Towards Trustable Intelligent Clinical Decision Support Systems: A User Study with Ophthalmologists

Robert Andreas Leist
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Saarland, Germany
robert.leist@dfki.de

Hans-Jürgen Profitlich
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
hans-juergen.profitlich@dfki.de

Tim Hunsicker
Universität des Saarlandes
Saarbrücken, Germany
tim.hunsicker@uni-saarland.de

Daniel Sonntag
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
Applied Artificial Intelligence
Oldenburg University
Oldenburg, Germany
daniel.sonntag@dfki.de

Abstract

Integrating Artificial Intelligence (AI) into Clinical Decision Support Systems (CDSS) presents significant opportunities for improving healthcare delivery, particularly in fields like ophthalmology. This paper explores the usability and trustworthiness of an AI-driven CDSS designed to assist ophthalmologists in treating diabetic retinopathy and age-related macular degeneration. Therefore, we created a CDSS and evaluated its impact on efficiency, informedness, and user experience through task-based semi-structured interviews and questionnaires with 11 ophthalmologists. The usability of the CDSS was rated highly, with a SUS of 81.75. Additionally, results show that participants felt like the CDSS would improve their efficiency and informedness with one major aspect being integrating Electronic Health Records (EHR) and Optical Coherence Tomography (OCT) data into a single interface. Additionally, we explored aspects of the trustworthiness of AI components, specifically OCT segmentation, treatment recommendation, and visual acuity forecasting. Through thematic analysis, we identified key factors influencing trustworthiness and clinical adoption. Results show that a larger degree of abstraction from input to output of a model correlates with decreased trust. From our findings, we propose three guidelines for designing trustworthy CDSS.

IUI '25, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712136

CCS Concepts

• **Applied computing** → **Health informatics**; • **Human-centered computing** → **Empirical studies in HCI**; **Interactive systems and tools**; Visualization.

Keywords

Clinical Decision Support Systems (CDSS), Interactive Machine Learning (IML), Human-AI collaboration, AI-assisted decision making, user trust, domain experts, ophthalmology

ACM Reference Format:
Robert Andreas Leist, Hans-Jürgen Profitlich, Tim Hunsicker, and Daniel Sonntag. 2025. Towards Trustable Intelligent Clinical Decision Support Systems: A User Study with Ophthalmologists. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3708359.3712136

1 Introduction

Artificial Intelligence (AI) has significantly transformed various domains, and its integration into healthcare, particularly in Clinical Decision Support Systems (CDSS), holds immense promise. In the field of ophthalmology, a variety of accurate and effective Deep Learning (DL) - a sub-field of AI - models have emerged due to the availability of high-quality medical imaging data [32, 37]. However, the trust of medical experts towards these tools remains underexplored, even though trust is crucial for human acceptance [20]. Studies indicate that lack of trust, alongside other factors, such as bad usability, halts the adoption of CDSS [28, 47].

Laato et al. [30] define **Trustworthiness** as "*end users' perception about the truthfulness and honesty of the system, as well as beliefs that the system works as intended*". Explainable AI (XAI) aims to improve trustworthiness by making AI decision-making processes more transparent. Model-centric approaches use post-hoc explanations or inherently interpretable models to explain decision-making [27, 39], while data-centric approaches revolve around information about

the training data [3, 4]. Several studies show that medical experts prefer data-centric explanations over model-centric counterparts for building trust [3, 4, 9]. However, while research focuses on how to explain AI, the question of what needs to be explained remains unexplored. Castelo et al. [10] suggest that the inherent objectivity of an algorithm's task plays a significant role in shaping its trustworthiness. We argue, that an AI model can be seen as an algorithm and, therefore, algorithm aversion applies. However, Langer et al. [31] have shown that people perceive trust in automated decision-making systems differently depending on the terminology used, i.e. AI or algorithm. Hence, it is unclear whether task-dependent algorithm aversion can be applied to AI. In the following paper, we want to probe into this concept by proposing the Research Question (RQ):

- RQ1: Are segmentation, classification, and time series forecast models perceived with varying levels of trustworthiness? What influences trust in these components?

Another key consideration is the usability of AI systems in real-world clinical environments. Ophthalmologists, like other specialists, often work under tight time constraints and need tools that integrate seamlessly into their workflows [28]. If an AI system is difficult to use or interpret, it can become a hindrance rather than a help. Studies have shown that the usability of AI tools can significantly influence whether practitioners embrace or dismiss them [28, 47]. By ensuring that these systems are user-friendly and aligned with clinicians' needs, developers can increase the likelihood that these tools will be adopted in clinical practice. In this work, we want to find out which factors of a CDSS in ophthalmology contribute to user experience, efficiency, and informedness:

- RQ2: How can an ophthalmologist's efficiency, informedness, and user experience be improved using a CDSS?

To address these questions, we developed an AI-driven CDSS supporting ophthalmologists in treating two prevalent eye diseases: Age-related Macular Degeneration (AMD) and Diabetic Retinopathy (DR). We implement three AI-driven components for the tasks of semantic segmentation, treatment recommendation, and time series forecasting. We evaluate our prototype in a qualitative user study with eleven ophthalmologists of varying experience. Through think-aloud task-based semi-structured interviews, we explored the perceived trustworthiness of the AI components and evaluated our CDSS in terms of efficiency, information content, and usability. Through thematic analysis, we identified several themes contributing to an AI system's acceptance. To summarize, there are three primary research contributions presented in this paper:

(1) We present a CDSS prototype with perceived efficiency, informedness, and user experience improvements. The prototype serves as a starting point for any researchers and developers working on CDSS in ophthalmology: https://github.com/DFKI-Interactive-Machine-Learning/ophthalmo-cdss
(2) We identify task-dependent algorithm aversion in AI models. Our findings indicate that larger discrepancies between input and output reduce trust in AI-driven components.
(3) We define three guidelines to make AI-driven components more trustworthy:
   (a) Remind users of the non-deterministic behaviour and limitations of AI to mitigate general algorithm aversion.
   (b) Implement XAI methods to reduce the degree of abstraction from input to output of AI components to mitigate task-dependent algorithm aversion.
   (c) Provide quickly accessible feedback options to give users a feeling of control over the models.

2 Related Work

In this section, we provide an overview of the relevant literature. First, we discuss the treatment practices of ophthalmologists and explore how AI methods can enhance their work. Next, we review key studies on trust in AI, which led to the formulation of our RQs.

2.1 Improving care in Ophthalmology

In ophthalmology, neovascular diseases like DR and AMD are characterized by the accumulation of fluids inside retinal layers. If left untreated, these conditions can lead to scars and other occlusions, ultimately causing vision loss [19, 33]. Treatment typically involves a series of intravitreal injections of medication (IVOM) when the presence of fluids is detected [46]. This is done manually by one or two doctors using Optical Coherence Tomographies (OCTs), a medical imaging technique similar to ultrasound. OCTs consist of an "en face" view of the retina (called IRSLO[1]) as well as an array of cross-sectional slices of the retina [24, 44]. The subjective evaluation of these slices needs prolonged training, and trainees often feel unconfident, as a study from 2019 shows [15]. Even between experts, there exist certain discrepancies in the annotations of biomarkers on OCTs [35, 38], highlighting the need for objective analysis.

2.1.1 Segmentation and Quantification. Recent developments in ML and semantic segmentation, facilitate the automatic quantification of medical images, specifically OCTs [32, 37]. Semantic segmentation describes the task of assigning each pixel on an image a class. Many DL architectures have been developed for this purpose. Most notable is UNet, which is a Convolutional Neural Network (CNN) with an encoder and decoder path connected via skip connections [41]. YNet builds upon the structure of UNet by adding another encoder branch, which first transforms the image into the Fourier domain [18]. The authors achieve state-of-the-art performance on OCT segmentation on the Duke [12] and UMN [40] data sets, especially regarding fluid detection. The model is trained to segment individual slices. Commercial software for quantifying fluids on OCTs exists (e.g. Fluid Monitor from RetInSight[2]), but it has yet to be implemented into a CDSS that gives therapy recommendations. Therefore,

[1] InfraRed Scanning Laser Ophthalmoscopy
[2] https://retinsight.com/fluid-monitor/ (Accessed: 05.01.2025)

in this work, we implemented a quantification algorithm based on segmentations and a recommender system that utilizes these quantifications.

