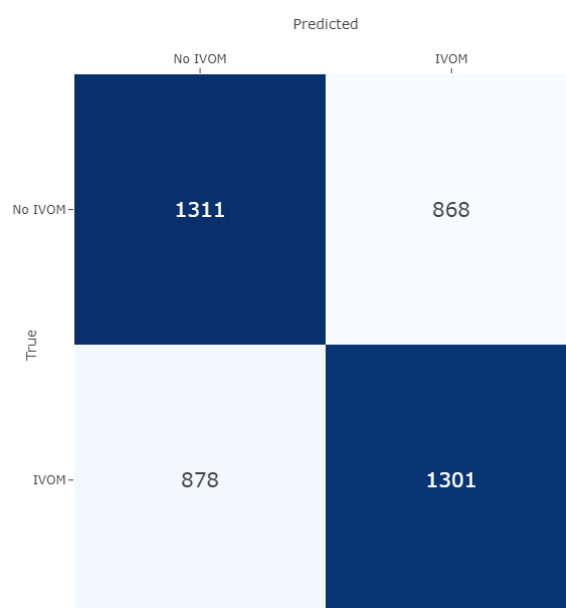


Figure 3.12: Confusion matrix of the binary classification of treatment versus no treatment.

python framework that facilitates the rapid development of interactive web applications. It allows for easy integration of data visualizations and user input functionalities. It also offers a wide range of publicly available, community created extensions for all kinds of functionalities. Hence, it was selected for the development of a prototypical CDSS. For visualizations Plotly was used [39]. Plotly is a python package for easy creation of high quality, interactive graphs. The integration of plotly graphs into streamlit made it especially useful.

### 3.9.1 Top bar

The top bar was designed to deliver the most crucial metadata about the patient in a compact fashion. Hence, it displays the age, gender, weight, height, body mass index (BMI), smoking behaviour as well as the blood pressure, if they are available in our database. Additionally, it highlights which disease the patient is being in treatment for.

Moreover, it showcases a treatment status. This treatment status informs about whether a patient is in a series of IVOM uploads or not and what number of IVOMs was already administered. Additionally, the active ingredient of the current series is shown. However, there is no information about the treatment status or the applied upload scheme inside the database and one has to apply an algorithm to the data in order to retrieve it. The algorithm looks at the last visits and whether or not an IVOM was given and then counts the previous visits until one visit falls out of the pattern of one month intervals inbetween visits. The scheme, PRN or TAE, could not be computed in the same way, which is mainly due to lots of deviations in the actual IVOM uploads. As the actual uploads do not strictly follow the scheme for unknown reasons and patients often do not come for extended periods, it was impossible to reliably compute the scheme.

To mitigate this problem the IVOM timeline was added, which shows a colored bar for

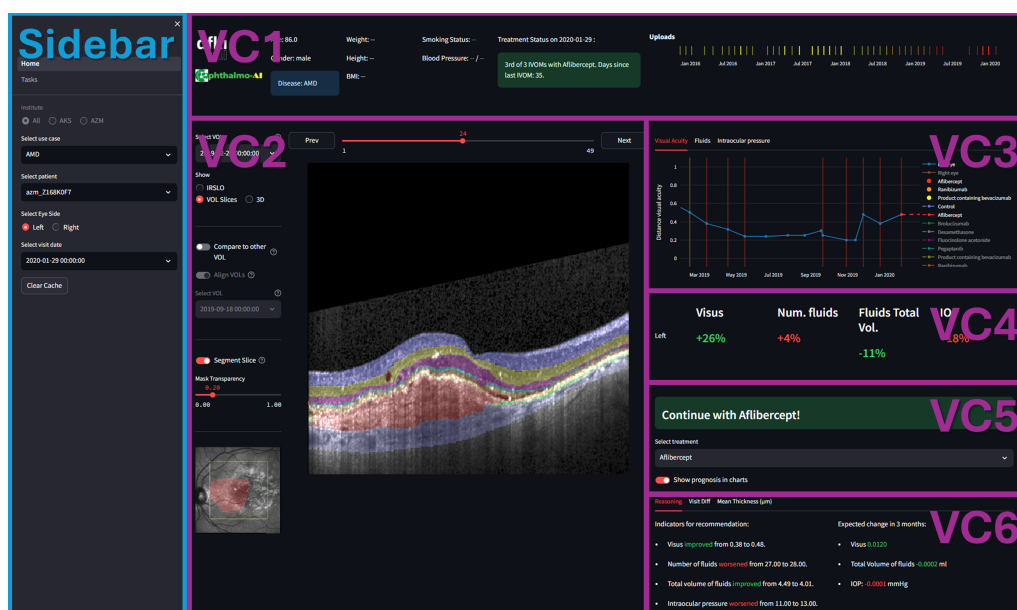


Figure 3.13: The six visual components (VC) of the dashboard: VC1 = Top bar, VC2 = OCT viewer, VC3 = History graphs, VC4 = Metrics, VC5 = Recommendation, VC6 = Infobox. Moreover, the sidebar can be seen, which yields functionality about patient selection.

every visit where an IVOM was given. The color of the bar represents one of the possible medications and is coherent with the history graphs in VC3 (see below), which also feature colored bars for IVOMs. In the IVOM timeline the medical expert can visually identify schemes and patterns. The timeline was implemented using plotly's express subpackage, which contains a timeline plot functionality.

### 3.9.2 OCT Viewer

The OCT viewer offers functionality to look at OCT slices, the IRSLO and the 3D reconstruction. It features a wide range of tools, which are discussed in the following section. In general, all features are divided into three spaces: The tool bar, an IRSLO overview and the plot area. The tool bar offers functionality to select, segment or compare the current OCT, while the IRSLO overview shows a small version of the IRSLO for the user to orient on. The plot area shows the different image information: IRSLO, Slice and 3D view. The segment and compare functions are shared across the IRSLO and Slice view. Segment shows the segmentation, while compare overlays two images. A slider can be used to move the overlap boundary in compare mode. An align function can be used to align the images in the compare slider, as explained in section 3.4. If the alignment is not good, a warning will be displayed.

## IRSLO

The IRSLO<sup>16</sup> is an "En face" representation of the OCT meaning that you see onto the retina. The IRSLO is automatically selected at the start, as it offers a quick overview over the retina. The IRSLO to be shown can be selected through a dropdown menu. The tools for the IRSLO include: Compare mode, Segmentation and a Thickness Mapping.

The compare mode will overlay two IRSLO images with a slider, which can be moved. Moving the slider means moving the boundary between the overlapping images. This allows for easy comparison between two images. The IRSLO for comparison can also be selected through a dropdown menu. Additionally, one can align the IRSLO images with the alignment algorithm described in section 3.4. Moreover, the segmentation mode gives a top down view of the lesion segmentations on top of the IRSLO. Furthermore, it shows a yellow rectangle indicating the position of the OCT recordings. A green line inside the rectangle highlights the currently selected slice, such that users can see, where this slice is located in the IRSLO. Figure 3.14 shows the IRSLO view with all of the aforementioned features selected. One can see the purple colored PED segmentations and compare them to a previous recording using the slider.

Additionally, the IRSLO view offers a thickness mapping, where the thickness of a selected retinal layer is overlayed over the IRSLO as a heatmap. Figure 3.14 shows this mode with the comparison to a previous visit. If the comparison tool is turned on, the heatmap does not show the thickness of that layer but how that layer changed from last visit to the current one. Blue colors indicate that the layer got thinner, while red colors indicate it got thicker. Additionally, the mean change is shown below the image. When hovering over any pixel on the heatmap, a hoverinfo will appear and show the thickness of that layer at that specific pixel.

## OCT Slices

The user can also compare the OCT slices using compare and the slider and show the segmentation mask of the segmentation model. Figure 3.16 shows the OCT view when both features are enabled. The segmentation model also segments both layers and lesions and not just the lesions unlike the IRSLO. The user can navigate through the different OCT slices using either the buttons on top or using the slider inbetween the buttons. When alignment is enabled but the selected OCTs are not follow ups, a warning will be displayed highlighting that alignment is experimental with this selection of OCTs.

## 3D Graph

The 3D graph makes use of the 3D reconstruction explained in section 3.6. It is supposed to give a "At a glance" overview over the whole OCT including layers and lesions, thus rendering scrolling through slices unnecessary and speeding up analysis of the OCT. Figure 3.17 shows the 3D graph inside the dashboard. One can see multiple transparent and colored layers. Additionally, lesion objects are shown as 3D volumes with less transparency. Moreover, the selected OCT slice is also displayed inside its position in the 3D graph. This was done such that the medical expert has an opportunity to control the reconstruction. Unfortunately, for performance reason no comparison could be implemented, that would allow to see two 3D graphs of different OCTs beside each other.

<sup>16</sup>Infra-Red Scanning Laser Ophthalmoscopy

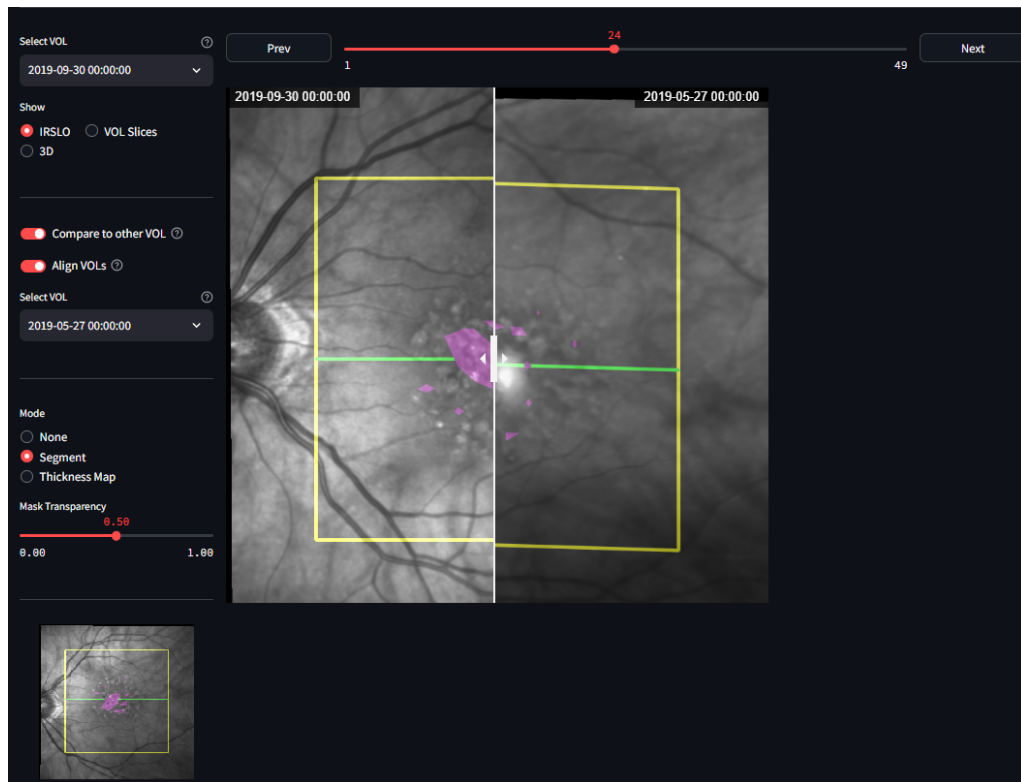


Figure 3.14: Example of VC2's IRSLO view with comparison and segmentation tool turned on. The slider on the image can be moved to make the right IRSLO overlap more or less over the left IRSLO. A segmentation mask is overlayed on top of the IRSLO. It shows only the segmentations of drusen, PEDs and fluids. Its transparency can be controlled by the transparency slider in the bottom right. Segmentation, the comparison slider and the alignment can be turned on and off. The left side IRSLO can be selected via the dropdown at the top and the right side IRSLO can be selected via the dropdown at the middle of the toolbar.

