An AI-driven Clinical Decision Support System for the Treatment of Diabetic Retinopathy and Age-related Macular Degeneration

Robert Andreas Leist[1,*], Hans-Jürgen Profitlich[1] and Daniel Sonntag[1,2]

[1]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
[2]University of Oldenburg, Germany

Abstract

Diabetic Retinopathy (DR) and Age-related Macular Degeneration (AMD) are among the leading causes of blindness worldwide. Despite the availability of treatments to prevent disease progression, the effectiveness of these interventions is often limited by inefficiencies in existing clinical software. Recent advancements in Artificial Intelligence (AI) offer the potential to enhance Clinical Decision Support Systems (CDSS), streamlining workflows and reducing the burden on healthcare providers. This paper introduces a CDSS designed to assist ophthalmologists in the management of DR and AMD, integrating three AI-driven components. First, we developed a segmentation model for automated analysis of medical imaging data. Second, we implemented a recommendation algorithm to guide treatment decisions. Finally, we utilized a time series forecasting model to enable predictive medicine. Our models were trained using real-world clinical data from 913 patients with AMD and 461 patients with DR. The system demonstrates promising performance, underscoring the importance of high-performing AI models in advancing CDSS for ophthalmology. The code for our CDSS is available here: https://github.com/DFKI-Interactive-Machine-Learning/ophthalmo-cdss.

Keywords

Health informatics, Interactive systems and tools, Visualization, Clinical Decision Support Systems (CDSS), Interactive Machine Learning (IML), Human-AI collaboration, AI-assisted decision making, ophthalmology

1. Introduction

In ophthalmology, neovascular diseases such as Diabetic Retinopathy (DR) and Age-related Macular Degeneration (AMD) are marked by fluid accumulation within the retinal layers. The detection and monitoring process relies on Optical Coherence Tomography (OCT), a medical imaging technique akin to ultrasound [1, 2]. OCT provides a series of cross-sectional slices of the retina [3]. Interpreting OCT images requires extensive training, and many trainees report feeling unconfident in their assessments, as evidenced by a 2019 study [4]. Even among experts, inconsistencies in biomarker annotations are common [5, 6], emphasizing the need for objective, standardized analysis.

Clinical Decision Support Systems (CDSS) are software solutions designed to assist medical professionals in tasks such as diagnostics, visualization, data collection, and decision-making [7]. These systems range from basic data visualization tools to sophisticated Artificial Intelligence (AI) driven applications. As medical data grows in scale and complexity, AI-powered CDSS has become increasingly critical. Deep learning (DL) techniques, in particular, excel with large datasets and have been shown to outperform traditional systems and, in some cases, even medical experts.

For example, Barnett et al. [8] developed a system that detects Multiple Sclerosis lesions with significantly higher sensitivity than conventional radiology reports. Similarly, Google's Med-PaLM and Med-PaLM 2, large language models for medical question answering, were found to provide responses preferred over those of physicians in user studies [9, 10]. Recently, Eisemann et al. [11] have shown in a prospective study, that AI assistance lead to a significant increase in breast cancer detection rate

Joint Proceedings of the ACM IUI Workshops 2025, March 24-27, 2025, Cagliari, Italy
*Corresponding author.
robert.leist@dfki.de (R. A. Leist); hans-juergen.profitlich@dfki.de (H. Profitlich); daniel.sonntag@dfki.de (D. Sonntag)
0009-0003-9918-2501 (R. A. Leist); 0000-0003-0929-5768 (H. Profitlich); 0000-0002-8857-8709 (D. Sonntag)

by 17.6%. A more recent review by Susanto et al. [12] highlights the potential of machine learning (ML)-based CDSS, particularly in image recognition, to outperform medical experts and improve patient care. Furthermore, these systems have been shown to reduce the time required for diagnosis.