2.1.2 Time Series Forecasting. Another critical aspect of therapy is scheduling appointments that effectively balance the need to catch every crucial symptom emergence and the burden of frequent visits for patients and doctors. Kim et al. [29] found that an increased frequency of doctor visits correlates with less life satisfaction. To address this issue, we imagine that a time series forecast model will help identify crucial points in the disease progression, leading to less frequent visits. We focus on one key metric: Visual Acuity (VA). VA measures the sharpness of vision and can be measured in several different units. There are several studies showcasing the application and benefit of time series forecast models in medicine [25, 45, 48, 49]. Schlosser et al. [43] evaluated several ML models on the task of VA forecasting and found that their best model outperformed a trained ophthalmologist by 19.7% on macro average F1-Score on a Winner Stabilizer Loser scheme. However, in a medical context, a regression might be more interesting, as the amplitude of change is important. For example, in the WSL scheme, a decrease of 0.1 and 0.9 would be equally important. Although Schlosser et al. [43] used regression for their models, they did not share regression performance. They found that a Multi-Layer Perceptron for VA regression with Linear Discriminant Analysis on the regressed value performed best for the classification. However, a meta-analysis of time series prediction models in healthcare by Morid et al. [36] has shown that Recurrent Neural Network (RNNs) [34] architectures are more performant. Especially Bidirectional Long Short Term Memory (BiLSTM) [21, 42] and Gated Recurrent Unit (BiGRU) [13, 42] networks performed well. In [43], BiLSTM and BiGRU performed worse in the classification task. However, it is unclear whether this is because of the nature of the task. We imagine that a forecast model can be included in the appropriate scheduling of appointments, leading to better patient satisfaction and better availability of doctors without loss of treatment quality.

2.1.3 Visualizing Electronic Health Records. Treatment of DR and AMD affords careful monitoring of many aspects of a patient's history. These are tracked in Electronic Health Records (EHRs). In a preliminary study, we found that EHRs in ophthalmology usually do not offer visualizations of the data they contain. Consequentially, doctors are required to memorize important patterns in the data in order for them to draw conclusions, a practice that is prone to errors. A systematic review from 2015 [50] found that EHR data is generally too large for manual identification of meaningful patterns. They claim that visualizations can help with this. However, they need to be interactive and handle several challenges, such as missing or incorrect data. In this work, we implemented several visualizations of EHR and OCT data into our CDSS and evaluated their usability, informedness, and efficiency aspects.

2.2 Trust in AI-driven Support

Although AI models can be very performant, they are not easily accepted by users. Especially in healthcare, physicians are hesitant to accept AI because of challenges such as inefficient incorporation into workflow and low initial trust in the system [28, 47]. Chen et al. [11] conducted an online questionnaire and found that 78% of their participants, consisting of doctors and medical students, agreed that AI would boost medicine. They identified seven factors contributing to the willingness to use AI, where accuracy, ease of use, and efficiency were most mentioned. Notably, only 64% of participants mentioned interpretability as an important factor. However, this study fails to analyse, whether doctors trust AI systems to begin with. Moreover, the participants were not presented with an actual AI system, rendering the results highly theoretical.

Juravle et al. [26] found in a series of online questionnaires that patients trust AI less than human doctors, which is in line with the general consent that humans prefer other humans over algorithms, also called algorithm aversion [16]. Task-dependent algorithm aversion describes the idea that the distrust in these systems depends on the type of task they are used for. Castelo et al. [10] found that aversion decreases with increasing objectiveness of the task. This suggests, that AI components' trustworthiness might vary based on their task. However, to the best of our knowledge, there are no studies examining the difference in trust in different AI components.

3 A Clinical Decision Support System in Ophthalmology

In this section, we explain the data set, AI components, the composition of visual components, and the visualization techniques used for the creation of our CDSS. The CDSS was implemented using streamlit[3]. All visualizations are made using Plotly[4]. The created graphs are interactive in the sense that the user can pan, zoom, select, and deselect data using the legend and hover over data points to get further information. The CDSS consists of six different Visual Components (VCs 1-6), which each display different data (see Figure 1) and a sidebar, which was used for patient selection. A video demonstration of the dashboard is available in the supplementary materials. The code for the CDSS is published on GitHub at: https://github.com/DFKI-Interactive-Machine-Learning/ophthalmo-cdss.

3.1 Data

For our CDSS, we used real-world clinical data ranging from 1993 to 2023. The data stems from two eye clinics in Germany. It includes the data of 913 patients with AMD and 461 patients with DR. The data can be separated into two categories: EHR and OCT data. EHR data includes all annotations done by the doctors before, during, and after patient visits. It includes measurements like VA, but also SNOMED-CT[5] codes. The codes have been analyzed using natural language processing techniques by a third-party company.

[3] https://streamlit.io/
[4] https://plotly.com/
[5] Systematized Medical Nomenclature for Medicine–Clinical Terminology

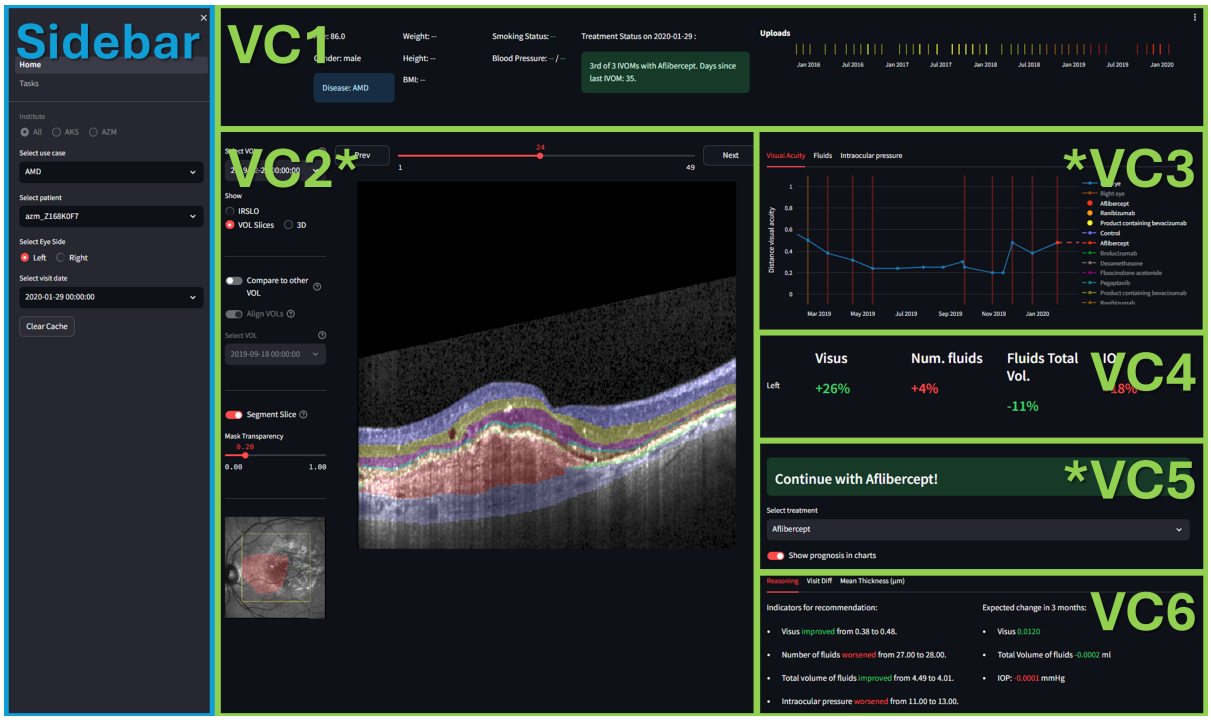


Figure 1: VCs of the evaluated CDSS. VCs that contain AI components are marked with an asterisk.

This led to 3192 different annotation features. Some of this data does not directly relate to diseases, such as age, gender, or smoking behavior, and will be called metadata in the following sections. The OCT data includes 53,410 OCTs, of which 45,389 were quantified using the algorithm described in section 3.2.1.