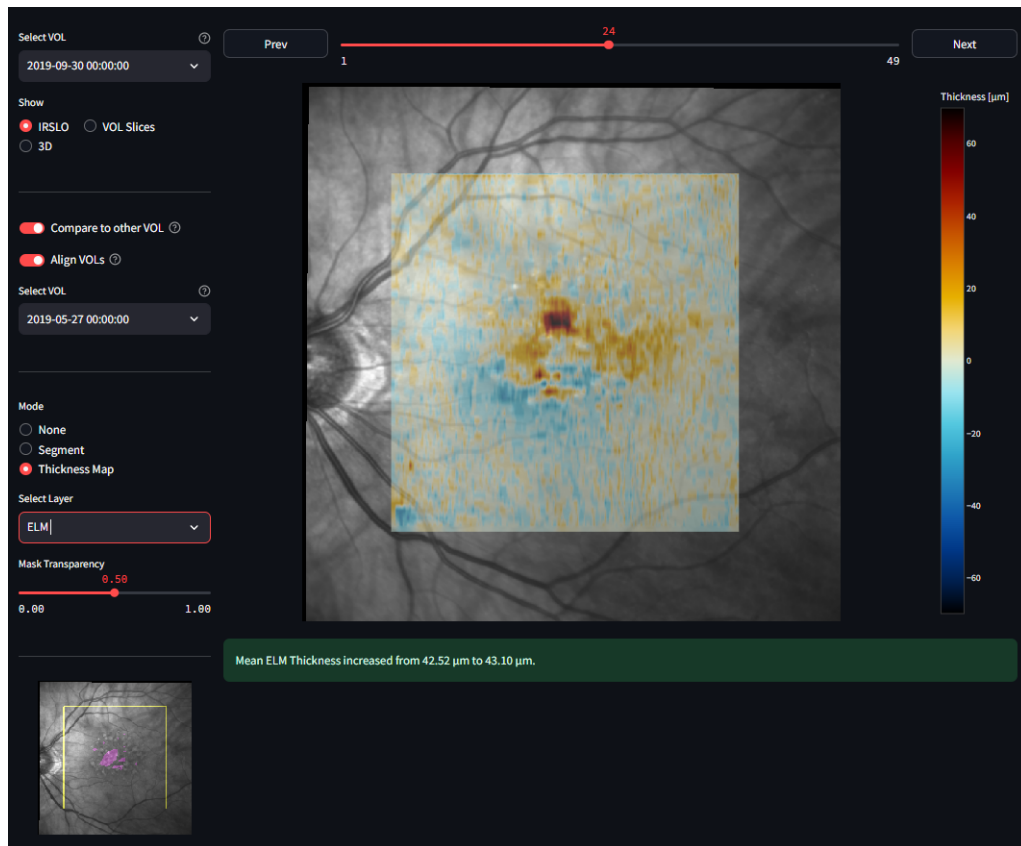


Figure 3.15: Example of VC2's IRSLO view with comparison and thickness map tool turned on. The heatmap shows the differences in thickness of the ELM layer between the two OCTs. Turning off compare will only show the thickness map of that layer from the main OCT. The "Align" button aligns the heatmaps according to the alignment of the underlying IRSLO images. The layer to be displayed can be selected through a dropdown menu at the bottom of the toolbar. Additionally, one can change the transparency of the heatmap through a slider. The mean thickness change of the selected layer will also be shown in a color coded info box below the plot. Green colors indicate thickening, while red colors indicate thinning of retinal layers.

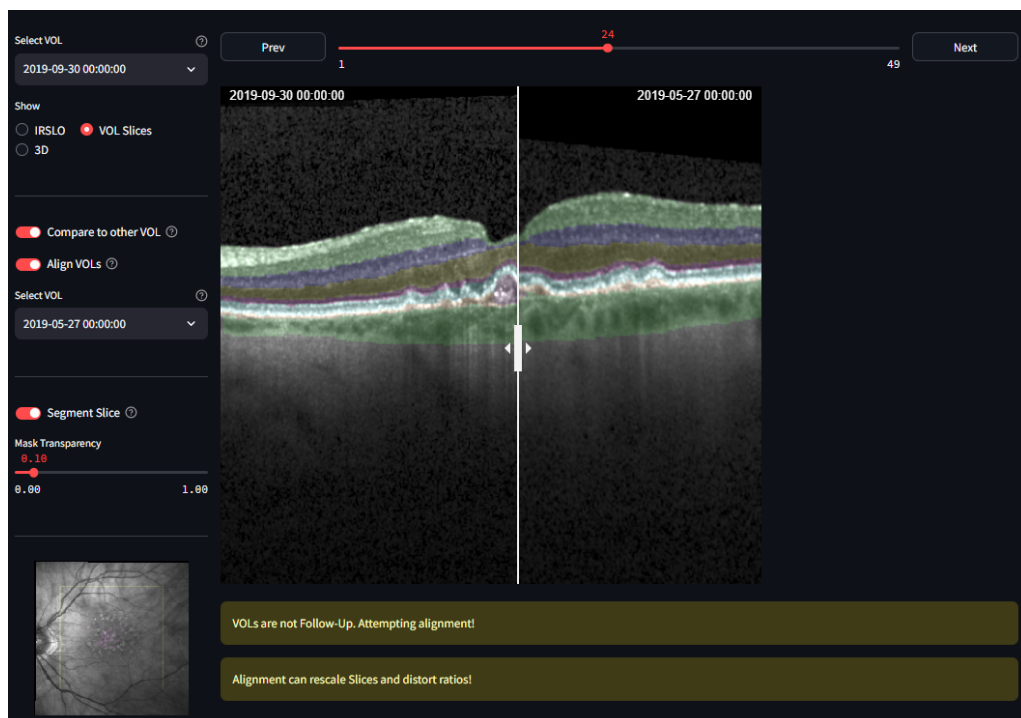


Figure 3.16: Example of VC2's Slice view with comparison and segmentation tool turned on. The slider on the image can be moved to make the right OCT overlap more or less over the left OCT. A segmentation mask is overlayed on top of the OCT. Its transparency can be controlled by the transparency slider in the bottom right.

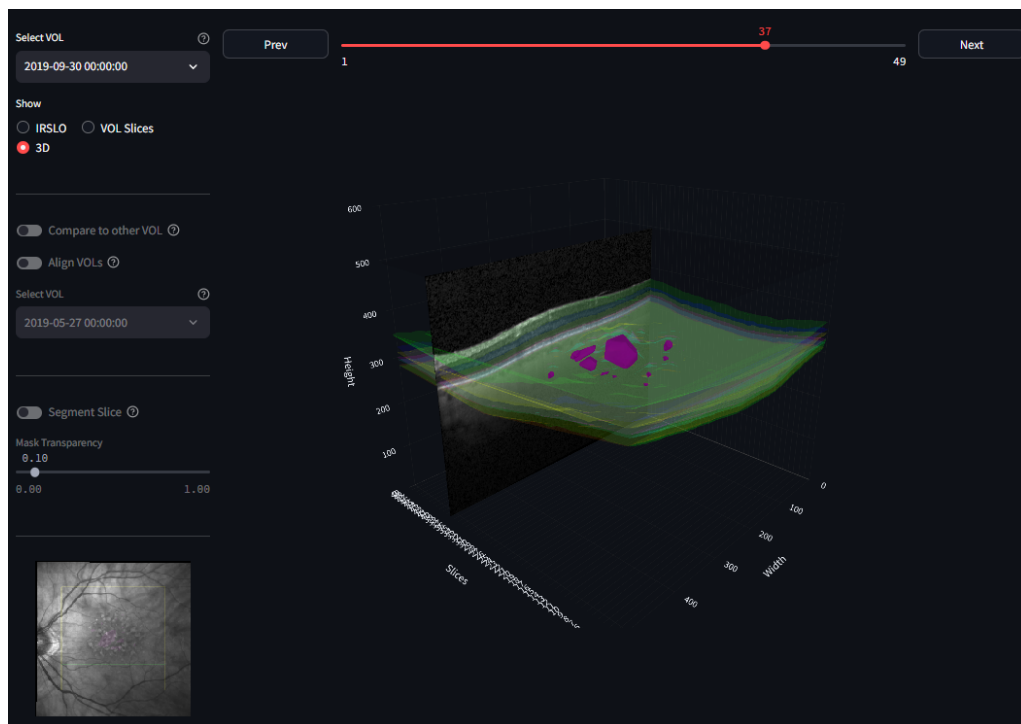


Figure 3.17: Example of VC2's 3D graph feature. The graph shows the top side of the segmented retinal layers as transparent layers colored by the same color scheme as the segmentation. Fluids, drusen and PEDs are shown as volume objects colored in the same scheme. Through the slider on top, one can move the OCT slice inside the 3D graph. The 3D graph can be rotated and zoomed using the mouse.



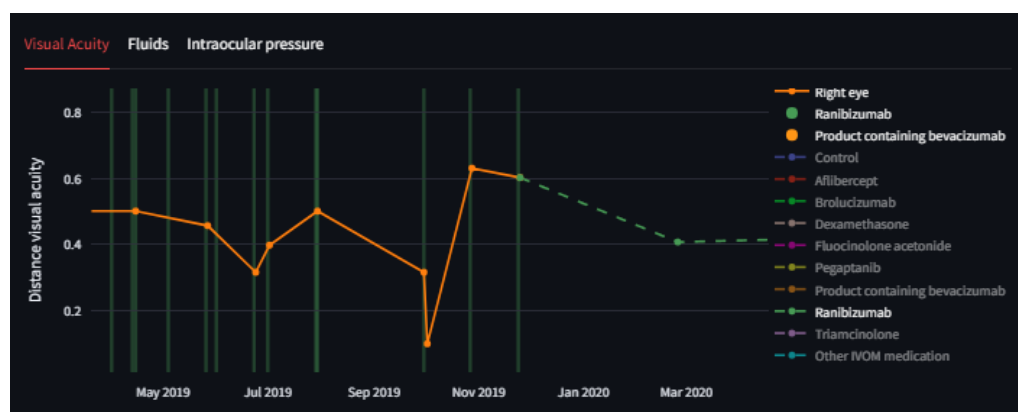


Figure 3.18: Example of VC3. The visual acuity history graph of the right eye for a randomly selected AMD patient. The orange markers are the recordings of visual acuities connected by the orange line. They are not interpolated in this graph. The green vertical lines indicate treatment with the medication ranibizumab shown in the legend. Moreover, this patient has been treated with bevacizumab according to the legend. These treatments are out of scope and can not be seen. By panning the user can make these treatments visible. The green dashed line shows the expected change when continuing treatment with this medicament. The transparent lines in the legend can be clicked to turn their prognosis visible in the graph.

### 3.9.3 History graphs

Traditionally, the medical experts have to search for metrics such as the visual acuity and IOP in a patient's electronic records manually. More specifically, they have to scroll through the records and find these values and keep them in mind, when making a treatment decision. An intuitive way of summarizing these values was plotting them in a line graph, which can be seen in figure 3.18. This line graph plots the recordings of visual acuity as markers and connects them via a line. Visits that involved the injection of a Anti-VEGF are marked as vertical lines colored in a specific scheme. The legend shows which medication has been given to the patient as dots. Moreover, the dashed line shows the prognosis of the recommended treatment. Other prognosis can additionally be shown through clicking the respective transparent name in the legend. This allows for easy comparison between prognosis. Moreover, the graph can be panned and zoomed out in order to see the whole development of this value. By selecting one of the tabs on top the user can switch between the graphs for visual acuity, fluids and intraocular pressure.