Despite the widespread adoption of CDSS in fields like radiology and general medicine [13], their application in ophthalmology, particularly for DR and AMD, remains limited [14]. Although models capable of outperforming human experts in retinal disease analysis have been developed, for example [15, 16, 17, 18, 19], they are often presented as isolated demonstrations of AI performance without integration into practical CDSS for clinical use.

This paper introduces an AI-driven CDSS designed to assist ophthalmologists in managing AMD and DR, two of the most common causes of vision loss. By integrating segmentation, time series forecast and recommendation models driven by AI, this prototype aims to bridge the gap between experimental research and practical application, laying the groundwork for future innovations in AI-supported ophthalmology. While we demonstrate decent performance, our models do not achieve state-of-the-art (SOTA). However, we present a refined CDSS with built in AI components, which assist ophthalmologists in the treatment of AMD and DR and serves as a foundation for future work in this field.

2. Related Work

In this section, we explain the therapy process for AMD and DR. Additionally, we provide an overview of the relevant literature on AI models for segmentation and time series forecasting in ophthalmology.

2.1. Treatment

Neovascular diseases such as Diabetic Retinopathy (DR) and Age-related Macular Degeneration (AMD) are marked by fluid accumulation within the retinal layers. If untreated, these conditions can result in scarring and occlusions, eventually leading to vision loss [1, 2]. Standard treatment involves a series of intravitreal injections of Anti-VEGF (Vascular Endothelial Growth Factor) medication (IVOM) upon detecting fluid presence [20]. Patients, then, undergo a series of three monthly IVOMs, before new OCTs are taken regularly, until treatment indication arises again. Since worsening VA and the accumulation of fluids are related [21], we argue that forecasting VA deterioration enables to predict fluid accumulation onset and, hence, can reduce monitoring visits.

2.2. Segmentation and Quantification

Developments in semantic segmentation, facilitate the automatic quantification of medical images, specifically OCTs [22, 23]. Semantic segmentation is an ML task, where every pixel on an image must be assigned one 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 by skip connections [24]. YNet builds upon the structure of UNet by adding another encoder branch, which first transforms the image into the Fourier domain [25]. The authors achieve state-of-the-art performance on OCT segmentation on the Duke [26] and UMN [27] data sets, especially with regard to fluid detection. The model is trained to segment individual slices. Commercial software for quantifying fluids on OCTs exists (e.g. Fluid Monitor from RetInSight[1]), but it has yet to be implemented into a CDSS that gives therapy recommendations. Therefore, in this work, we implemented a quantification algorithm based on segmentations and a recommender system that utilizes these quantifications.