3.2 AI Components

The AI components can be divided into three categories depending on their task: Segmentation, Classification, and Time Series Forecast. We selected these tasks as they are highly relevant to ophthalmologists and illustrate varying degrees of input-output discrepancy. For semantic segmentation, the input (e.g., an image) and the output (e.g., a mask) are closely related, as the resulting mask can be overlaid on the input image. Treatment recommendation involves a more complex relationship, as it synthesizes diverse inputs, such as OCT scans and EHR data, to generate a recommendation that is markedly different from the input. Time series forecasting exhibits the greatest discrepancy, as it expands the discrepancy of the recommendation by predicting future values, introducing an additional layer of abstraction.

3.2.1 Segmentation and Quantification. For the segmentation task, we used a YNet architecture [18], which was trained to segment images into eleven classes, which are shown in table 1. The training and test set contained 1023 and 400 images, respectively. The model has an overall mean dice score of 0.66, and specifically for fluids, it also has a dice score of 0.66. The quantification directly depended on the segmentation. It used the segmented slices to reconstruct lesions in three dimensions by going through them iteratively

Table 1: Class dependent Dice scores of the trained YNet model

Class	Dice
IPL	0.91
OPL	0.78
ELM	0.55
EZ	0.46
RPE	0.56
BM	0.52
Choroidea	0.85
Drusen	0.33
PED	0.62
Fluids	0.66
Background	0.98

and connecting lesion areas that lie within a distance of 50 μm. Consequently, we get multiple point clouds per lesion, which were then reconstructed to a volume by computing their convex hull using the quickhull algorithm [2]. For the retinal layers, we only reconstructed the surface pointing to the top of each layer. Through this reconstruction in three dimensions, we were able to create a 3D visualization and quantify the thickness of layers and volumes of lesions.

3.2.2 Time Series Forecast. To predict patients' developments and forecast critical points for therapeutic intervention, we trained several BiLSTM models on the available EHR data as well as the quantifications of the OCTs. We forecasted VA developments for one, three, six, nine, and twelve months and trained a model for each time target. For

Table 2: Performance of the time series forecasting model.

Forecast time	MAE	STD
1 month	0.248	0.173
3 months	0.226	0.137
6 months	0.224	0.151
9 months	0.230	0.158
12 months	0.269	0.153

the training, we used the data of 1100 patients (80%) and created between 51,098 and 58,767 data points depending on the forecasted time using a sliding window approach. We always included the last twelve visits each time, and no incomplete windows were used. The missing data was addressed using moving average interpolation on numeric variables, such as VA or fluid volume. In this method, the interpolated data point is positioned at the center of the moving average window, with weights decreasing as the distance from the center increases. Categorical data was imputed by using nearest neighbour imputation. Table 2 shows the performance of our models on the test set of 274 unseen patients containing between 10,778 and 16,290 data windows.

3.2.3 Classification. The classification task entailed a treatment recommendation. It can be seen in two granularity levels: First, the model decides whether the patient should be treated, and second, which medication should be used. The model was realized using if-then-else conditions, which were modeled after clinical guidelines. The flowchart of this algorithm can be seen in figure 2. Although this implementation is not an ML model, it is based on computations from the aforementioned models, and additionally, study participants were not told how the model works beforehand. The classification model's agreement with historical data was about 60% on the treatment decision and 85% on the medication decision. The low agreement on whether to administer IVOMs might be impacted by the fact that patients do not strictly follow the therapy plan perfectly, but postpone treatment for various reasons.

3.3 Workflow Assessment & Design Rationale

The design of our dashboard is loosely inspired by the approach outlined by Bhattacharya et al. [3]. To refine this foundation, we conducted a preliminary workflow assessment involving an assistant doctor and an expert from one of the clinics. During this process, the doctor provided a detailed overview of their workplace and software. The workspace consists of an examination room equipped with a desktop workstation featuring two 27-inch monitors. The software includes various programs for examining EHRs and medical imaging. Moreover, we asked them to describe their usual workflow. Three key tasks were identified, which can be seen in table 3. The doctor was then asked to describe a perfect CDSS tackling these tasks. A low-fidelity prototype was developed in Word [6] and was refined in one feedback iteration. An image of the prototype can be seen in the appendix.

3.4 Visual Components

We developed six VCs to address the challenges identified during the workflow assessment and low-fidelity prototyping process, as detailed in Table 3. These components are designed to enhance the current solutions by providing better visualization, data interpretation, and decision-making support. Below, we describe each VC and its functionality in detail:

- **VC1: Metadata top bar**. Metadata is displayed in this VC. It serves as a quick overview of general patient information, treatment status, and IVOM history.
- **VC2: OCT Viewer**. We visualize the OCT data in this VC. Users can look at the IRSLO and the slices of the OCT or the 3D reconstruction. This VC comes with functionality for segmentations and comparisons for the first two visualizations. The OCT slices can be fully segmented and compared against slices from an older OCT, which can be seen in figure 3a. In the IRSLO, lesions can be segmented and it can be compared against older IRSLOs. Additionally, the layer thickness can be inspected, which can be seen in figure 3b. The 3D view, shown in figure 3c, shows the 2D layers for the surface of retinal layers and 3D volumes for lesions. An OCT slice is visualized inside the 3D model such that users can compare the model to the actual OCT.
- **VC3: Line graphs**. Important metrics such as VA, the volume of fluids (a quantification from the OCT segmentation, see section 3.2.1), and intraocular pressure are displayed in line graphs in this VC. Additionally, a dotted line represents the forecast model's prognosis of VA given the selected treatment. Vertical, colored lines represent IVOM interventions.
- **VC4: Quick Metrics**. Color-coded percentages show the change from last to current visit for the VA, the volume of fluids, and intraocular pressure, whereas red indicates negative influence and green positive influence on disease parameters.
- **VC5: Recommendation**. In this VC, we display the treatment recommendation based on clinical guidelines (see section 3.2.3). Recommendations are color-coded, such that green represents therapy, red represents aborting therapy, yellow represents that more information is needed (e.g. when the last medical image was taken more than a month ago), and blue represents that no therapy is needed.
- **VC6: Diverse Utility**. In this VC, we display less important data from the EHR and 3D reconstruction in three different tabs. The first tab called "Reasoning" shows the change from last to current visit, but also the expected change in three months of our forecast model. Visit Diff utilizes other annotations from doctors for symptoms like bleeding, edema, and more.

[6] https://www.microsoft.com/de-de/microsoft-365/word?market=de (Accessed: 09.01.2025)

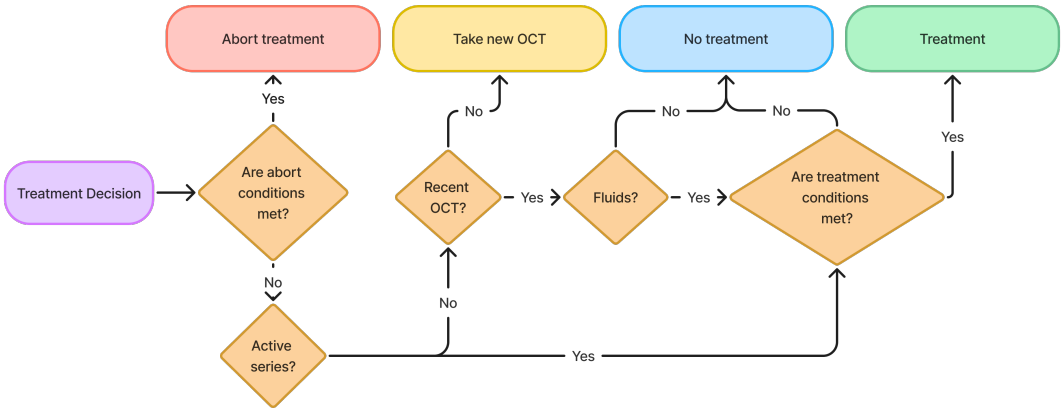


Figure 2: Flowchart of the treatment recommendation algorithm based on clinical guidelines.

Table 3: Task analysis and suggested improvements.