### 3.9.4 Metrics

The metrics are supposed to give an "at a glance" development view from last to current visit. They show the development of visual acuity, number and volume of fluids and the IOP in percent. Figure 3.19 shows an example of the metrics for a randomly selected AMD patient. The development is computed by dividing the current value by the last value and subtracting 100%. For cases, where the last value was zero and a Division-ByZero exception would be thrown or the last value was nearing zero and would cause the percentage to become extremely large, the current value was simply taken as the

	Visus	Num. fluids	Fluids Total Vol.	IOP
Right	-5%	-100%	-100%	+7%

Figure 3.19: Example of VC4. The metrics show the difference between current and last visit in percent.

percentage. In general, this was only a problem for the volume of fluids, as its values fluctuate greatly. The metrics were color coded in green to highlight good or healthy changes and in red to highlight bad or unhealthy changes. They were displayed in grey, if there was no change.

### 3.9.5 Recommendation

VC5 shows the result of the recommendation algorithm described in section 3.8. It was color coded to highlight the importance of the recommendation. Figure 3.20 shows four examples of possible recommendations. A) shows a recommendation to abort treatment as the visual acuity values are too low and treatment bears unnecessary risks without possibility for much improvement. B) shows a recommendation to treat with an already used medication. However, it also recommends to switch to a different drug, as it promises better results. C) shows a simple recommendation to continue an IVOM series with the same medication. D) recommends to not treat, as there are no fluids detected on the OCT. The colors of the recommendation are similar to that of a traffic light. Red symbolizes that one needs to stop treatment, while green symbolizes to continue or start treatment. Since the absence of fluids does not necessarily mean that treatment must be stopped, it is colored in blue.

### 3.9.6 Infobox

The infobox (VC6) features three tabs, each providing additional, non-critical information. Although this data is not essential for treatment decisions, it is included for completeness without occupying major screen space.

#### Reasoning

The reasoning tab shows a data-centric explanation for the recommendation. More specifically, it highlights, how critical metrics have changed from the last to the current visit. Additionally, it shows, how the system expects these values to change given a specific treatment. This treatment can be select via a dropdown menu. Figure 3.21 shows an example of the reasoning. Again, color coding was used to highlight healthy and unhealthy changes.

#### Visit Diff

The "Visit diff" tab shows additional differences between the last and the current visit in the form of a table. More specifically, it shows the preprocessed feature values discussed

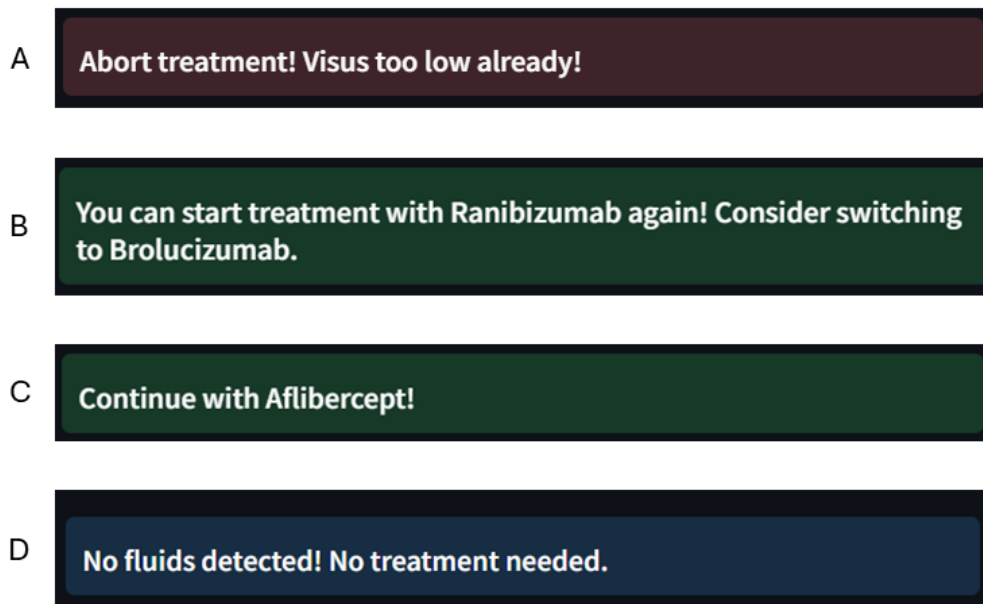


Figure 3.20: Four examples of VC5. A: Recommendation to abort is highlighted in red. B: Recommendation to switch medication. C: Recommendation to continue started series. D: Recommendation to not treat as no fluids are present.

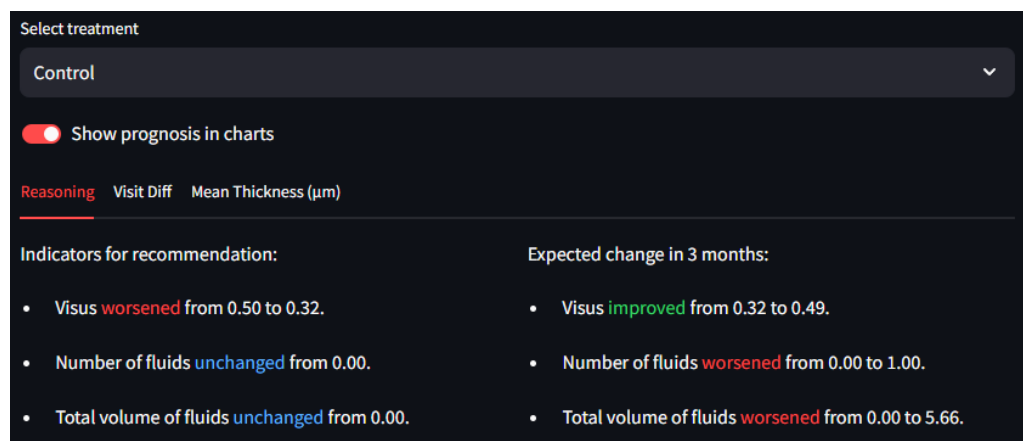


Figure 3.21: Example of VC6' "Reasoning" feature. The reasoning tab shows, how critical metrics changed from the last to the current visit and how the system expects these values to change in the future. Through the "Select treatment" dropdown menu the user can select, for which treatment they want to see the expected values.

Reasoning	Visit Diff	Mean Thickness ( $\mu\text{m}$ )	
	2015-03-10 00:00:00	Changes	2019-08-28 00:00:00
edema	Yes	unchanged	Yes
bleeding	Yes new	unchanged	Yes new
inflammation		emerged	Yes
scar	Yes   Macula	increased	Yes   Macula   Optic disc
atrophy	Retina	increased	Macula   Retina
drusen	Yes	unchanged	Yes

Figure 3.22: Example of VC6' "Visit Diff" feature. The "Visit Diff" tab shows the difference in annotations of the last and current visit and how they changed. The differences are highlighted through color coding and phrasing.

in section 3.7.1, that are not already displayed on the dashboard. However, it only shows those annotations that are present. Additionally, it highlights healthy and unhealthy changes through color coding and phrasing. Figure 3.22 shows an example of this tab. One can see that emerged drusen are highlighted in red, while the decrease in volume of PED is highlighted in green. Unchanged annotations are not highlighted.

### Mean Thickness Table

The last tab of the infobox shows a table of the mean thicknesses of the segmented retinal layers, their total thickness and how these changed. Again, color coding was used to highlight healthy and unhealthy changes, where retinal layer thinning is generally considered unhealthy and thickening healthy. The mean thickness tab can be seen in figure 3.23.

Reasoning	Visit Diff	Mean Thickness ( $\mu\text{m}$ )	
	2021-08-31 00:00:00	Change	2022-02-06 00:00:00
IPL	88.3	0.2	88.6
OPL	49.8	0.7	50.6
ELM	42.5	0.6	43.1
EZ	19.5	-0.2	19.3
RPE	37.1	-0.2	36.9
BM	10.5	-0.3	10.3
Choroidea	99.2	6.1	105.3
Total	346.933492	7.088953	354.022445

Figure 3.23: Example of VC6' Mean Thickness feature. The Mean Thickness feature shows a table of the development of retinal layers from the last to the current visit as well as by how much they increased or decreased.

---

## Chapter 4

### User Study

In this chapter, the details and results of the conducted qualitative user study will be discussed. At first, the preliminary workflow assessment user study will be presented and evaluated. Finally, the main study will be explained and its results will be demonstrated.

#### 4.1 Preliminary Workflow Study

The first step in the development of the dashboard was a preliminary workflow study involving one assistant doctor and a manager. The work place was investigated and the doctor showed their usual workflow. This study was a one hour long interview session conducted at the eye clinic in Sulzbach. It took place in one of the examination rooms, which is where the medical experts analyze patient data and decide on new treatment before welcoming the patient for their visit. Since this was merely one interview, the following part summarizes the findings.

##### 4.1.1 Findings

The work station consists of a large desk, a computer and two 27 inch monitors indicating a lot of screen space. The workflow comprises two phases: The preparation phase and the patient visit. However, it's during the preparation phase that medical experts heavily rely on their systems. Consequently, this phase is most important for identifying and improving weaknesses of the standard system. The preparation phase covers three main tasks.

In the first Task (T1), the medical expert analyzes metadata like age, gender, BMI and other risk factors from the patient records. Additionally, they check, which diseases the patient is being treated for: AMD or DR. As a next step the patient's history will be analyzed: How have metrics like visual acuity changed over time? How many and which IVOMs have already been administered? How did medication influence visual acuity? Recovering this information from the patient's history was labelled as tedious. The expert criticed that having to scroll through a medical file and finding the values manually

requires an intensive mental effort. Furthermore, it bears the risk of the ophthalmologist forgetting or overlooking crucial information consequently leading to a less informed treatment decision.

The second phase (T2) was valued as the most important for the treatment decision and consists of analyzing the most recent OCT and comparing it to older ones. With known patients, the expert does not look thoroughly through the history of OCTs but just quickly scrolls through the most recent one. If they detect fluids, they order a new cycle of IVOMs. However, with unknown patients it is necessary to get a holistic view over the development of fluids. This can only be done by viewing and comparing multiple OCTs. Whereas follow ups can be compared in a specifically designed program, other OCTs must be compared by opening two windows. While the first variant was seen as advantageous, the second was again labelled as tedious. Moreover, there was no mention of other programs or features for segmentation, 3D reconstruction, thickness maps or quantifications.

For the third and last task (T3) the ophthalmologist needs to decide on the treatment by combining guidelines and the extracted information from the first two tasks. There was no direct complaint about this task. However, it was said that the whole process could be made more efficient.

The second phase, when the patient is actually in the room, only consists of the conduction of one optional test and the explanation of the diagnosis and treatment decision. In this phase, the computer system is of very little use to the doctor. However, they sometimes do use it to show the patient their OCTs and highlight the importance of treatment.