2.3. 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. A study on

[1] https://retinsight.com/fluid-monitor/

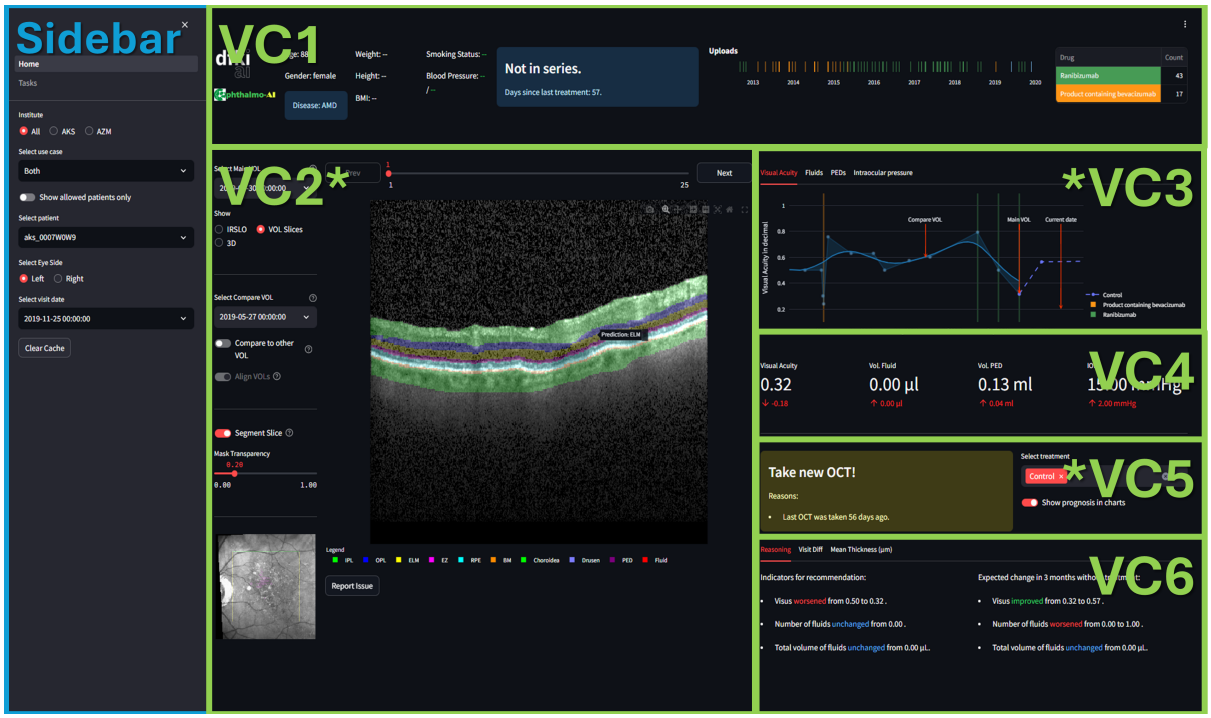


Figure 1: VCs of the evaluated CDSS. VCs that contain AI components are marked with an asterisk.

Americans older than 50 years found that an increased frequency of doctor visits correlates with less life satisfaction [28]. 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)[2]. VA measures the sharpness of vision. We decided to forecast VA, as it is the only metric in ophthalmology, which has been used for forecasts before. Additionally, while fluid volumes might be a better indicator for treatment, they depend on the segmentation model's performance, whereas VA was assessed by the doctors. In 2024, Schlosser et al. [29] 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 (WSL) scheme. However, in a medical context, a regression might be more interesting, as the amplitude of change is important, and it leaves more room for the expert to decide how to use this information. Although Schlosser et al. [29] used regression for their models, they did not share regression performance. They found that a Multi-Layer Perceptron (MLP) for VA regression with Linear Discriminant Analysis (LDA) on the regressed value performed best for the classification. However, a meta-analysis of time series prediction models in healthcare by Morid et al. [30] from 2022 has shown that Recurrent Neural Network (RNNs) [31] architectures are more performant. Especially Bidirectional Long Short Term Memory (BiLSTM) [32, 33] and Gated Recurrent Unit (BiGRU) [32, 34] networks performed well. In Schlosser et al.'s evaluation, 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 such 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.

3. Methods

In this section, we explain the data set, AI components and the composition of visual components and the visualization techniques used for the creation of our CDSS. The CDSS was implemented using

[2]In this work we use decimal VA, as the data from the clinics was in this format.

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 as well as a video of it will be made available here: https://github.com/DFKI-Interactive-Machine-Learning/ophthalmo-cdss.

3.1. Data

For our CDSS, we used real-world clinical data[5] ranging from 1993 to 2023. The data stems from the eye clinic in Sulzbach and the St. Franziskus hospital in Münster. 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[6] codes. This lead to 3192 different annotation features. Some of this data does not directly relate to diseases, such as age, gender, or smoking behaviour 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.