Task	Current Solution	Problem	Improvement	Involved VC
Assessment of Therapy Status & Patient Information	Scroll through EHR to assess information	Tedious scrolling, Risk of missing crucial information	Display metadata; Visualize important metrics such as VA in line graphs; Give a quick overview for changes from last to current visit	VC1, VC3, VC4, VC6
OCT Analysis	Separate program; Multiple instances needed for comparison	Confusing, not easy to compare; No support for determining fluid levels	Integrated OCT Viewer with comparison functionality and automatic segmentation and quantification	VC2, VC3
Treatment Decision	Subjective decision making	No support from software	Decision support through recommendation based on clinical guidelines	VC5

The Mean Thickness tab shows a table of the mean thickness of each layer from the previous and the current OCT.

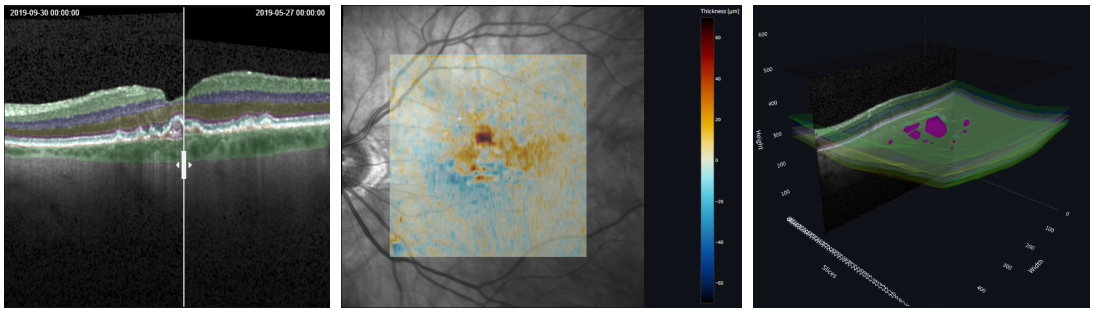
4 Evaluation

In this section, we present the user study and its evaluation. We start with an overview of the demographics of the participants, followed by a detailed description of the study procedure and the apparatus used. Finally, we outline the evaluation method applied to analyze the results. The transcripts and codes can be made available at reasonable request.

4.1 Participants

Eleven ophthalmologists (three female, 27%; eight male, 73%) participated in the study on three different days. Nine were assistant doctors with less than five years of experience (**A1-9**), one was a specialized ophthalmologist with five years of experience (**O**), and one was a senior ophthalmologist with 21 years of experience (**S**). All women were among **A**. **A1** claimed to have ten years of experience. However, they classified themselves as **A**, which leads to the suspicion that the stated experience in years is wrong. Furthermore, **A7** missed filling out the questionnaire, which is why we do not have sentiment on AI data or SUS ratings.

Participants rated their sentiment toward AI, their experience with AI, and their experience with software using a five-point Likert scale, where 1 represented strong opposition or no experience, and 5 represented strong support or advanced experience. The results of this questionnaire are presented in Table 4. The data reveals a generally positive sentiment toward AI. However, experience with AI is moderate overall, with only **S** and **A3** reporting advanced experience. In contrast, participants rated their experience with software as above average, except **A2**, **A5**, and **A6**, who indicated intermediate experience.



(a) Slice with Segmentation (b) IRSLO with Thickness Map (c) 3D Reconstruction

Figure 3: Examples of the three different views of VC2.

Table 4: Summary of the questionnaire answers. Sentiment and Experience with AI and software were rated on a Likert Scale from 1 (negative sentiment or no experience) to 5 (positive sentiment or advanced experience). The experience in years marked with an asterisk (*) is expected to be a mistake. Additionally, one A did not submit the questionnaire (/).

Code	Experience in years	Sentiment towards AI	Experience with AI	Experience with software	SUS
S	21	5	5	4	85
O	5	4	3	5	90
A1	10*	4	3	5	90
A2	4	4	3	3	62.5
A3	4	4	4	4	82.5
A4	3	4	3	4	95
A5	3	5	3	3	85
A6	1.6	5	3	3	67.5
A7	1.2	/	/	/	/
A8	1	4	2	4	77.5
A9	0.9	5	3	4	82.5
avg.	4.47	4.40	3.20	3.90	81.75

4.2 Study procedure

We evaluated the dashboard with eleven doctors of varying levels of experience in ophthalmology. Before starting, participants were informed about the study procedure and asked to sign informed consent forms. They were also instructed to think aloud during the session. For a full study guide, see appendix B. The study procedure was as follows:

- **Tutorial:** Participants were briefly introduced to the various components of the CDSS. Then they were encouraged to explore the CDSS and ask any questions about its usage. The tutorial should familiarize participants with the CDSS.
- **Task 1 (T1):** The first task involved analyzing metadata and the therapy status. Participants were asked to provide information about the patient that might impact therapy. They were also asked to specify how many and which IVOMs the patient had received and whether these treatments were part of a series.
- **Task 2 (T2):** The second task required the doctors to analyze OCT data. Specifically, they were asked to identify key biomarkers and describe trends in the patient's condition by comparing different OCTs. They were also instructed to review each mode of the OCT viewer and its functionalities (see Section 3.4).
- **Task 3 (T3):** Finally, in the third task, participants were asked to make a therapy decision using any functionality of their choice.

These tasks were developed based on the workflow assessment interview (see section 3.3) and represent scenarios that could occur in clinicians' daily workflows. We opted for a task-based approach due to its ability to provide valuable insights as demonstrated by Bhattacharya et al. [3]. The patient data shown was preselected to ensure that each task features a different patient with all relevant information available. Additionally, for tasks 2 and 3, we selected patients with high-quality OCTs. All participants saw the same data for each task. After each task, participants were asked a fixed set of yes or no questions: "Did the CDSS improve your efficiency, level of informedness, and user experience compared to your current system?". Additionally, we conducted semi-structured interviews during the completion of the tasks to explore relevant themes, such as the trustworthiness of AI components, in more depth. The idea of a feedback system was always brought up whenever participants spotted an error in a prediction or after T2

when looking at the segmentation of VC2. The interviewers asked whether a feedback system could improve trust and what this system should look like.

At the end of the session, participants completed a questionnaire to get demographic data and their sentiment towards AI (See section 4.1. Additionally, they filled in the System Usability Scale (SUS). The SUS is a ten-item questionnaire that rates the usability of a system using Likert scales [8]. We used the SUS to confirm our interview findings and obtain an objective measure of the system's usability.

4.3 Apparatus

The study was conducted in an examination office at the eye clinic of Sulzbach[7]. Participants accessed the CDSS via a laptop. A 27-inch monitor displaying a web page interface of the CDSS and a computer mouse were connected to the laptop for ease of use and to mimic their usual setup. Audio of interviews was recorded for later transcription. Two interviewers were present: one conducted the study, while the other observed, took notes on notable behaviours, and assisted as a co-interviewer.

4.4 Thematic Analysis

Thematic analysis is a qualitative research method used to identify, analyze, and report patterns or themes within data. It was introduced by Clarke and Braun in 2006, who have continually refined and improved the method [5–7]. This approach is particularly effective for interpreting complex textual data, such as interview transcripts, 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 the codes, reviewing and refining the themes, and producing the final report. In this study, thematic analysis was conducted on the interview transcripts by two authors in a joint coding session. Since thematic analysis is inherently subjective, Cohen's Kappa (κ) [14] was used to evaluate the level of agreement between two independent raters, each of whom classified the codes into themes and assigned text passages to codes. Unlike simple percentage agreement, Cohen's Kappa accounts for agreement occurring by chance, providing a more accurate measure of interrater reliability.

5 Results

In the following section, we present the results of the user study, starting with the results of the questionnaire and the SUS, then show the answers to the fixed set of questions and, finally, show the uncovered codes and themes of the thematic analysis.