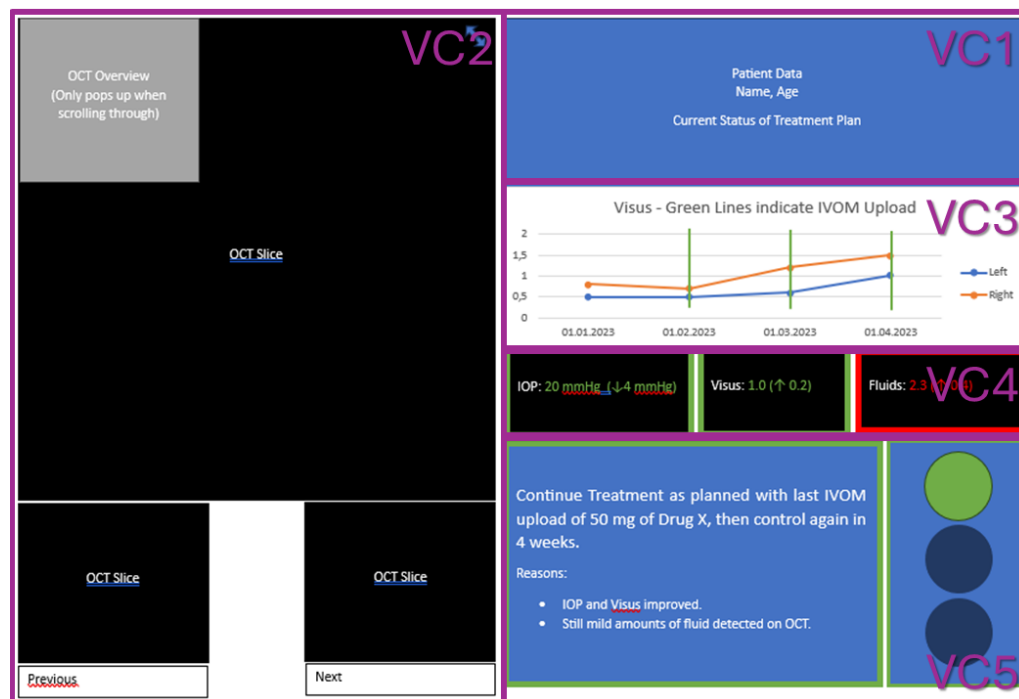


Figure 4.1: Low fidelity prototype developed for this thesis and its visual components (VC): VC1 = Infobox, VC2 = OCT viewer, VC3 = History graphs, VC4 = Metrics, VC5 = Recommendation

### 4.1.2 Low Fidelity Prototype

After the interview a low fidelity prototype was developed and afterwards evaluated by the same ophthalmologist. Figure 4.1 shows this prototype. Its visual components are an infobox (VC1), a viewer for OCTs (VC2), history graphs (VC3), changes from the most important metrics (VC4) and the recommendation (VC5). The ophthalmologist agreed that the low fidelity prototype would be useful and only criticized the reasoning for the recommendation.

## 4.2 Study Protocol

To evaluate whether the dashboard could improve efficiency, informedness and user experience, a series of semi-structured interviews followed by a questionnaire was conducted at the eye clinic in Sulzbach. The study design followed an approach from Bhattacharya et al. [10], who created several tasks, which participants had to solve. This makes the participants actively use the system and encourages interaction. The tasks were developed from the preliminary workflow study. The interviews were audio recorded and the recordings were automatically transcribed using Google's Live Transcribe application<sup>17</sup>. The dashboard as a web application was opened inside a browser on a standard work station. The protocol for the study was as follows:

1. **Recording Agreement:** Participants were first asked to sign a consent form agreeing to the recording of their voice.
2. **Preparation:** Participants were given a disclaimer, that the dashboard is a prototype and certain loading times are expected. However, loading times should be disregarded for the evaluation.
3. **Tutorial:** One interviewer gave a short introduction on the usage of the dashboard showing each feature quickly. Participants were then given ten minutes to explore the dashboard themselves thinking aloud and asking questions about the usage.
4. **Tasks:** The participants were presented with three tasks (T1-3). For each task the dashboard was set to unseen data of a new patient. However, each participant was presented the same patient data for the same tasks. Participants were asked to think aloud during completion of the task. After completion they were asked a set of fixed questions (Q1-3) as well as questions that fit the situation. It was always asked whether this task could be solved more efficiently, more informed or with better user experience than with their usual software.
  - **T1:** Find out the age, gender, BMI, smoking behaviour, disease (AMD or DR) and how often, with what medication and how effective a patient has been treated. Analyze visual acuity and IOP history. Find any other differences since the last visit.
  - **Q1:** Is the data visualized and presented in an intuitive and understandable way?
  - **T2:** Analyze and compare the current OCT with other OCTs from the past. Describe the evolution of biomarkers.

<sup>17</sup><https://support.google.com/accessibility/android/answer/9158064?hl=en>



- **Q2:** Would you prefer the 3D reconstruction over the slices under ideal circumstances (Fast loading times, perfect segmentation)?
  - **T3:** Decide the next step in the treatment of a patient.
  - **Q1:** Do you trust the recommendation? After explanation of the recommendation and prognosis model: Does the explanation increase your trust in the system?
5. **Questionnaire:** Participants were asked to complete a questionnaire that included demographic information, opinions and experience in AI and questions from the Systems Usability Scale (SUS). You can find the german questionnaire in appendix B

## 4.3 Results

### 4.3.1 Demographics

Eleven ophthalmologists (three female, 27%; eight male, 73%) participated in the study on three different days. Their experience ranges from less than one year to 21 years, whereas nine were assistant doctors, one was a specialised ophthalmologist (also just called ophthalmologist in the following) and one was a senior ophthalmologist. To avoid confusion, these groups will be called assistant, specialist and senior. One assistant claimed to have 10 years of experience. However, they answered that they are an assistant doctor, which leads to the suspicion that the stated experience in years is wrong. Furthermore, another assistant missed filling in the questionnaire, which is why we only have the demographic data but no sentiment on AI or SUS ratings for that person. The participants were asked to rate their opinion on AI and ML from completely against it (0) to completely pro AI (5). Four people said that they were completely pro AI (5), while six people said they were somewhat pro AI (4). Their experience with AI was also rated on a scale from no experience (0) to expert experience(5) and, while they were all pro AI, their experience with it was rated intermediate (3) by seven out of ten people. One person said they only had some experience (2), while another said they had advanced experience (4). The senior physician claimed to be an expert in AI (5). Additionally, their general experience with software and computers was rated in the same manner. However, here the senior physician only rated themselves as advanced (4). Three assistant doctors, who had some to intermediate experience with AI, said that their general knowledge with software was advanced (4). Two even said that they were experts (5) in that field, while they were only intermediates with AI. The remaining three rated themselves again as intermediates (3).

### 4.3.2 Systems Usability Scale

The average SUS score from ten participants was 81.75 with a standard deviation of 10.1. This indicates that the dashboard was generally found to be well usable. The highest rating was 95 and the lowest 62.5. Two participants that rated themselves as intermediates in software knowledge gave the lowest ratings of 62.5 and 67.5, while those that rated themselves to have advanced or expert knowledge gave generally higher ratings. The lowest scores also rated their need for a technical person to help them use the dashboard higher than all others (3 and 4 on the Likert scale). They were both assistant doctors with four and 1 years of experience. The highest ratings of 95 and two times 90

came from two assistant doctors and the one specialist. The senior physician's answer summed up to a 85 SUS score.

### 4.3.3 Interview Questions

A fixed set of yes or no questions was asked after the completion of each task. For the first task T1, ten out of eleven participants answered that they felt more efficient than usual, while one person did not answer this questions. Nine of eleven participants also said that they were more informed with the new dashboard, whereas the senior physician did not feel more informed and one other person did not give a clear answer. Furthermore, user experience was improved for eight doctors, while three did not give a distinguished answer. Ten participants answered with yes to Q1, while one person did not answer clearly.

The analysis of the OCT data from T2, however, was rated as more efficient by only six out eleven participants, while two did not answer clearly and the remaining three said they do not feel more efficient with the dashboard. However, eight felt more informed, whereas two did not sufficiently answer and the senior physician again did not feel that way. Only four people reported a better user experience. However, the rest did not answer clearly. Only three people said they would prefer the 3D view under perfect circumstances (Q2). Three people did not sufficiently answer and the remaining five reported that they would not prefer the 3D view to the slices. It seemed that the more experience doctors had, the less likely they were to prefer the 3D view.

For the last task, T3, nine of eleven medical experts said they felt more efficient, whereas one did not feel more efficient and the senior physician failed to answer clearly. Moreover, the senior ophthalmologist did not provide clear answers for informedness nor user experience. From the remaining ten participants nine said that they felt more informed and seven had a better user experience. The specialised ophthalmologist was in both groups. One person did not feel more informed, and three did not answer sufficiently for the user experience. Only three people initially trusted the recommendation, while the rest did not. However, after explanation of the recommendation system five assistant doctors trusted the system. The senior physician as well as two assistant doctors did not trust the system before and after the explanation.

### 4.3.4 Qualitative Analysis

Thematic analysis was conducted on the transcripts of the interview sessions. In total 66 codes were generated with 320 annotations in the transcripts. Generation of codes was done in a joint coding session with a Psychology Phd. student from Saarland University. They were then categorized into 16 supercodes, which again could be divided into four major categories: Trust, Efficiency, Informedness and User experience. Two unaffiliated participants were asked to fill in a questionnaire for the interrater-reliability measurement. They were asked to assign 15 codes to the respective categories and 21 text passages to their respective codes. The assignment of codes to categories had a very good Cohen's  $\kappa$  of 0.91 and the coding of transcript snippets a good one of 0.57. This shows that raters agree with the given categorization and coding of the transcripts.

## Trust

In the trust category, about half of the participants mentioned a low perceived trustworthiness of the system (6, 54.5%). Three participants mentioned that the prognosis specifically was less trustable, since its course is too individual to be predictable. Moreover, three participants found the 3D view not trustworthy, while one participant mentioned not trusting the segmentation in general. The senior physician was among both groups calling the prognosis "reading coffee grounds", although it was initially determined in the OphthalmoAI project, that this is one of the main wishes of ophthalmologists. Moreover, two assistant doctors mentioned that they do not understand the recommendation and its reasoning, hence, trusting it less (2, 18.2%). However, all participants said that the system must be controllable in order to establish trust (11, 100%). Moreover, most participants agreed that feedback options would increase trust in the system (8, 72.7%). The senior physician and one assistant doctor with ten years experience claimed that a feedback system would fail due to the lack of usage under time pressure. However, the senior and two assistant doctors also said they would use the feedback when errors occur. While in general the question was about a feedback button, three participants mentioned that they would even correct segmentation lines manually. Four doctors amongst them the specialist said knowing that the system is controlled through other doctors increases their trust immensely. Lastly, one person mentioned that the visualizations of the system can be used to establish better trust of the patient in the doctor's treatment (1, 9.1%).

## Efficiency

For the efficiency aspect, three people complained that having to read unnecessary, non-critical factors decreases efficiency (3, 27.3 %). They said that specifically the IOP should only be mentioned when it is critical to the situation. For example when it is too high or too low and treatment has to be aborted. Almost all but one assistant doctor mentioned that they felt more efficient using the dashboard (10, 90.9%). However, note that the interview questions have not been coded as they were evaluated separately. Six participants including the specialist said that the metadata analysis leads to enhanced efficiency. Eight participants including the senior physician reported that the enhanced overview improved their efficiency. Moreover, two assistant doctors claimed that coming to the same treatment decision as the recommendation speeds up the decision process. For the OCT analysis task, three assistant doctors felt neither an improvement nor a decline in efficiency (3, 27.3%). Similarly, about half of the participants including the two most experienced claimed that for the standard patient new features are unnecessary and only the OCT slices and the visual acuity values would be needed (6, 54.4%). Finally, seven doctors mentioned that loading speeds play a big role in efficiency (7, 63.6%).