3.2. Segmentation and Quantification

We used the YNet architecture for a semantic segmentation of the OCT slices. YNet is a dual-encoder autoencoder-based network with a spatial encoder, a spectral encoder, and a spatial decoder [25]. The spatial encoder, similar to U-Net [24], extracts local features using convolutional blocks with batch normalization, ReLU activation, and max pooling. The spectral encoder introduces Fast Fourier Convolutional (FFC) blocks to capture global frequency domain features. These blocks use Fourier units to transform features into the frequency domain via Fast Fourier Transform (FFT), process them with convolutions, and return them to the spatial domain using inverse FFT (IFFT), enabling the model to capture frequency-based patterns. The spatial decoder combines outputs from both encoders using skip connections and transpose convolutional blocks to reconstruct the segmentation map. This integration of spatial and spectral features allows YNet to effectively segment complex patterns in medical images.

Each pixel was classified into one of eleven classes, which are shown in table 1. The training and test set contained 1023 and 400 images from 221 patients, respectively. Hyperparameter tuning was done only for the learning rate and batch size, as YNet was already fine-tuned for OCTs [25]. Additionally, we trained on different losses, such as only dice loss [35], only weighted or unweighted cross entropy loss, and a sum of both. Weighted cross entropy loss has been used in the training of the original UNet for example [24]. For the weighted cross entropy loss, we chose the weights as 0.1 for layers, 1.0 for fluids, and 0.3 and 0.5 for two other types of lesions in order to increase detection rate of lesions. We found that a batch size of 32, a learning rate of $5e^{-4}$ and the sum of the dice loss and the weighted cross entropy loss yielded the best model in terms of average dice score. We trained for 500 epochs with early stopping, if the validation loss did not improve for 25 epochs.

The quantification directly depended on the segmentation. It used the segmented slices to reconstruct lesions in three dimensions by going through them iteratively 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 [36]. 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]https://streamlit.io/
[4]https://plotly.com/
[5]This data was part of the OphthalmoAI project (https://www.interaktive-technologien.de/service/ergebnissteckbriefe/meki/ophthalmo-ai) and as such can not be published.
[6]Systematized Medical Nomenclature for Medicine–Clinical Terminology

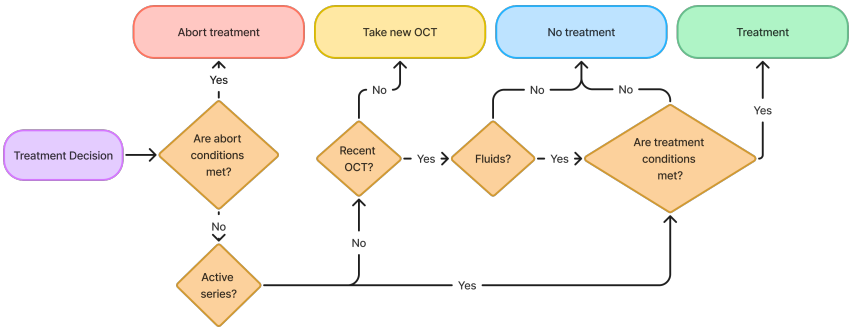


Figure 2: Flowchart of the treatment recommendation algorithm based on clinical guidelines.

3.3. Time Series Forecast

To predict patients' developments and forecast critical points for therapeutic intervention, we trained several Bidirectional Long Short Term Memory (BiLSTM) models on the available EHR data as well as the quantifications of the OCTs. We chose the BiLSTM architecture, as it demonstrated good performance in forecasting of biomedical data [30]. Long Short-Term Memory (LSTM) networks, introduced by Hochreiter [33], are a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data by using memory cells with input, forget, and output gates to regulate information flow selectively [33]. BiLSTM networks extend standard LSTMs by processing sequential data in both forward and backward directions, combining two LSTM layers to capture dependencies from both past and future contexts [37].