5.1 Systems Usability Scale

The average SUS score from ten participants was 81.75, with a standard deviation of 10.1, indicating good usability [1]. The individual SUS scores can be found in table 4. The highest rating was 95 (**A4**). **A2** and **A6** with intermediate software knowledge (i.e. 3 on the Likert scale) gave the lowest

[7]https://www.augenklinik-sulzbach.de/ (Accessed: 05.01.2025)

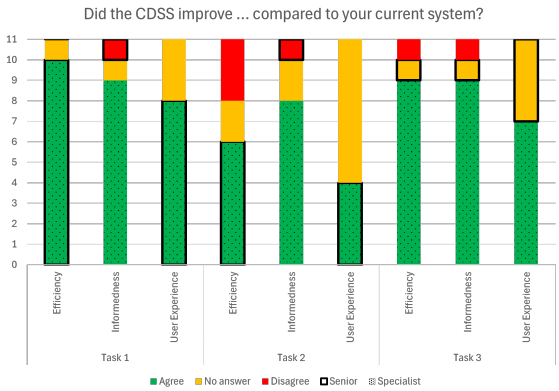


Figure 4: Agreement to the question of whether the CDSS could improve efficiency, informedness or usability for the specific task. Answers from S and O are marked by bold frames and dotted patterns, respectively.

ratings of 62.5 and 67.5, while those who rated themselves as having advanced or expert knowledge (i.e. 4 and 5 on the Likert scale) gave generally higher ratings. They also rated their need for a technical person to help them use the dashboard higher than all others (i.e. 3 and 4 on the Likert scale). **O** and **S** gave ratings of 90 and 85, respectively, which is above the average rating among all participants.

5.2 Task Ratings

A fixed set of yes or no questions was asked after the completion of each task, whose results can be seen in figure 4. Although the questions were yes or no in nature, many times, participants did not answer with yes or no. Hence, through coding, we categorized their answers into yes, no, or no sufficient answers.

For **T1**, all except **A2**, who did not sufficiently answer, felt more efficient than usual. **S** did not feel more informed, and **A7** did not give a clear answer, while the remaining participants felt more informed. Furthermore, **A4**, **A6**, and **A7** did not give a clear answer to whether their user experience was improved. The rest felt like their user experience was improved.

The analysis of the OCT data from **T2** was rated as more efficient by everybody but **A2-4** and **A6-7**. **A2** and **A4** did not give clear answers to any of the questions. **A3** and **A6-7** did not feel more efficient. **S** did not feel more informed, while the rest, that sufficiently answered, did. For user experience, **A2-8** did not answer clearly, while the rest felt that it improved.

For **T3**, **S** did not answer clearly for all three aspects of this task. Everybody but **A3** felt more efficient. Furthermore, only **A2** said they did not feel more informed, while the rest did. Everybody but **S**, **A3**, and **A5-6** felt like the user experience was improved. **S**, **A3** and **A5-6** did not give a clear answer here.

5.3 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. They were then categorized 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 and transcript snippets to codes had Cohen's κ of 0.91 and 0.57, respectively. This shows that raters agree with the given categorization and coding of the transcripts. Quotes are translated from the German transcripts, and only grammar is corrected. Parts that were already included in section 5.2 were not considered during the thematic analysis unless they contained information which goes beyond what was already established.

5.3.1 RQ1: Are segmentation, classification, and time series forecast models perceived with varying levels of trustworthiness? We found that overall, everybody mentioned low perceived trustworthiness of the system in some way. The most common reason for distrust was a lack of experience with the CDSS. This was highlighted by everyone except **O**, **A3**, and **A5-6**. As **A1** explained: "*The more I would use it, the more I would get a feeling for what the program is doing. And then it would increase my trust.*". **A1-4**, **A6-7** and **O** mentioned that the system needs to provide the raw data to be controllable: "*[...] If I only saw this [quantification] data [without the OCT], then I would be missing a control opportunity for the system*" (**A4**); "*As I have seen from this patient, I could trust the system. [...] But [...] I would always try to clinically comprehend it.*"(**O**). All participants agreed that they would never rely on the system but primarily on their skills and the clinical guidelines.

A6 did not initially trust the segmentation: "*It seems useful [...], but [...] I would not rely on it*". **S** and **A1-2** mentioned low trust in the 3D view after discovering errors, which also implies mistrust in the segmentation: "*If there was a mistake in the computation, [it will make] the same mistake for each [OCT]. [...] I would not use the 3D view to make a decision. I would use the OCT slices.*" (**S**). Other participants did not mention low trust in the segmentation. We proposed the idea of a feedback system to the participants, where they could provide feedback when dissatisfied with the segmentation. **S**, **O** and **A1**, **A3**, **A6** and **A8-9** said that this would increase trust. However, **S** and **A1** said that such a system would not be used due to clinicians' time constraints: "*The time just is not there*" (**S**). Furthermore, a federate feedback system was mentioned as beneficial for trust by **A3**, **A6**, **A8** and **O**: "You would have an additional control, which would be good." (**A8**).

All participants initially checked treatment parameters themselves before consulting the recommendation implying a low perceived trustworthiness of this component. **A1**, **A4**, and **O** mentioned they trusted the recommendation after validating it but would keep cross-checking. **A2**, **A5**, and**A7** said they do not trust the recommendation because they do not understand how it works.

The forecast model was widely ignored in the treatment decision. Furthermore, **S** had strong negative sentiment towards the forecast: "*That is total bullshit. [...] How can someone make a prognosis over clinical findings? That is basically like reading coffee grounds*". **A7** echoes similar concerns: "*One can maybe use it as an idea [...], but it is too individual [...], how people react to therapy*".

After an explanation of the treatment recommendation, the forecast model and their training processes, **A1**, **A3-4** and **A9** said that their trust in the recommendation system was increased: "*If I have a system, where I know it makes decisions similarly to myself, then that makes it easier to rely on it*" (**A4**). **A9** was the only participant who mentioned including the forecast in their decision-making after the explanation.

5.3.2 RQ2: How can an ophthalmologist's efficiency, informedness, and user experience be improved using a CDSS?. Everybody but **A2** mentioned that they felt more efficient using the dashboard in some way. **A1**, **A3-4**, **A6**, **A8** and **O** mention the improved metadata display as a reason for this: "*[In the EHR] you have to search significantly longer than here*" (**A4**). Moreover, **A1**, **A4-9** and **S** say that the integration of OCT and EHR into one clear overview would bring an efficiency improvement: "*I find the most useful, that I have a complete overview, such that I do not need to open the OCT in a separate program.*" (**A9**). **A5** and **A9** also say that agreement with the recommendation would improve efficiency. **A1**, **A7** and **A9** mentioned that having to read unnecessary, non-critical factors decreases efficiency: "*The intraocular pressure is not relevant for treatment [...][and it is] an additional information I have to read, and that costs me time*" (**A1**). For OCT analysis, **A3** and **A6-7** said that the CDSS brings no efficiency bonus if we disregard the loading times and the missing "Scrolling through slices" feature.

In the informedness category, all except **A7** found that the dashboard provides them with more information than they would usually have. This is due to the practical 3D view (**A1** and **A9**), the quantifications (**A1-4**, **A6**, **A8** and **S**), the segmentation (**A1**, **A3**, **A5**, **O** and **S**) and the forecast (**A9**): "*I think [the 3D view] is handy*" (**A9**), "*Such solid data [...] that is going to be the future [of indication]*" (**S**), "*I can see [multiple biomarkers] with the segmentation much better, than without.*" (**A5**). **A3**, **A5-7**, **A9** and **O** said that the CDSS is particularly helpful in borderline cases: "*For many patients, it is fairly obvious, but especially in those borderline cases [...] I think it is useful.*" (**A9**). Furthermore, **A3**, **A6-7** and **A9** think that disagreement with the CDSS would push for a more thorough analysis: "*This brings a certain treatment quality if the system says something else than I do. Then I can cross-check and maybe consult someone with more experience than me.*" (**A3**). More information was wished for regarding drug prices (**A1**), reason for drug switches (**A2**), number and times of IVOMs (**A9**, **O**) and concurrent diseases or surgeries (**A5**, **A6**).

Finally, everybody but **A2** mentioned some improvement in user experience. Similar to efficiency, this was largely due to the integration of OCT and EHR data, but also due to other factors such as intuitiveness of the CDSS (**A1**, **O**), easy comparison of scans (**A1**, **A5**, **A9**, **O** and **S**), good looking

Table 5: Trustworthiness codes and which participants mentioned them.