## Informedness

In the informedness category, all but one assistant doctor found that the dashboard provides them more information than usual (10, 90.9%). Three participants among them the specialised ophthalmologist said that the 3D view offered practical insights. Seven doctors including the senior physician reported the quantifications of fluids as useful. The senior physician called them "unavoidable" and "the future" of treatment indication. They expect that in the future precise volume levels will be used to decide which drug will be used. Five people including the specialist reported that segmentations were a good feature to detect the position and morphology of lesions. Furthermore, one assistant

doctor said that they would incorporate the prognosis into their treatment decision. Most doctors reported that in some way that the dashboard would improve the quality of treatment (8, 72.7%). Four assistant doctors said that having a diverging opinion from the system's recommendation would lead to more intensive control and, hence, improving patient care. Moreover, six out of eleven participants including the specialist said that the dashboard and its functionality is particularly useful in borderline cases and unsafe decisions. Three participants including both the senior and the trained specialist reported that the segmentation feature would be even more useful for other diseases such as glaucoma. However, eight participants wanted even more data in the dashboard (8, 72.7%). In this manner, one assistant doctor wanted drug prices to be included in the recommendation process, as this is an important feature for the clinic and insurances. Moreover, two assistants wanted the total retinal thickness as a biomarker, another wanted the exact time since the last IVOM. Other wishes were more information about medical history data such as surgical procedures, the reason for drug switches, the reason for prolonged absence of the patient, more detailed information on the diagnosis (wet or dry AMD for example) or on other concurrent diseases like diabetes. All these wishes stem from assistant doctors. However, the specialised ophthalmologist also mentioned that they would like to see the total number of given IVOMs. The senior physician did not miss any data, but they also reported that they usually will get metadata presented by their assistant doctors.

### User Experience

Finally, user experience was also rated positively in almost all cases (10, 90.9%) except by one assistant doctor. This positive rating bases itself on multiple reasons. The senior and three assistant doctors reported that the combination of OCT viewer and patient medical history data contributes positively on the user experience. Moreover, the senior, the specialist and three other assistant doctors liked the comparison using a slider a lot, as it allowed for easy comparison compared to having multiple windows open. Additionally, they said that it was a better way to compare these images than the flickering they are used to. Flickering is a method to compare OCT slices. The slices are layed on top of each other and the transparency of the top slice will be turned on and off very quickly creating the flicker effect. Only participant four did not like this way of comparing. Four other assistant doctors rated the user experience positively, since they liked the overview of the 3D graph. One assistant doctor and the specialised ophthalmologist found the dashboard to be very intuitive. Moreover, two assistant doctors rated the system recommendation as positive, as it is reassuring to see that the system's recommendation agrees with the doctor's. However, seven assistant doctors also experienced uncertainty with the new system (7, 63.6%). One found the dashboard to be overwhelming, especially the 3D view. Two doctors reported that they were unsure, which OCT is the older and which the recent one in the compare slider. Two other doctors thought that changing the comparison OCT also changed all the metrics to compare to that visit date. Another three doctors were unsure, which eye side was displayed on the dashboard as there was no marker for that. All doctors reported that they do not want to get used to new features (11, 100%). Whereas six assistants and the senior doctor directly reported this, the doctors also mentioned this indirectly. In this manner, eight doctors including the senior and the specialist reported that they were missing the ability to use the scroll wheel to change the slices. One assistant doctor said that getting used to the new comparison slider takes too much time and that they would rather use the old system. They and two more assistants also said that they liked the tabular information of the patient's historical data better.

This doctor was also the only one that did not rate the user experience positively. One other assistant doctor said that using the active ingredients instead of the medication names was unusual. Finally two assistant doctors and the trained specialist said that they would not use the 3D view just because they do not want to get used to it.

### 4.3.5 Feature specific Feedback

While the SUS Score shows that the system is very usable, the thematic analysis highlights the potential for improvement of some features. However, due to time constraints thematic analysis was not specifically conducted to evaluate certain features of the dashboard and, hence, also not validated through Cohen's  $\kappa$ . In this section, I still want to summarize the main feedback points for some visual components. The remaining visual components were neither specifically critiqued nor were changes or new functionalities requested.

#### Top bar

The top bar (VC1) served as a starting point highlighting important metadata about the patient as well as a "at a glance" look at the treatment status and the history of IVOMs. Generally, this feature was positively rated. The only complaints were, that the treatment status should show exactly, how long ago the last IVOM was instead of showing "Over one month ago", and, that the IVOM timeline is confusing. The IVOM timeline should show, which dates are selected as separate lines, and it should better highlight the temporal distances between IVOMs. However, no participant could propose a way of achieving the latter. Moreover, the timeline should show the same time range as the history graphs, as this lead to confusion with one participant. Finally, there should be a table summarizing the amount of given medications.

#### OCT Viewer

The OCT viewer (VC2) as the main tool of ophthalmologists was critiqued the most. The most mentioned problem was that participants could not scroll through the slices. Additionally, some missed a legend for the segmentation and the 3D objects. Moreover, a few participants requested a comparison feature for the 3D view. In the 3D view, one participant requested a feature to turn off certain layers. The loading times were also almost always negatively mentioned.

#### Metrics

Participants wanted the metrics (VC5) to represent the comparison between the selected date and the current date instead of the last date and the current one. Also the visualization using percentage instead of actual values was confusing for some users.

#### Recommendation

Since most users trusted the recommendation more after learning about its internal workings, they were confused about, why the system does not list the reasons from the guidelines. For example, if the model aborts because of extreme IOP, this should be listed. Moreover, the recommendation would always suggest a control or therapy, but doctors

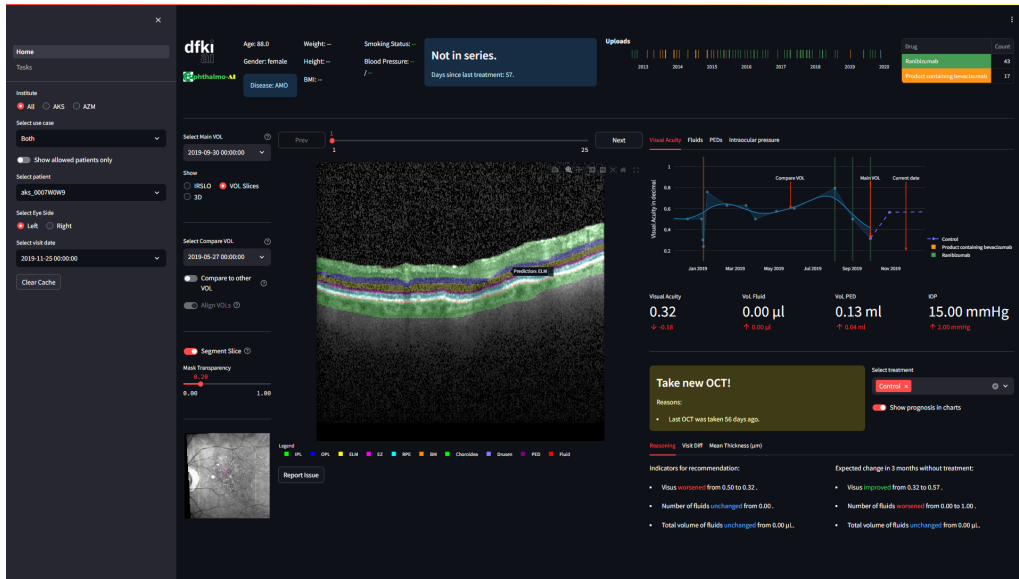


Figure 4.2: Overview of the final dashboard prototype.

often would take a new OCT, if none was available to assess the current situation. The system should also recommend new medical imaging before deciding on treatment.

## 4.4 Final Improvement Iteration

After the evaluation of the user study, the dashboard was improved once more regarding the collected feedback. The final version of the CDSS can be seen in figure 4.2. One can observe that there are many small differences to the evaluated dashboard shown in figure 1.1. Starting from the top bar (VC1), the treatment status now shows exactly, how many days have passed since the last IVOM. Additionally, a table summarizing the amount of each medication given has been added. The OCT viewer (VC2) has been mostly left the same, although legends for the segmentations were added to each plot. The "Report issue" button represents a feedback option. The button opens a dialog, where the user can give feedback on the segmentation and recommendation, which is then sent to an AL backend, such that models can be adapted. However, no API exists yet. The history graphs (VC3) have been adapted to show the current date and selected OCT dates. Furthermore, actual measurements are now represented as scatter points and a line shows the interpolated and smoothed data. The metrics (VC4) now show actual values instead of percentages. The recommendation (VC5) was adapted to list reasons as they appear in the guidelines. It also now considers that new OCTs should be taken after some time.

---

# Chapter 5

## Conclusion

In this chapter, the key findings of this thesis will be summarized, and potential future research will be outlined. The chapter will begin with a discussion of the prognosis model's results, emphasizing its advancements and shortcomings, followed by an outlook on future developments. Next, a concise summary of the recommendation system's performance and potential will be provided. Finally, the outcomes of the user study will be examined, accompanied by suggestions for future research directions. This comprehensive overview will not only show the thesis's primary contributions but also highlight areas for further research.

### 5.1 Prognosis model

#### 5.1.1 Discussion

In this thesis, an advanced time series forecasting model for ophthalmology health metrics was developed and evaluated. The model utilizes a Bidirectional Long Short-Term Memory (BiLSTM) network, demonstrating robust performance in predicting visual acuity metrics. Furthermore, it exhibited exceptionally high accuracy in forecasting the number and volume of intraocular fluids. However, it is important to note that the latter results may be subject to bias due to the highly skewed distribution of the target variables.

To ensure a comprehensive understanding of the model's decision-making process, the feature importance was assessed using SHAP values. The analysis revealed that the BiLSTM network effectively identifies and leverages clinically relevant features in its predictions. Notably, across all prediction tasks, the target variable emerged as the most significant feature, underscoring the model's reliance on historical data for accurate forecasting.

The performance of this predictive model underscores its potential applicability in real-world Clinical Decision Support Systems (CDSS). Such systems could be instrumental in enabling early intervention and preventive treatment strategies for patients suffering from Age-related Macular Degeneration (AMD) or Diabetic Retinopathy (DR).

### 5.1.2 Future Work

For future research, it is recommended to explore alternative architectures, including bidirectional GRUs, Transformers, and Autoencoders, among others. Incorporating a broader range of features from the existing database could further enhance model accuracy and generalizability. Additionally, feeding raw OCT data directly into the network, as opposed to using pre-processed quantifications from another network, could help mitigate potential biases introduced by the segmentation and quantification algorithms. Further investigations could also explore the impact of varying sequence lengths and target time frames on model performance. This could provide deeper insights into the temporal dynamics of ophthalmology health metrics and improve the accuracy of the model. By addressing these aspects, future studies can refine and extend the utility of time series forecasting models in ophthalmological care, ultimately contributing to better patient outcomes and more efficient clinical workflows.

## 5.2 Recommendation System

### 5.2.1 Discussion

For ophthalmologists dealing with DR and AMD, the critical decision is whether to administer Anti-VEGF medication to a patient. This decision follows clear guidelines, making its automation highly desirable. However, developing a recommendation system in this field has proven to be challenging. The algorithmic approach taken in this thesis achieved only 60% accuracy when evaluated against historical data. While the system was notably effective in recommending the same medication as actually given—suggesting that the medication selection process is sound—its overall treatment recommendation accuracy is barely above chance. Consequently, it may not yet be suitable for clinical application.

Despite its suboptimal performance, the system's internal logic is clinically well-grounded, as it strictly applies official therapy guidelines. This discrepancy suggests that the historical data may not fully adhere to these guidelines, or that there may be errors in the quantification of fluids through the segmentation network and reconstruction algorithm, or inaccuracies in the clinical annotations.

### 5.2.2 Future Work

Future research should focus on examining the identified flaws and their impact on the model's performance to validate the recommendation logic. Additionally, a more sophisticated evaluation could be conducted by comparing the system's recommendations against a validated gold standard developed by ophthalmology experts. Enhancing the dataset quality and ensuring it adheres closely to clinical guidelines could also improve the system's accuracy. Incorporating more comprehensive patient data, including surgical histories and other relevant factors, might further refine the model's decision-making process. Moreover, integrating advanced machine learning techniques such as RNNs could enhance the system's ability to handle outliers and varied clinical scenarios. One could train the forecast model to predict the historical medication. However, it is unclear if this target aligns with guidelines. Hence, a validated groundtruth must be established first.