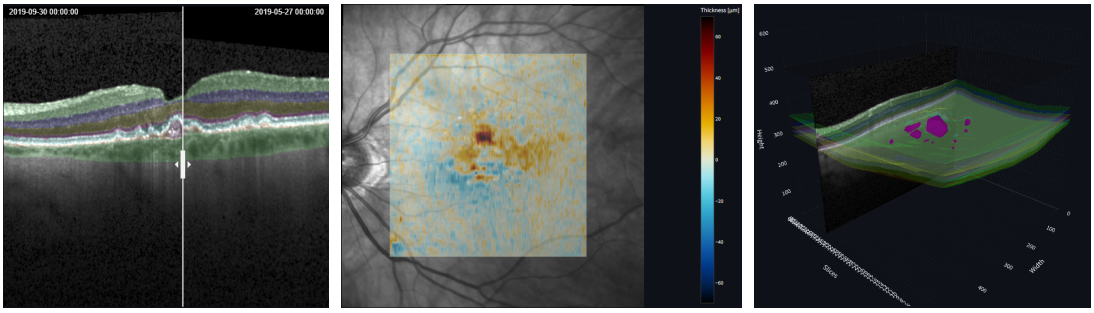
The models consisted of two layers: A BiLSTM layer[7], and a fully connected layer. We forecasted VA developments for one, three, six, nine and twelve months and trained a model for each time target. Hyperparameter tuning was done for learning rate, learning rate reduction factor, batch size, dropout rates and hidden size of the BiLSTM layer, and the number of stacked BiLSTM layers. Only the three and six month forecast models were fully tuned, while the other models were tuned on the four best parameter combinations of the former. Models were trained for 500 epochs with an early stopping threshold of 100 epochs. The learning rate was reduced, if there was no improvement after five epochs.

For the training data, 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. The test set contained data from 274 unseen patients summing up to between 10,778 and 16,290 data windows. For data window creation, we included the last twelve visits no matter the time difference between. As input features, we used EHR annotations and OCT quantifications. Missing data was addressed using moving average interpolation with a triangular window on numeric variables, such as VA or fluid volume. Windows included the data of 45 days before and after the missing data point. If there was not at least one data point before and after the missing data point, we discarded the window. Categorical data was imputed by using nearest neighbour imputation.

3.4. Recommendation

The recommendation task 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 segmentation and time series forecast models and, hence, AI driven.

[7] https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html (Accessed: 10.01.2025)



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

Figure 3: Three of the main visualization components of our Clinical Decision Support System.

3.5. Design Rationale

Our dashboard design is loosely inspired by the approach of Bhattacharya et al. [38] and refined through a workflow assessment interview with an assistant doctor and an expert from a German eye clinic. The doctor described their ideal CDSS, which informed the creation of a low-fidelity prototype in Microsoft Word[8]. After one feedback iteration with the same doctor, the prototype was refined (see appendix A). A high fidelity prototype was developed and evaluated with eleven ophthalmologists.[9]. The feedback from the evaluation was built into this final prototype.

3.6. Visual Components

We display metadata in **VC1**, including treatment status and an IVOM timeline. The treatment status provides an overview of therapy progression, while the timeline graph shows IVOMs color-coded by medication. In **VC3**, important metrics such as VA are presented in line graphs, featuring medication lines and a dotted forecast model prediction from **VC5**. Below, **VC4** shows color-coded percentage changes between visits. **VC5** displays treatment recommendations color-coded (green: therapy, red: stop therapy, yellow: imaging needed, blue: no therapy) with reasons listed below. **VC6** includes EHR data across three tabs: Reasoning (key metrics and forecasts), Visit Diff (symptom annotations), and Mean Thickness (layer thickness comparisons).

The OCTs are visualized in **VC2** with three views: Slices, IRSLO[10], and 3D reconstruction. The Slices view allows navigation through OCT slices, segmentation overlay, and comparison with older OCTs via an alignment slider (Figure 3a). The 3D view (Figure 3c) enables cross-checking slice data against 3D reconstructions but excludes comparisons due to performance limits. The IRSLO view provides an ”En Face” retina view with tools for comparing lesion areas and projecting retinal layer thickness as a heatmap, which can display thickness differences between OCTs (Figure 3b).