Code	Participant
Relies primarily on skills and guidelines	A1-9, O, S
Recommendation initially ignored	A1-9, O, S
Forecast initially ignored	A1-9, O, S
Experience builds trust	A1, A2, A4, A7, A8, A9, S
Controllable via raw data	A1, A2, A3, A4, A6, A7, O
Feedback increases trust	A1, A3, A6, A8, A9, O, S
No trust in segmentation	A1, A2, A6, S
Explanation of model enables trust	A1, A3, A4, A9
Manual agreement improves trust	A1, A4, O
Lack of understanding reduces trust	A2, A5, A7
Feedback not adoptable due to time	A1, S
Forecast cannot work	A7, S
Negative sentiment towards forecast	S
Would include forecast in decisions	A9

visualizations, which can also be shown to patients (**A1**, **A3**) or positive reinforcement through the recommendation (**A3**, **A7**). As a feature, everybody but **A4-5** and **A8** missed scrolling through the slices.

6 Discussion

In this section, we discuss the results of our user study, relate them to current literature, and answer the proposed RQs. Then, we propose guidelines for designing trustworthy and efficient AI-driven CDSS. Finally, we cover limitations and future work.

6.1 RQ1: Are segmentation, classification, and time series forecast models perceived with varying levels of trustworthiness?

In this study, we examined the impact of an AI components task domain on the trust of experts. Segmentation was the least initially distrusted AI component, although mistakes in this component made users lose trust quickly, as demonstrated by **S**: "*If there was a mistake in the computation, [it will make] the same mistake for each [OCT]*". Although it has been known that errors decrease trust in AI models much more than they would in humans [17], this also indicates, contrary to Langer et al. [31], that doctors might view AI as algorithms with defined rules, which it cannot deviate from. This point of view hinders the adoption of AI components. We argue that CDSS should implement reminders that highlight the non-deterministic nature of AI and that one error does not conclude the model to be inherently erroneous.

Since the recommendation was not used until after participants had formed their own decision, we think that the recommendation was widely distrusted. However, this could also be attributed to not being used to the CDSS, as six **A** and **S** said that they would only build trust by using the CDSS and cross-checking whether they align with its decisions. **A2**, **A5**, and **A7** said that they cannot trust the recommendation because they do not understand it, which again shows a missing understanding of how AI works to be an issue for trust. Provided an explanation of the model and its training, trust increased for four **A**. It seems especially important for experts to know that the set of considered features was complete. Hence, we can support the findings of Bhattacharya et al. [3] or Cai et al. [9], that users need to know about the AI's training process and data to build a mental model of it.

The recommendation was distrusted at first, but participants were open to trusting it in the future, given it performs as they expect. The forecasting component, however, while also being widely ignored for completion of **T3**, received negative sentiment by **S**. Additionally, **S** and **A7** mentioned that its goal is impossible and, hence, we find its perceived trustworthiness is worse than the recommendations. We argue that the abstraction from input to output plays a role in how trustworthy AI components are perceived. The degree of abstraction between OCT slices and segmentation masks is minimal, which is why participants seem to trust it more. Expanding to a 3D reconstruction, segmentation already loses trust, as now the abstraction degree is larger. The recommendation abstracts from EHR data and OCT quantifications of one visit to a treatment decision and, hence, receives less trust, as again abstraction increases. Finally, the forecasting expands this by also including past visits and predicting a metric for a future date. Consequently, it has the largest degree of abstraction and the worst perceived trustworthiness of the evaluated AI components. Hence, this finding identifies task-dependent aversion [10] in the degree of abstraction of AI components.

Feedback systems were mentioned as beneficial for trust, but concerns about their usage due to time constraints were raised. This is contrary to Honeycutt et al. [23], who found that giving feedback decreases trust in the system. However, in this study, users were forced to give feedback on errors, while in our study, they had the option to give feedback. We argue that the inconvenience of having to deal with system errors decreases trust. According to our study participants, federated feedback would increase trust because users feel like the AI is being controlled. These systems must be optional and fast to not be a burden for the user. The option to give more detailed feedback should also be given.

To conclude, we propose three guidelines:

- **Remind users of the non-deterministic nature of AI components**: Users might associate AI with erroneous algorithms and develop algorithm aversion. Reminding them that AI is non-deterministic might mitigate this issue.
- **Explain AI components with a large degree of abstraction**: Models, whose output is significantly abstracted from their input, must be explained to a degree where users can easily verify the correctness of the prediction.
- **Implement quick feedback options**: Provide options for feedback that can be used swiftly without interrupting clinicians' workflow. Mandatory feedback might be perceived as an inconvenience and, hence, decreases trust, while optional feedback gives a feeling of control over the system without being a burden on the user.

6.2 RQ2: How can an ophthalmologist's efficiency, informedness, and user experience be improved using a CDSS?

We found that the developed CDSS has improved efficiency, informedness, and user experience in multiple aspects.

The integrated display of OCT data and the visualizations of EHR data decrease the time and effort needed for analysis. We argue that CDSS should integrate standard functionality and AI components to improve efficiency. However, this also represents a barrier for researchers and developers, as integration into existing systems is rarely possible, and integrating the entire workflow means additional development work.

As anticipated, participants felt more informed due to the visualization of the EHR data. Additionally, AI components were mentioned as helpful for improving patient care. The segmentation locates regions of interest and helps with borderline cases. The recommendation motivates users to cross-check with more experienced clinicians when disagreeing with the system. However, the forecast was not positively received. User experience also improves when efficiency or informedness improves. Additionally, our study found that good-looking visualizations, positive reinforcement of decisions, and an intuitive system enhanced users' experience.

In conclusion, the main perceived improvement of our CDSS over clinicians' standard setup was the integration of EHR and OCT data, as well as the visualization of EHR data. While other factors also influenced efficiency, informedness, and user experience, it is hard to generalize these findings. Especially the influence of AI components needs to be studied further. Our CDSS reaches an average SUS score of 81.75 and, hence, according to Bangor et al. [1], has good usability. We provide the code of our CDSS as a foundation for future research at: https://github.com/DFKI-Interactive-Machine-Learning/ophthalmo-cdss.

6.3 Limitations

This study provides an overview of key considerations when designing an AI-supported CDSS that integrates into the therapy workflow of ophthalmologists. However, there are several limitations to our study:

(1) **Indirect comparison**: It was not possible to integrate our prototype into a real hospital information system and compare it with clinicians' actual setup. Instead, our study reveals only perceived improvements, which might not exist. Additionally, it lacks objective measurements, such as completion times of tasks.
(2) **Imperfect system**: Our CDSS came with several issues, such as slow processing times and missing fine-tuning of ML models. Both could have affected usability and trust ratings, as demonstrated by the frequent wish for "scrolling through the OCT slices", a feature that the current setup supports but ours could not do performantly.
(3) **Replicability**: Our data is not publicly available and, hence, it is difficult to re-run the experiments.
(4) **Generalizability**: While our guidelines should facilitate trust in AI components of CDSS, it is unclear how well they generalize to other fields of medicine. Also, all our participants come from one institution.
(5) **Missing quantifications**: We did not quantify our findings using questionnaires. For example, the trustworthiness of AI components could be quantified using the Cahour-Forzy scale questionnaires [22]. Additionally, it remains unclear which specific explanation methods are most effective in increasing trust for different AI components.
(6) **SUS limitation**: The SUS provides only a general assessment of usability and does not capture specific usability issues or nuances in user experience, such as the ease of learning or efficiency in specific tasks. Additionally, the SUS scores may be influenced by users' familiarity with similar systems or their biases toward AI-based tools, which could skew the results.

6.4 Future Work

In future work, we hope to address the limitations while also exploring ways to properly explain different AI components to achieve maximal perceived trustworthiness without making doctors overly reliant on the system. The results of such a study could offer valuable insights for designing more trustworthy and user-friendly CDSS, ultimately improving adoption and clinical outcomes. With further development, our tool could be integrated into hospitals and compared to standard setups to further quantify the impact.

7 Conclusion

In this study, we developed an AI-driven CDSS to support ophthalmologists in the treatment of AMD and DR. Through semi-structured, task-based interviews with eleven ophthalmologists we evaluated our CDSS in terms of efficiency, informedness, and user experience. Additionally, we explored the trustworthiness of AI-driven segmentation, treatment recommendation, and VA forecasting.

Our CDSS was perceived as more efficient, more informative, and more usable by the study participants compared to their standard setup. Major factors were the integration

of OCT and EHR data, visualizations of EHR data, and fluid quantification. Additionally, we identified key factors influencing the trustworthiness of AI in CDSS: Users need to be reminded of the non-deterministic nature of AI to not lose trust when discovering errors, a large degree of abstraction from input to output decreases perceived trustworthiness, and feedback can increase trust, when it is fast and optional.