## 5.3 User Study

### 5.3.1 Discussion

The user study found that the developed CDSS has improved efficiency, informedness and user experience in multiple aspects. First and foremost, participants found that the fusion of all systems into a single dashboard yielded a clear overview over all the data, which enhances all three aspects. This overview mainly decreases the time and effort needed for the metadata analysis. However, alignment with the recommendation was found to speed decision processes up and increase the user experience by giving a sense of security. Notably many participants requested more information to be displayed, although even more reported that the dashboard already provides more information than their usual setup. The primary reason for that being the quantification of the 3D reconstructed fluids, which was even titled as "the future of indication" by the senior physician. Moreover, the resulting 3D graph as well as the segmentations enhanced informedness because of their ability to show regions of interest at a glance. Although most said that the dashboard offers more information, some participants complained about the display of non-critical data making them waste time. Therefore, when creating a CDSS one needs to carefully consider, which data to display. Participants also reported that the dashboard would increase the quality of treatment overall. Some contributed this to more intense control of their own decision, when diverging from the treatment recommendation. Others mentioned that the dashboard was particularly useful for borderline cases or even different ophthalmological areas such as glaucoma treatment. Moreover, the new features improved user experience, as they allowed for easier comparisons, which also reflects on the very good SUS score of 81.75.

However, there was also critique on the dashboard. Primarily participants were hesitant to trust the dashboard, especially its AI features. In this sense, most people did not trust the recommendation at the start. Although, after an explanation of the internal workings of the recommendation algorithm about half of them changed their mind and said they would trust the recommendation more now. This low perceived trustworthiness also affects the segmentation, 3D reconstruction and the prognosis. Generally, participants said that the only way to gain their trust was by using the dashboard for extended periods of time and controlling the accuracy of the dashboards predictions and recommendations. This means that the CDSS must always feature the raw data for comparison. Some participants quickly lost faith in the system after encountering errors. However, participants reported that feedback options would reestablish trust in the system in these cases. The collective supervision of many medical experts through this feature would improve confidence even more, although some raised concerns about the adoption of it because of users' timely constraints. Hence, feedback options must be fast and simple. However, more fine grained options should be offered as well, as some users reported they would manually correct segmentations to get more precise quantifications. Although in general efficiency was improved, the analysis of the OCT data was unchanged according to some participants. This was probably due to the fact, that trust in the system was not established yet and there was some uncertainty when using a new system. All participants agreed that they do not like having to get used to new features, as it costs too much time. Hence, one can conclude that a proper evaluation of a CDSS needs its participants to use the system for an extended period of time, such that they can establish trust and get used to the new features.

### 5.3.2 Future work

This study delivers an overview on consideration when designing an AI supported CDSS, that integrates the therapy workflow of a clinician. However, many aspects remain unexplored and the dashboard could be improved and evaluated in a second prototyping round and user evaluation study. Future work could include a qualitative comparison of physicians using their usual setup and physicians using the CDSS, given that the second group gets enough time to familiarize with the CDSS. This could bring deeper insights into the effect of the dashboard onto efficiency, informedness and user experience. Other possible research questions could revolve around efficient ways to establish trust in the system. One could offer different versions of the CDSS showing different explanations like SHAP values and compare how users evaluate the trust in the system. By addressing these aspects, future studies can contribute to the creation of reliable and trustable AI supported CDSS, which inadvertently improve patient care and healthcare as a whole.

## 5.4 Potential Applications

The developed CDSS has shown significant promise in enhancing ophthalmologists' user experience, efficiency, and decision-making capabilities. With further development, the CDSS could be fully integrated into clinical settings, particularly through the implementation of a human-in-the-loop system. In this framework, physicians would utilize the dashboard to aid their decision-making process while providing feedback on any incorrect predictions. The CDSS would learn from this feedback, continuously refining and improving its predictive accuracy.

Figure 5.1 illustrates the scheme of such a system. Initially, clinics send OCT and EHR data to the CDSS database. The database supplies this data to the models, which either segment and quantify the data or provide forecasts, depending on the model. These predictions, along with the raw data, are then fed to the recommendation algorithm, which computes the treatment decision. All this information is displayed by the visual components, where medical experts can interact.

Additionally, these experts can provide feedback through various granular options. For instance, a button could allow users to mark a segmentation or recommendation as incorrect. The model can then be retrained on these samples if ground truth values are available. If not, the OCTs will be sent to an annotation tool, where doctors can manually segment them. The annotation tool feeds these new ground truths back to the database, triggering the feedback system to initiate retraining. Active Learning models like EdgeAL can make use of this data, because the marking provides a direct measure of uncertainty, although not very fine-grained. Integrating the annotation tool directly into the CDSS would enable direct manual correction of incorrectly segmented slices. The difference between the two masks could provide a more detailed measure of uncertainty for an AL model.

By consistently incorporating feedback, the dashboard can iteratively improve, offering more reliable and precise predictions. This, in turn, will enhance the clinician's efficiency, informedness, and overall user experience. Although AL models exist in the OphthalmoAI project, integrating them in the dashboard was out of the scope of this thesis.

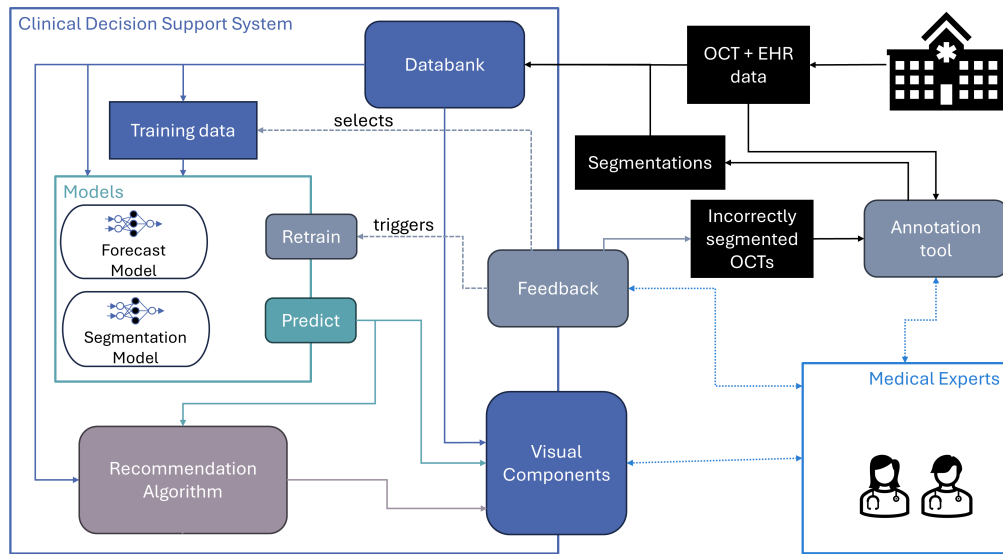


Figure 5.1: Scheme of the fully integrated Clinical Decision Support System with Feedback functionality. Dashed, dotted and continuous lines indicate feedback, interaction and data flow, respectively. The data comes from the clinic and enter the databank of the CDSS. This data is fed to the models for training, to the recommendation system for treatment recommendation and to the visual components for visualization. The models provide predictions for both the recommendation algorithm and the visual components. The recommendation is also displayed in the visual components. The user (medical expert) is interacting with the visual components, the feedback system and the annotation tool. The feedback system captures the users feedback and can trigger a retraining of the models, select training data or provide bad model predictions to the annotation tool. The user can interact with the annotation tool to segment OCTs from the eye clinic or from the feedback system to feed segmentation data into the databank.

---

## Bibliography

- [1] 2019. IDF Diabetes Atlas 10th Edition. [www.diabetesatlas.org](http://www.diabetesatlas.org). (2019). Accessed: 2023-10-26.
- [2] 2021. Age-Related Macular Degeneration (AMD). <https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/age-related-macular-degeneration>. (June 2021). Accessed: 2023-10-27.
- [3] 2023. Age-Related Macular Degeneration (AMD): Symptoms, Causes, Treatment. <https://www.webmd.com/eye-health/macular-degeneration/age-related-macular-degeneration-overview>. (June 2023). Accessed: 2023-10-27.
- [4] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 75, 13 pages. DOI: <http://dx.doi.org/10.1145/3411764.3445736>
- [5] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A. Becker, and Catherine Mooney. 2021. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences* 11, 11 (2021). DOI:<http://dx.doi.org/10.3390/app11115088>
- [6] Rajendra S. Apte. 2021. Age-Related Macular Degeneration. *New England Journal of Medicine* 385, 6 (2021), 539–547. DOI:<http://dx.doi.org/10.1056/NEJMc2102061> PMID: 34347954.
- [7] C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. 1996. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)* 22, 4 (1996), 469–483.
- [8] Michael Barnett, Dongang Wang, Heidi Beadnall, Antje Bischof, David Brunacci, Helmut Butzkueven, J William L Brown, Mariano Cabezas, Tilak Das, Tej Dugal, and others. 2023. A real-world clinical validation for AI-based MRI monitoring in multiple sclerosis. *NPJ Digital Medicine* 6, 1 (2023), 196.
- [9] Eta S Berner. 2007. *Clinical decision support systems*. Vol. 233. Springer.
- [10] Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. 2023. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. ACM. DOI: <http://dx.doi.org/10.1145/3581641.3584075>
- [11] Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*. Vol. 4. Springer.

- [12] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 807–819. DOI:<http://dx.doi.org/10.1145/3490099.3511139>
- [13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [14] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.
- [15] Virginia Braun and Victoria Clarke. 2021. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and psychotherapy research* 21, 1 (2021), 37–47.
- [16] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* 189 (11 1995).
- [17] Michael Brownlee, Lloyd P. Aiello, Mark E. Cooper, Aaron I. Vinik, Jorge Plutzky, and Andrew J.M. Boulton. 2016. Chapter 33 - Complications of Diabetes Mellitus. In *Williams Textbook of Endocrinology (Thirteenth Edition)* (thirteenth edition ed.), Shlomo Melmed, Kenneth S. Polonsky, P. Reed Larsen, and Henry M. Kronenberg (Eds.). Elsevier, Philadelphia, 1484–1581. DOI:<http://dx.doi.org/https://doi.org/10.1016/B978-0-323-29738-7.00033-2>
- [18] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O’Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–12. DOI:<http://dx.doi.org/10.1145/3290605.3300789>
- [19] Stephanie Chiu, Michael Allingham, Priyatham Mettu, Scott Cousins, Joseph Izatt, and Sina Farsiu. 2015. Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. *Biomedical Optics Express* 6 (04 2015). DOI:<http://dx.doi.org/10.1364/BOE.6.001172>
- [20] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
- [21] Nicolás Cuenca, Isabel Ortuño-Lizarán, and Isabel Pinilla. 2018. Cellular characterization of OCT and outer retinal bands using specific immunohistochemistry markers and clinical implications. *Ophthalmology* 125, 3 (2018), 407–422.
- [22] Isabel De la Torre-Díez, Borja Martínez-Pérez, Miguel López-Coronado, Javier Rodríguez Díaz, and Miguel Maldonado López. 2015. Decision support systems and applications in ophthalmology: literature and commercial review focused on mobile apps. *Journal of medical systems* 39 (2015), 1–10.
- [23] Statistisches Bundesamt (Destatis). 2021. Körpermaße nach Altersgruppen und Geschlecht. <https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Gesundheitszustand-Relevantes-Verhalten/Tabellen/liste-koerpermasse.html#>. (2021).