4. Results

4.1. Model performance

Our segmentation model achieves an average dice score of 0.66. The dice score is a segmentation measurement. It is computed by taking two times the overlap of predicted and actual region divided by the sum of both regions, and, hence, falls into a range between zero and one, whereas a score of one is perfect, while a score of zero is completely wrong [39]. As shown in table 1, certain classes, such as drusen, the ellipsoid zone (EZ), the Retinal Pigment Epithelium (RPE), and Bruch’s Membrane (BM), are

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

[9] The findings of our qualitative evaluation are currently under conditional acceptance at IUI25 and we will add a reference, once they are accepted (23.01.2025)

[10] InfraRed Scanning Laser Ophthalmoscopy.

Table 1
Class dependent Dice scores of the trained YNet model

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

Table 2
Performance of the time series forecasting model.

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

particularly challenging to predict accurately. While the model's Dice score for fluid segmentation is also 0.66, it demonstrates strong performance in identifying the presence or absence of fluids, achieving a 94.5% accuracy. This accuracy measures the model's ability to correctly identify at least one fluid pixel in a slice or correctly classify a slice with no fluid by not predicting any fluid pixels.

We evaluate our time series forecast models on our test set in terms of Mean Absolute Error (MAE). We decided for MAE, because it gives a direct impression of the scale of the error on the metric, whereas for example mean squared error is less interpretable. Note that data in the test set comes from patients not present in the training set. As can be seen in table 2, our time series forecast has mean absolute errors of 0.224 to 0.269 with standard deviations of 0.137 to 0.173, whereas the untuned models, expectedly, performed worse. With decimal VA values ranging from 0 to 3, this equates to errors of about 7.4% to 8.9% with standard deviations of 4.6% to 5.7%.

The recommendation system was evaluated in terms of its agreement with historical data. In about 60% of cases, recommendation and historical decision agree for, whether to treat or not. In 85% of the cases, where a patient needs treatment, they agree on the medication decision. The low agreement on whether to administer IVOMs, might be impacted by the fact that patients do not strictly follow a perfect therapy plan.

5. Discussion

Our AI tools demonstrate decent performance, but cannot reach SOTA. For instance, we trained YNet on our data, but only achieve a dice score of 0.66 for fluids. We expect more elaborate hyperparameter tuning to significantly improve this metric. Additionally, Farshad et al. [25] train on data of ten patients annotated by a single expert, while our data comes from 221 different patients from two eye clinics with annotations done by multiple experts. Moreover, during training we noticed an error in the evaluation code of Farshad et al. [25], where correctly predicting a class to be absent gave a dice score of 2. This lead to an overestimation of the class dependent dice score. We think that this mistake could have influenced the YNet evaluation. An improved segmentation model could contribute to objective, measurable fluid quantifications and, hence, improve patient care. Furthermore, since the time series model and recommendation system rely on the segmentation quantifications as input features, any improvements in the segmentation process could potentially enhance the performance of these systems as well.

Moreover, our time series forecasting model demonstrates good performance. This highlights the potential of such models in preventive and predictive medicine. However, further studies are needed to finetune the architecture, evaluate the impact of data preprocessing and compare different model architectures. Especially, DL architectures are yet to be extensively applied in the forecasting of VA. A major challenge was the data preparation step, as missing data had to be appropriately imputed. We advise future research to start with simple imputation techniques like nearest neighbour for categorical features and a rolling window interpolation for numeric features. However, this must also be considered

as a hyperparameter for tuning.

Our recommendation system is only 10% above chance level. However, we argue that it would still improve clinicians' decision, as it adheres to clinical treatment guidelines, and our evaluation was only done in retrospective not accounting for patient preferences or scheduling complications. A prospective study, similar to [11], could show whether patients and medical experts can profit of a CDSS, similar to ours. Alternatively, creating a new data set, where experts give a gold standard treatment decision for a set of visits, could improve the evaluation of such a system.