We highlighted the limitations of our work and discussed potential future research directions. Additionally, we made our CDSS publicly available as a foundation for future developments and research on GitHub at: https://github.com/DFKI-Interactive-Machine-Learning/ophthalmo-cdss.

Acknowledgments

This work was funded, in part, by the German Federal Ministry of Education and Research (BMBF) under grant number 16SV8639 (OphthalmoAI) and grant number 01IW23002 (No-IDLE).

We also thank Abdulrahman M. Selim for his thorough proofreading of our paper.

References

[1] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[2] C. Bradford Barber, David P. Dobkin, and Hannu Huhdanpaa. 1996. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.* 22, 4 (dec 1996), 469–483. doi:10.1145/235815.235821

[3] 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*. 204–219.

[4] Aditya Bhattacharya, Simone Stumpf, Lucija Gosak, Gregor Stiglic, and Katrien Verbert. 2024. EXMOS: Explanatory Model Steering Through Multifaceted Explanations and Data Configurations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–27.

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[6] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[7] 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.

[8] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* 189 (11 1995).

[9] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[10] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.

[11] Mingyang Chen, Bo Zhang, Ziting Cai, Samuel Seery, Maria J Gonzalez, Nasra M Ali, Ran Ren, Youlin Qiao, Peng Xue, and Yu Jiang. 2022. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. *Frontiers in medicine* 9 (2022), 990604.

[12] Stephanie J Chiu, Michael J Allingham, Priyatham S Mettu, Scott W Cousins, Joseph A Izatt, and Sina Farsiu. 2015. Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. *Biomedical optics express* 6, 4 (2015), 1172–1194.

[13] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555 [cs.NE] https://arxiv.org/abs/1412.3555

[14] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[15] William H. Dean, Susannah Grant, Jim McHugh, Oliver Bowes, and Fiona Spencer. 2019. Ophthalmology specialist trainee survey in the United Kingdom. *Eye* 33, 6 (June 2019), 917–924. doi:10.1038/s41433-019-0344-z Publisher: Nature Publishing Group.

[16] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144, 1 (2015), 114.

[17] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144, 1 (2015), 114.

[18] Azade Farshad, Yousef Yeganeh, Peter Gehlbach, and Nassir Navab. 2022. Y-Net: A Spatiospectral Dual-Encoder Networkfor Medical Image Segmentation. arXiv:2204.07613 [eess.IV] https://arxiv.org/abs/2204.07613

[19] Donald S Fong, Lloyd P Aiello, Frederick L Ferris III, and Ronald Klein. 2004. Diabetic retinopathy. *Diabetes care* 27, 10 (2004).

[20] David Gefen, Elena Karahanna, and Detmar W Straub. 2003. Trust and TAM in online shopping: An integrated model. *MIS quarterly* (2003), 51–90.

[21] S Hochreiter. 1997. Long Short-term Memory. *Neural Computation MIT-Press* (1997).

[22] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[23] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 63–72.

[24] 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, et al. 1991. Optical coherence tomography. *science* 254, 5035 (1991), 1178–1181.

[25] Cheng Jin, Heng Yu, Jia Ke, Peirong Ding, Yongju Yi, Xiaofeng Jiang, Xin Duan, Jinghua Tang, Daniel T Chang, Xiaojian Wu, et al. 2021. Predicting treatment response from longitudinal images using multi-task deep learning. *Nature communications* 12, 1 (2021), 1851.

[26] Georgiana Juravle, Andriana Boudouraki, Miglena Terziyska, and Constantin Rezlescu. 2020. Chapter 14 - Trust in artificial intelligence for medical diagnoses. In *Progress in Brain Research*, Beth Louise Parkin (Ed.). Real-World Applications in Cognitive Neuroscience, Vol. 253. Elsevier, 263–282. doi:10.1016/bs.pbr.2020.06.006

[27] Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag. 2023. Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*. 000111–000124. doi:10.1109/INES59282.2023.10297629

[28] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. 2018. Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis. *JMIR Medical Informatics* 6, 2 (April 2018), e8912. doi:10.2196/medinform.8912 Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada.

[29] Eric S. Kim, Nansook Park, Jennifer K. Sun, Jacqui Smith, and Christopher Peterson. 2014. Life Satisfaction and Frequency of Doctor Visits. *Psychosomatic Medicine* 76, 1 (Jan. 2014), 86. doi:10.1097/PSY.0000000000000024

[30] Samuli Laato, Miika Tiainen, AKM Najmul Islam, and Matti Mäntymäki. 2022. How to explain AI systems to end users: a systematic literature review and research agenda. *Internet Research* 32, 7 (2022), 1–31.

[31] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J König, and Nina Grgić-Hlača. 2022. "Look! It'sa computer program! It's an algorithm! It's AI!": Does terminology affect human perceptions and evaluations of algorithmic decision-making systems?. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–28.

[32] Zhongwen Li, Lei Wang, Xuefang Wu, Jiewei Jiang, Wei Qiang, He Xie, Hongjian Zhou, Shanjun Wu, Yi Shao, and Wei Chen. 2023. Artificial intelligence in ophthalmology: The path to the real-world clinic. *Cell Reports Medicine* 4, 7 (2023).

[33] Laurence S Lim, Paul Mitchell, Johanna M Seddon, Frank G Holz, and Tien Y Wong. 2012. Age-related macular degeneration. *The Lancet* 379, 9827 (2012), 1728–1738.

[34] Larry R Medsker and LC Jain. 2001. Recurrent neural networks. *Design and Applications* 5, 64-67 (2001), 2.

[35] Martina Melinščak, Marin Radmilović, Zoran Vatavuk, and Sven Lončarić. 2021. Annotated retinal optical coherence tomography images (AROI) database for joint retinal layer and fluid segmentation. *Automatika : časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije* 62, 3-4 (Oct. 2021), 375–385. doi:10.1080/00051144.2021.1973298 Publisher: KoREMA - Hrvatsko društvo za komunikacije,računarstvo, elektroniku, mjerenja i automatiku.

[36] Mohammad Amin Morid, Olivia R. Liu Sheng, and Joseph Dunbar. 2022. Time Series Prediction using Deep Learning Methods in Healthcare. arXiv:2108.13461 [cs.LG]

[37] Mehmood Nawaz, Adilet Uvaliyev, Khadija Bibi, Hao Wei, Sai Mu Dalike Abaxi, Anum Masood, Peilun Shi, Ho-Pui Ho, and Wu Yuan. 2023. Unravelling the complexity of Optical Coherence Tomography image segmentation using machine and deep learning techniques: A review. *Computerized Medical Imaging and Graphics* (2023), 102269.
[38] Timm Oberwahrenbrock, Ghislaine L. Traber, Sebastian Lukas, Iñigo Gabilondo, Rachel Nolan, Christopher Songster, Lisanne Balk, Axel Petzold, Friedemann Paul, Pablo Villoslada, Alexander U. Brandt, Ari J. Green, and Sven Schippling. 2018. Multicenter reliability of semiautomatic retinal layer segmentation using OCT. *Neurology Neuroimmunology & Neuroinflammation* 5, 3 (May 2018), e449. doi:10.1212/NXI.0000000000000449 Publisher: Wolters Kluwer.
[39] P Jonathon Phillips, P Jonathon Phillips, Carina A Hahn, Peter C Fontana, Amy N Yates, Kristen Greene, David A Broniatowski, and Mark A Przybocki. 2021. Four principles of explainable artificial intelligence.
[40] Abdolreza Rashno, Behzad Nazari, Dara D Koozekanani, Paul M Drayna, Saeed Sadri, Hossein Rabbani, and Keshab K Parhi. 2017. Fully-automated segmentation of fluid regions in exudative age-related macular degeneration subjects: Kernel graph cut in neutrosophic domain. *PloS one* 12, 10 (2017), e0186949.
[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs.CV] https://arxiv.org/abs/1505.04597
[42] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533–536.
[43] 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.
[44] Peter Frederick Sharp, Ayyakkannu Manivannan, Heping Xu, and John Vincent Forrester. 2004. The scanning laser ophthalmoscope—a review of its role in bioscience and medicine. *Physics in Medicine & Biology* 49, 7 (2004), 1085.
[45] 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.
[46] Vikas Tah, Harry O. Orlans, Jonathan Hyer, Edward Casswell, Nizar Din, Vishnu Sri Shanmuganathan, Louise Ramskold, and Saruban Pasu. 2015. Anti-VEGF Therapy and the Retina: An Update. *Journal of Ophthalmology* 2015, 1 (2015), 627674. doi:10.1155/2015/627674 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2015/627674.
[47] Honoka Tamori, Hiroko Yamashina, Masami Mukai, Yasuhiro Morii, Teppei Suzuki, and Katsuhiko Ogasawara. 2022. Acceptance of the Use of Artificial Intelligence in Medicine Among Japan's Doctors and the Public: A Questionnaire Survey. *JMIR Human Factors* 9, 1 (March 2022), e24680. doi:10.2196/24680 Company: JMIR Human Factors Distributor: JMIR Human Factors Institution: JMIR Human Factors Label: JMIR Human Factors Publisher: JMIR Publications Inc., Toronto, Canada.
[48] Devon Watts, Rafaela Fernandes Pulice, Jim Reilly, Andre R Brunoni, Flávio Kapczinski, and Ives Cavalcante Passos. 2022. Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis. *Translational psychiatry* 12, 1 (2022), 332.
[49] Devon Watts, Rafaela Fernandes Pulice, Jim Reilly, Andre R Brunoni, Flávio Kapczinski, and Ives Cavalcante Passos. 2022. Predicting treatment response using EEG in major depressive disorder: A machine-learning meta-analysis. *Translational psychiatry* 12, 1 (2022), 332.
[50] Vivian L West, David Borland, and W Ed Hammond. 2015. Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association* 22, 2 (2015), 330–339.