- [24] Azade Farshad, Yousef Yeganeh, Peter Gehlbach, and Nassir Navab. 2022. Y-Net: A Spatiospectral Dual-Encoder Network for Medical Image Segmentation. (2022).
- [25] German Research Center for Artificial Intelligence (DFKI). 2021-2024. Ophthalmology AI. <https://iml.dfki.de/ophthalmo-ai/>. (2021-2024).
- [26] Haojie Fu, Bin Zhang, Jianliang Tong, Harold Bedell, Hecheng Zhang, Yating Yang, Chaochao Nie, Yingdong Luo, and Xiaoling Liu. 2017. Relationships of orientation discrimination threshold and visual acuity with macular lesions in age-related macular degeneration. *PLOS ONE* 12 (09 2017), e0185070. DOI:<http://dx.doi.org/10.1371/journal.pone.0185070>
- [27] Bundesministerium für Bildung und Forschung. 2016-2021. (2016-2021). <https://www.gesundheitsforschung-bmbf.de/de/xploit-semantische-unterstuetzung-fur-die-pradiktive-modellierung-in-der-systemm.php>
- [28] Deutsche Ophthalmologische Gesellschaft, Retinologische Gesellschaft, and Berufsverband der Augenärzte Deutschlands. 2019. Therapie des diabetischen Makulaödems. (August 2019).
- [29] Deutsche Ophthalmologische Gesellschaft, Retinologische Gesellschaft, and Berufsverband der Augenärzte Deutschlands. 2022. Anti-VEGF-Therapie bei der neovaskulären altersabhängigen Makuladegeneration. (October 2022).
- [30] Maryellen L Giger. 2018. Machine learning in medical imaging. *Journal of the American College of Radiology* 15, 3 (2018), 512–520.
- [31] Heidelberg Engineering GmbH. 2016. Know your retinal layers. (2016). <https://www.heidelbergengineering.com/int/news/know-your-retinal-layers-33401465/> Accessed on: 13-06-2024.
- [32] VISYO Qualitätsnetzwerk Saar GmbH. 2022. Treat & Extend (T&E)-Schema bei AMD, DMÖ und RVV. (December 2022).
- [33] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Vol. 4. IEEE, 2047–2052.
- [34] Theodore Grosvenor and Theodore P Grosvenor. 2007. *Primary care optometry*. Elsevier Health Sciences.
- [35] Mareike Hartmann, Han Du, Nils Feldhus, Ivana Kruijff-Korbayová, and Daniel Sonntag. 2022. XAINES: explaining AI with narratives. *KI-Künstliche Intelligenz* 36, 3-4 (2022), 287–296.
- [36] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [37] David Huang, Eric A. Swanson, Charles P. Lin, Joel S. Schuman, William G. Stinson, Warren Chang, Michael R. Hee, Thomas Flotte, Kenton Gregory, Carmen A. Puliafito, and James G. Fujimoto. 1991. Optical Coherence Tomography. *Science* 254, 5035 (1991), 1178–1181. DOI:<http://dx.doi.org/10.1126/science.1957169>

- [38] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. 2020. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 1055–1059.
- [39] Plotly Technologies Inc. 2024a. Plotly. <https://plotly.com/>. (2024). Accessed: 2024-05-05.
- [40] Snowflake Inc. 2024b. Streamlit - A faster way to build and share data apps. <https://streamlit.io/>. (2024). Accessed: 2024-05-05.
- [41] Monique W M Jaspers, Marian Smeulders, Hester Vermeulen, and Linda W Peute. 2011. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *Journal of the American Medical Informatics Association* 18, 3 (03 2011), 327–334. DOI:<http://dx.doi.org/10.1136/amiajnl-2011-000094>
- [42] Cheng Jin, Heng Yu, Jia Ke, Peirong Ding, Yongju Yi, Xiaofeng Jiang, Xin Duan, Jinghua Tang, Daniel T Chang, Xiaojian Wu, and others. 2021. Predicting treatment response from longitudinal images using multi-task deep learning. *Nature communications* 12, 1 (2021), 1851.
- [43] Md Kadir, Hasan Md Tusfiqur Alam, and Daniel Sonntag. 2023. EdgeAL: An Edge Estimation Based Active Learning Approach for OCT Segmentation. (07 2023).
- [44] Sharif Amit Kamran, Sourajit Saha, Ali Shihab Sabbir, and Alireza Tavakkoli. 2019a. Optic-net: A novel convolutional neural network for diagnosis of retinal diseases from optical tomography images. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 964–971.
- [45] Sharif Amit Kamran, Sourajit Saha, Ali Shihab Sabbir, and Alireza Tavakkoli. 2019b. Optic-net: A novel convolutional neural network for diagnosis of retinal diseases from optical tomography images. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 964–971.
- [46] Sharif Amit Kamran, Alireza Tavakkoli, and Stewart Lee Zuckerbrod. 2020. Improving robustness using joint attention network for detecting retinal degeneration from optical coherence tomography images. In *2020 IEEE International Conference On Image Processing (ICIP)*. IEEE, 2476–2480.
- [47] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, and others. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell* 172, 5 (2018), 1122–1131.
- [48] Jeany Q Li, Thomas Welchowski, Matthias Schmid, Matthias Marten Mauschwitz, Frank G Holz, and Robert P Finger. 2020. Prevalence and incidence of age-related macular degeneration in Europe: a systematic review and meta-analysis. *British Journal of Ophthalmology* 104, 8 (2020), 1077–1084. DOI:<http://dx.doi.org/10.1136/bjophthalmol-2019-314422>
- [49] Robert Lindsey, Aaron Daluiski, Sumit Chopra, Alexander Lachapelle, Michael Mozer, Serge Sicular, Douglas Hanel, Michael Gardner, Anurag Gupta, Robert Hotchkiss, and others. 2018. Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences* 115, 45 (2018), 11591–11596.

- [50] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. (2017).
- [51] Asimina Mataftsi, Dimitrios Koutsimpogeorgos, Periklis Brazitikos, Nikolaos Ziakas, and Anna-Bettina Haidich. 2019. Is conversion of decimal visual acuity measurements to logMAR values reliable? *Graefe's Archive for Clinical and Experimental Ophthalmology* 257 (2019), 1513–1517.
- [52] Larry R Medsker and LC Jain. 2001. Recurrent neural networks. *Design and Applications* 5, 64-67 (2001), 2.
- [53] Valentyn Melnychuk, Evgeniy Faerman, Ilja Manakov, and Thomas Seidl. 2020. Matching the Clinical Reality: Accurate OCT-Based Diagnosis From Few Labels. *arXiv preprint arXiv:2010.12316* (2020).
- [54] Lucenteforte E Oddone F Brazzelli M Parravano M Franchi S Ng SM Michelessi, M and G Virgili. 2015. Optic nerve head and fibre layer imaging for diagnosing glaucoma. *Cochrane Database of Systematic Reviews* 11 (2015). DOI:<http://dx.doi.org/10.1002/14651858.CD008803.pub2>
- [55] M. D. moff, M. K. Garvin, and M. Sonka. 2010. Retinal imaging and image analysis. *IEEE Rev Biomed Eng* 3 (2010), 169–208.
- [56] Christoph Molnar. 2022. *Interpretable Machine Learning* (2 ed.). <https://christophm.github.io/interpretable-ml-book>
- [57] Mohammad Amin Morid, Olivia R. Liu Sheng, and Joseph Dunbar. 2022. Time Series Prediction using Deep Learning Methods in Healthcare. (2022).
- [58] Ovidiu Musat, Corina Cernat, Mahdi Labib, Andreea Gheorghe, Oana Toma, Madalina Zamfir, and Ana Maria Boureanu. 2015. Diabetic macular edema. *Romanian journal of ophthalmology* 59, 3 (2015), 133.
- [59] Quan Dong Nguyen, Arup Das, Diana V. Do, Pravin U. Dugel, Andre Gomes, Frank G. Holz, Adrian Koh, Carolyn K. Pan, Yasir J. Sepah, Nikhil Patel, Heather MacLeod, and Patrik Maurer. 2020. Brolucizumab: Evolution through Preclinical and Clinical Studies and the Implications for the Management of Neovascular Age-Related Macular Degeneration. *Ophthalmology* 127, 7 (2020), 963–976. DOI:<http://dx.doi.org/https://doi.org/10.1016/j.ophtha.2019.12.031>
- [60] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *International conference on machine learning*. Pmlr, 1310–1318.
- [61] Annika Christina Perlich. 2022. Nutzung und Akzeptanz klinischer Entscheidungsunterstützungssysteme-Entwurf eines Modells für die medizinische Lehre. (2022).
- [62] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. (2018).
- [63] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 535, 14 pages. DOI:<http://dx.doi.org/10.1145/3491102.3501967>



- [64] Markus Rohm, Volker Tresp, Michael Müller, Christoph Kern, Ilja Manakov, Maximilian Weiss, Dawn A. Sim, Siegfried Priglinger, Pearse A. Keane, and Karsten Kortuem. 2018. Predicting Visual Acuity by Using Machine Learning in Patients Treated for Neovascular Age-Related Macular Degeneration. *Ophthalmology* 125, 7 (2018), 1028–1036. DOI:<http://dx.doi.org/https://doi.org/10.1016/j.ophtha.2017.12.034>
- [65] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. (2015).
- [66] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: an efficient alternative to SIFT or SURF. *Proceedings of the IEEE International Conference on Computer Vision* (11 2011), 2564–2571. DOI:<http://dx.doi.org/10.1109/ICCV.2011.6126544>
- [67] Clearfield E Soliman MK Sadiq MA Baldwin AJ Hanout M Agarwal A Sepah YJ Do DV Sarwar, S and QD Nguyen. 2016. Aflibercept for neovascular age-related macular degeneration. *Cochrane Database of Systematic Reviews* 2 (2016). DOI:<http://dx.doi.org/10.1002/14651858.CD011346.pub2>
- [68] Tobias Schlosser, Frederik Beuth, Trixy Meyer, Arunodhayan Sampath Kumar, Gabriel Stolze, Olga Furashova, Katrin Engelmann, and Danny Kowerko. 2024. Visual acuity prediction on real-life patient data using a machine learning based multistage system. *Scientific Reports* 14, 1 (2024), 5532.
- [69] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (oct 2019), 336–359. DOI:<http://dx.doi.org/10.1007/s11263-019-01228-7>
- [70] Burr Settles. 2009. Active learning literature survey. (2009).
- [71] Lloyd S Shapley. 1951. Notes on the n-person game—ii: The value of an n-person game. (1951).
- [72] Ruijie Shi, Xiangjie Leng, Yanxia Wu, Shiyin Zhu, Xingcan Cai, and Xuejing Lu. 2023. Machine learning regression algorithms to predict short-term efficacy after anti-VEGF treatment in diabetic macular edema based on real-world data. *Scientific Reports* 13, 1 (2023), 18746.
- [73] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*. PMLR, 3145–3153.
- [74] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguerre y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large Language Models Encode Clinical Knowledge. (2022).