6. Conclusion

We have developed an AI-driven CDSS that integrates data visualization and decision support functionalities, offering a solid foundation for future CDSS advancements. Our tool has undergone two refinement iterations and demonstrates promising performance. However, the lack of hyperparameter tuning and comparisons with other model types represents a significant limitation of our study.

Despite these constraints, our time series forecasting results highlight considerable potential for preventive and predictive medicine, underscoring its value in CDSS applications. We hope our prototype serves as a useful starting point for future research and development in this domain.

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).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] L. S. Lim, P. Mitchell, J. M. Seddon, F. G. Holz, T. Y. Wong, Age-related macular degeneration, The Lancet 379 (2012) 1728–1738.

[2] D. S. Fong, L. P. Aiello, F. L. Ferris III, R. Klein, Diabetic retinopathy., Diabetes care 27 (2004).

[3] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, et al., Optical coherence tomography, science 254 (1991) 1178–1181.

[4] W. H. Dean, S. Grant, J. McHugh, O. Bowes, F. Spencer, Ophthalmology specialist trainee survey in the United Kingdom, Eye 33 (2019) 917–924. URL: https://www.nature.com/articles/s41433-019-0344-z. doi:10.1038/s41433-019-0344-z, publisher: Nature Publishing Group.

[5] T. Oberwahrenbrock, G. L. Traber, S. Lukas, I. Gabilondo, R. Nolan, C. Songster, L. Balk, A. Petzold, F. Paul, P. Villoslada, A. U. Brandt, A. J. Green, S. Schippling, Multicenter reliability of semiautomatic retinal layer segmentation using OCT, Neurology Neuroimmunology & Neuroinflammation 5 (2018) e449. URL: https://www.neurology.org/doi/full/10.1212/NXI.0000000000000449. doi:10.1212/NXI.0000000000000449, publisher: Wolters Kluwer.

[6] M. Melinščak, M. Radmilović, Z. Vatavuk, S. Lončarić, 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 (2021) 375–385. URL: https://hrcak.srce.hr/clanak/391401. doi:10.1080/00051144.2021.1973298, publisher: KoREMA - Hrvatsko društvo za komunikacije,računarstvo, elektroniku, mjerenja i automatiku.

[7] E. S. Berner, Clinical decision support systems, volume 233, Springer, 2007.

[8] M. Barnett, D. Wang, H. Beadnall, A. Bischof, D. Brunacci, H. Butzkueven, J. W. L. Brown, M. Cabezas, T. Das, T. Dugal, et al., A real-world clinical validation for ai-based mri monitoring in multiple sclerosis, NPJ Digital Medicine 6 (2023) 196.

[9] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, N. Scharli, A. Chowdhery, P. Mansfield, B. A. y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, V. Natarajan, Large language models encode clinical knowledge, 2022. arXiv:2212.13138.

[10] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. A. y Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, V. Natarajan, Towards expert-level medical question answering with large language models, 2023. arXiv:2305.09617.

[11] N. Eisemann, S. Bunk, T. Mukama, H. Baltus, S. A. Elsner, T. Gomille, G. Hecht, S. Heywang-Köbrunner, R. Rathmann, K. Siegmann-Luz, et al., Nationwide real-world implementation of ai for cancer detection in population-based mammography screening, Nature Medicine (2025) 1–8.

[12] A. P. Susanto, D. Lyell, B. Widyantoro, S. Berkovsky, F. Magrabi, 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 (2023) ocad180. URL: https://doi.org/10.1093/jamia/ocad180. doi:10.1093/jamia/ocad180. arXiv:https://academic.oup.com/jamia/advance-article-pdf/doi/10.1093/jamia/ocad180/51309090/ocad180.pdf.