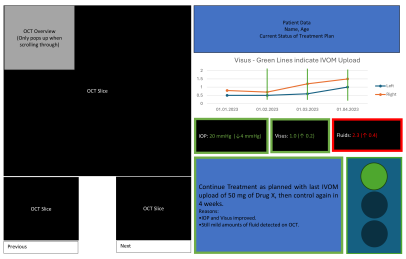


Figure 5: The low-fidelity prototype was developed in collaboration with one doctor from the eye clinic in Sulzbach.

A Low-Fidelity Prototype

B Interview Guide

We conducted a semi-structured task-based interview to find out whether our CDSS could improve efficiency, informedness, and user experience. Additionally, we evaluated the trustworthiness of several AI components. In the following section, we outline the interview process and the instructions provided to participants. All instructions and information given to participants are presented as quotes (translated into English).

B.1 Introduction

We introduced the purpose and the concept of the study to the participants, i.e. gathering participants' feedback and insights into the CDSS and its usability. We explained that they would have to complete three tasks and instructed them to think aloud.

> "Thank you for participating in our study. The purpose of our study is to evaluate a Clinical Decision Support System for ophthalmologists in the treatment of Diabetic Retinopathy and Age-related Macular Degeneration. In the course of this study, you will complete three tasks, which are modeled after tasks, that you would encounter in your workflow. Please always say, what you think, aloud. We will ask you open-ended questions. There are no right or wrong answers; we are interested in your honest opinions and perspectives. Feel free to take your time to consider your responses, and don't hesitate to ask any questions you may have. Is there anything you would like to know before we begin?"

B.2 Tutorial

Before the participants started with the Tasks, they got five to ten minutes to familiarize themselves with the CDSS. One of the interviewers gave a short introduction to the CDSS followed by the participant using the CDSS as they wished while asking questions about its usage:

> "Here you can see the developed CDSS. It separates into six distinct visual components. The first one gives you an overview of metadata and the treatment status. The second one shows

you the OCT data. It has functionality in the toolbox to the left of it, one of which is the segmentation. Segmentation is an AI tool. The third visual component shows line graphs of some selected metrics such as visual acuity. The dotted line represents a forecast of another AI component of the CDSS. The fourth visual component shows how some metrics changed from the last to the current visit. In the fifth, you can see the treatment recommendation, which is another AI tool. The last visual component shows you other relevant data, namely a reasoning for our recommendation, a thickness table, and visit diff. Feel free to explore all of these components on your own now and ask questions, whenever you feel the need to. Remember to always speak your mind, as we are interested in any insight you can provide."

B.3 Tasks

The tasks were modeled after actual tasks from an ophthalmologist's everyday workflow. They were determined in a preliminary workflow study with one ophthalmologist. The patient data shown was preselected to ensure that each task features a different patient with all relevant information available. Additionally, for tasks 2 and 3, we selected patients with high-quality OCTs. All participants saw the same data for each task. After each task, we asked three questions:

(1) "Could the CDSS improve your efficiency regarding this task compared to your usual setup? Why/ why not?"
(2) "Could the CDSS improve your informedness regarding this task compared to your usual setup? Why/ why not?"
(3) "Could the CDSS improve your user experience regarding this task compared to your usual setup? Why/ why not?"

B.3.1 Task 1: Metadata Analysis. The first revolved around metadata analysis. The instructions for the participants were:

> "Imagine you got a new patient. Find out all relevant metadata. Specifically, find out:
> - The patient's gender, age, BMI, and smoking behavior.
> - Which disease does the patient have?
> - How is the patient's visual acuity? How has it changed?
> - Has the patient already received treatment? How often and which medication has been given? Since when has the patient been in treatment? Was treatment effective regarding visual acuity?
> - What else has changed since the last visit?
>
> Please use the CDSS to solve this task. For this task, please ignore visual component 2. If you have any questions or need the task to be repeated, feel free to ask."

After completion, we asked about efficiency, informedness, and user experience as mentioned above. Additional follow-up questions were:

- "Was the data visualized in an intuitive and comprehensible way? Why/ why not?"
- "How would your answers change, if you already knew this patient?"

B.3.2 Task 2: OCT Analysis. The second task revolved around the analysis of OCT data. The instructions were as follows:

> "In this task, you want to analyze the OCT data of the patient. For that please look at the current OCT and describe biomarkers. Afterward, find out how those biomarkers changed over time. Feel free to use any features from visual component 2."

If the participants did not use all functionalities by themselves, the interviewers showed them the missed functionality. Again we asked about efficiency, informedness, and user experience for this task. Additional follow-up questions were:

- "How do segmentation, thickness map, and 3D reconstruction help you? Would you use them in your everyday workflow?"
- "Do you trust the segmentation? Why/ why not?"

B.3.3 Task 3: Treatment Decision. In the last task, participants were asked to decide on treatment for this patient. The instructions were:

> "For the final task, please use the full CDSS to decide whether to treat this patient or not. Which medication would you administer?"

After they finished the task, we asked the following questions:

- "Do you agree with the recommendation? Why/ why not? Do you understand, why the system gives this recommendation? Is its reasoning valid? Why/ why not?"
- "How reliable do you find the segmentation, recommendation, and forecast? Do you trust these components? Would you include them in your decision-making? Why/ why not?"

Afterward, we explained the AI components:

> "We would like to explain how the AI components work to you. Please ask any questions, if you do not understand something. The segmentation and forecast are both Deep Learning models. The segmentation was trained on 1023 images to predict segmentations. The ground truth masks came from your peers.
> For the forecast model, we extracted 17 features such as bleeding, edema, etc. from the EHR data. Additionally, we fed the forecast model the quantifications of lesions and retinal thickness as well as the visual acuity, intraocular pressure,e and metadata such as age, gender, and so on. The forecast model was trained on about 50,000 samples and has seen the data of about 1000 different patients.

The recommendation algorithm uses the EHR data, the quantifications from the 3D reconstruction, and the time series forecast. Its base algorithm resembles clinical guidelines. First, it checks whether abort criteria are fulfilled. If not, it checks whether the OCT is active. If yes, then it recommends treatment, where the recommended medication can deviate from the previously given medication, if the forecast predicts a significantly better outcome. Otherwise, it will simply recommend to not treat.
Does this explanation influence your trust in these components? Please explain why."

B.4 Questionnaire

At the end of the interview, we thanked the participants again and asked them to fill in an online questionnaire regarding the demographics, sentiment on AI and SUS.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009