- [75] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaeckermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. (2023).
- [76] Daniel Sonntag, Fabrizio Nunnari, and Hans-Jürgen Profitlich. 2020. The Skincare project, an interactive deep learning system for differential diagnosis of malignant skin lesions. Technical Report. (2020).
- [77] Daniel Sonntag, Colette Weihrauch, Oliver Jacobs, and Daniel Porta. 2010. *THE-SEUS CTC-WP4 Usability Guidelines for Use Case Applications*. Technical Report. Bundesministerium für Wirtschaft und Technologie.
- [78] Letizia Squarcina, Filippo Maria Villa, Maria Nobile, Enrico Grisan, and Paolo Brambilla. 2021. Deep learning for the prediction of treatment response in depression. *Journal of affective disorders* 281 (2021), 618–622.
- [79] Anindya Pradipta Susanto, David Lyell, Bambang Widyanoro, Shlomo Berkovsky, and Farah Magrabi. 2023. Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. *Journal of the American Medical Informatics Association* (08 2023), ocad180. DOI: <http://dx.doi.org/10.1093/jamia/ocad180>
- [80] Reed Sutton, David Pincock, Daniel Baumgart, Daniel Sadowski, Richard Fedorak, and Karen Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. 3 (02 2020). DOI:<http://dx.doi.org/10.1038/s41746-020-0221-y>
- [81] Kenji Suzuki. 2017. Overview of deep learning in medical imaging. *Radiological physics and technology* 10, 3 (2017), 257–273.
- [82] Daniel SW Ting, Yong Liu, Philippe Burlina, Xinxing Xu, Neil M Bressler, and Tien Y Wong. 2018. AI for medical imaging goes deep. *Nature medicine* 24, 5 (2018), 539–540.
- [83] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 25, 1 (2019), 44–56.
- [84] Volker Tresp, Sonja Zillner, Maria Costa, Yi Huang, Alexander Cavallaro, Peter Fasching, Andre Reis, Martin Sedlmayr, Thomas Ganslandt, Klemens Budde, Carl Hinrichs, Danilo Schmidt, Philipp Daumke, Daniel Sonntag, Thomas Wittenberg, Patricia Oppelt, and Denis Krompass. 2013. Towards a New Science of a Clinical Data Intelligence. (11 2013).
- [85] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, and others. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* 17, 3 (2020), 261–272.
- [86] David F Watson. 1981. Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes. *The computer journal* 24, 2 (1981), 167–172.

- [87] Devon Watts, Rafaela Fernandes Pulice, Jim Reilly, Andre R Brunoni, Flávio Kapczinski, and Ives Cavalcante Passos. 2022a. Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis. *Translational psychiatry* 12, 1 (2022), 332.
- [88] Devon Watts, Rafaela Fernandes Pulice, Jim Reilly, Andre R Brunoni, Flávio Kapczinski, and Ives Cavalcante Passos. 2022b. Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis. *Translational psychiatry* 12, 1 (2022), 332.
- [89] Moritz Wolf, Dana Ruiter, Ashwin Geet D'Sa, Liane Reiners, Jan Alexandersson, and Dietrich Klakow. 2020. HUMAN: Hierarchical Universal Modular ANnotator. (2020).
- [90] Huiyuan Ding Yixuan Liu Xin He, Xi Zheng and Hongling Zhu. 2023. AI-CDSS Design Guidelines and Practice Verification. *International Journal of Human-Computer Interaction* 0, 0 (2023), 1–24. DOI:<http://dx.doi.org/10.1080/10447318.2023.2235882>
- [91] Joanne WY Yau, Sophie L Rogers, Ryo Kawasaki, Ecosse L Lamoureux, Jonathan W Kowalski, Toke Bek, Shih-Jen Chen, Jacqueline M Dekker, Astrid Fletcher, Jakob Grauslund, and others. 2012. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes care* 35, 3 (2012), 556–564.
- [92] Shiri Zayit-Soudry, Iris Moroz, and Anat Loewenstein. 2007. Retinal pigment epithelial detachment. *Survey of ophthalmology* 52, 3 (2007), 227–243.
- [93] Seyedeh Maryam Zekavat, Sayuri Sekimitsu, Yixuan Ye, Vineet Raghu, Hongyu Zhao, Tobias Elze, Ayellet V Segrè, Janey L Wiggs, Pradeep Natarajan, Lucian Del Priore, and others. 2022. Photoreceptor layer thinning is an early biomarker for age-related macular degeneration: epidemiologic and genetic evidence from UK Biobank OCT data. *Ophthalmology* 129, 6 (2022), 694–707.
- [94] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 3–11.
- [95] A.P. Zijdenbos, B.M. Dawant, R.A. Margolin, and A.C. Palmer. 1994. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Transactions on Medical Imaging* 13, 4 (1994), 716–724. DOI:<http://dx.doi.org/10.1109/42.363096>

---

# Appendix A

## 3D Reconstruction algorithm & Code

The algorithm (a simplified version of the code can be seen in listing A.1) starts by creating two empty lists: one for work in progress (WIP) reconstructions and the other for finished reconstructions. It then starts iterating over all segmented masks of the OCT. The mask is turned into a binary mask, where true values stand for pixels that were annotated as the label that should be 3D reconstructed and false values stand for every other label. If every pixel of the binary mask is false, we can skip this mask and move on to the next. However, if there are some true values, then the algorithm tries to reconstruct the 3D objects. First the masks contours will be computed using the OpenCV python library. Contours are used as we only care about outer points of the object. It also means one needs to save less data in memory. The contours are transformed into 3D contours using the grid position of the current mask. Initially, since the WIP reconstructions list is empty, all contours will be added to this list and for each contour a new ReconstructionBuilder class (which can be seen in listing A.3) is initiated. However, if there are already WIP reconstructions, then these need to be updated.

Hence, the updateWipReconstructions function, which is shown in a simplified version in listing A.2, is called. It first computes the smallest possible distance of each reconstruction to each new contour, which is done in the getDistanceMatrix function. It then iterates over all contours sorted by their smallest possible distance to any reconstruction and checks whether this distance is smaller than the maximum allowed distance. If this is the case, then we add the new contour to reconstruction builder. If not, then we can create a new reconstruction builder without having to check the other reconstructions, as the contours were already sorted by their smallest distance. The maximum allowed distance used in the dashboard was simply the distance between the last two slices times the square root of two, as this seemed to yield the best reconstructions. However, this parameter is experimental and needs to be studied and finetuned. A fixed distance would also clinically make more sense, since some resolutions of OCTs have rather large distances between slices, which do not allow for accurate reconstruction.

Lastly, the algorithm iterates over all WIP reconstructions and checks, whether they still have any points in the current grid position, which means that they were just updated. If not, then the reconstruction is finished and can be build. The building process only assures that the reconstruction is actually a three dimensional object and not just for

example a single point and one contour that lies completely inside a plane. If the object is only one or two dimensional, then the building function adds artificial points around the centroid to make it three dimensional. The parameter "CENTROID\_OFFSET" determines how far off the centroid these points will be set. In y direction, so along the scan grid of the OCT, half the slice distance was used, while for the other directions an offset of 5 pixels was used. Again, this parameter can be adjusted. However, these lower dimensional annotations did not appear very often.

```

1  def reconstruct_objects(oct, label):
2      finished_reconstructions = []
3      wip_reconstructions = [] # WIP = Work in Progress
4
5      for mask in oct.masks:
6          label_mask = mask == label
7
8          if not any label_mask:
9              # No annotations, can be skipped
10             skip
11
12         contours_2D = findContours(label_mask)
13         contours = transformTo3DContours(contours_2D, oct.grid_position)
14
15         if any wip_reconstructions:
16             updateWipReconstructions(wip_reconstructions, contours)
17         else:
18             for contour in contours:
19                 wip_reconstructions.append(ReconstructionBuilder(contour))
20
21         for reconstruction in wip_reconstruction:
22             if not anyPointsInCurrentGridPosition(reconstruction,
23                                                    oct.grid[slice_index]):
24                 finished_reconstructions.append(reconstruction.build())
25                 wip_reconstruction.pop(reconstruction)

```

Listing A.1: Pseudocode of the 3D reconstruction.

```

1  def updateWipReconstructions(wip_reconstructions, contours):
2      distances = getDistanceMatrix(wip_reconstructions, contours)
3      for distance, contour in sortedBySmallestDistance(distances, contours):
4          reconstruction = getClosestReconstructionOfContour(contour,
5                                                              distances,
6                                                              wip_reconstructions)
7
8          if distance < MAXIMUM_ALLOWED_DISTANCE:
9              # Closest reconstruction is inside the allowed distance ->
10             # Add contour to reconstruction
11             reconstruction.add(contour)
12         else:
13             # No reconstruction is inside the allowed distance ->
14             # Create new reconstruction
15             wip_reconstructions.append(ReconstructionBuilder(contour))
16
17  def getDistanceMatrix(wip_reconstructions, contours):
18      distances = []
19      for rec_builder in wip_reconstructions:
20          rec_distances = []
21          for contour in contours:
22              rec_distances.append(rec_builder.smallest_distance_to(contour))
23          distances.append(rec_distances)
24      return distances

```

Listing A.2: Pseudocode of the updateWipReconstructions and getDistanceMatrix function inside the 3D reconstruction code

```

1  class ReconstructionBuilder:
2      def __init__(contour):
3          self.contours = []
4          self.contours.append(contour)
5
6      def add(contour):
7          self.contours.append(contour)
8
9      def smallest_distance_to(contour):
10         distances = []
11         for point in contour:
12             distances.append(np.linalg.norm(self.contours - point, axis=1))
13         return min(distances)
14
15     def build():
16         if pointsAreNot3DimensionalObject(self.contours):
17             addFakePointsCloseToCentroid(self.contours, CENTROID_OFFSET)
18         return Reconstruction(self.contours)
19
20
21 class Reconstruction:
22     def __init__(contours):
23         points = toCartesianPoints(contours)
24         self.hull = ss.ConvexHull(points)
25
26     @property
27     def points:
28         return self.hull.points[self.hull.vertices]
29
30     def get_volume():
31         return self.hull.volume
32
33     def get_top_down_view():
34         2d_hull = computeConvexHullOfXandYPoints(self.points)
35         return transformToContours(2d_hull.points[2d_hull.vertices])

```

Listing A.3: Simplified Pseudocode of the ReconstructionBuilder and Reconstruction class

---

# Appendix B

## Questionnaire

You can find the questions for the questionnaire here. The questions are in german, which was the language that was spoken during interviews.

### B.1 Demographics

1. Bitte geben Sie ihre Test-Personen ID an.
2. Bitte geben Sie ihr Alter an.
3. Bitte geben Sie ihr Geschlecht an.
4. Bitte geben Sie ihren Berufstitel an.
5. Wie lange arbeiten Sie in diesem Beruf?

### B.2 Experience with AI and Software

1. Wie sind Sie generell gegenüber Künstlicher Intelligenz (KI) und Maschinellern Lernen (ML) eingestellt? (Bewerte auf einer Skala von 1 bis 5; 1: Komplette dagegen, 5: Komplette dafür)
2. Wie würden Sie ihre Erfahrung mit KI und ML einordnen? (Bewerte auf einer Skala von 1 bis 5; 1: Gar keine Erfahrung, 5: Experte)
3. Wie würden Sie ihre generelle Erfahrung im Umgang mit Software und Computern einordnen? (Bewerte auf einer Skala von 1 bis 5; 1: Gar keine Erfahrung, 5: Experte)

### B.3 System Usability Scale

All questions were rated from one to five, where one means totally disagree and five means totally agree.

1. Ich denke, dass Ich dieses Dashboard oft benutzen würde.
2. Ich fand, dass das Dashboard unnötig komplex ist.
3. Ich fand, dass das Dashboard einfach zu benutzen war.
4. Ich denke, dass Ich die Unterstützung eines technisch versierten Menschen brauche, um das Dashboard zu benutzen.
5. Ich fand, dass die verschiedenen Funktionen gut in das Dashboard integriert waren.
6. Ich fand, dass es zu viel Inkonsistenz im Dashboard gab.
7. Ich denke, dass die meisten Menschen sehr schnell lernen würden das Dashboard zu benutzen.
8. Ich fand das Dashboard sehr umständlich zu benutzen.
9. Ich fühlte mich beim Verwenden des Dashboards sehr zuversichtlich.
10. Ich musste eine Menge Dinge lernen, bevor ich mit diesem Dashboard loslegen konnte.