[13] A. C. Perlich, Nutzung und akzeptanz klinischer entscheidungsunterstützungssysteme-entwurf eines modells für die medizinische lehre (2022).

[14] I. De la Torre-Díez, B. Martínez-Pérez, M. López-Coronado, J. R. Díaz, M. M. López, Decision support systems and applications in ophthalmology: literature and commercial review focused on mobile apps, Journal of medical systems 39 (2015) 1–10.

[15] S. A. Kamran, S. Saha, A. S. Sabbir, A. Tavakkoli, 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, 2019, pp. 964–971.

[16] S. A. Kamran, A. Tavakkoli, S. L. Zuckerbrod, 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, 2020, pp. 2476–2480.

[17] V. Melnychuk, E. Faerman, I. Manakov, T. Seidl, Matching the clinical reality: Accurate oct-based diagnosis from few labels, arXiv preprint arXiv:2010.12316 (2020).

[18] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, cell 172 (2018) 1122–1131.

[19] D. S. Ting, Y. Liu, P. Burlina, X. Xu, N. M. Bressler, T. Y. Wong, Ai for medical imaging goes deep, Nature medicine 24 (2018) 539–540.

[20] V. Tah, H. O. Orlans, J. Hyer, E. Casswell, N. Din, V. Sri Shanmuganathan, L. Ramskold, S. Pasu, Anti-VEGF Therapy and the Retina: An Update, Journal of Ophthalmology 2015 (2015) 627674. URL: https://onlinelibrary.wiley.com/doi/abs/10.1155/2015/627674. doi:10.1155/2015/627674, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2015/627674.

[21] V. Chaudhary, F. G. Holz, S. Wolf, E. Midena, E. H. Souied, H. Allmeier, G. Lambrou, T. Machewitz, P. Mitchell, A. study investigators, Association between visual acuity and fluid compartments with treat-and-extend intravitreal aflibercept in neovascular age-related macular degeneration: an aries post hoc analysis, Ophthalmology and therapy 11 (2022) 1119–1130.

[22] M. Nawaz, A. Uvaliyev, K. Bibi, H. Wei, S. M. D. Abaxi, A. Masood, P. Shi, H.-P. Ho, W. Yuan, Unravelling the complexity of optical coherence tomography image segmentation using machine and deep learning techniques: A review, Computerized Medical Imaging and Graphics (2023) 102269.

[23] Z. Li, L. Wang, X. Wu, J. Jiang, W. Qiang, H. Xie, H. Zhou, S. Wu, Y. Shao, W. Chen, Artificial intelligence in ophthalmology: The path to the real-world clinic, Cell Reports Medicine 4 (2023).
[24] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, 2015. URL: https://arxiv.org/abs/1505.04597. arXiv:1505.04597.
[25] A. Farshad, Y. Yeganeh, P. Gehlbach, N. Navab, Y-net: A spatiospectral dual-encoder networkfor medical image segmentation, 2022. URL: https://arxiv.org/abs/2204.07613. arXiv:2204.07613.
[26] S. J. Chiu, M. J. Allingham, P. S. Mettu, S. W. Cousins, J. A. Izatt, S. Farsiu, Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema, Biomedical optics express 6 (2015) 1172–1194.
[27] A. Rashno, B. Nazari, D. D. Koozekanani, P. M. Drayna, S. Sadri, H. Rabbani, K. K. Parhi, Fully-automated segmentation of fluid regions in exudative age-related macular degeneration subjects: Kernel graph cut in neutrosophic domain, PloS one 12 (2017) e0186949.
[28] E. S. Kim, N. Park, J. K. Sun, J. Smith, C. Peterson, Life Satisfaction and Frequency of Doctor Visits, Psychosomatic Medicine 76 (2014) 86. URL: https://journals.lww.com/psychosomaticmedicine/abstract/2014/01000/life_satisfaction_and_frequency_of_doctor_visits.12.aspx. doi:10.1097/PSY.0000000000000024.
[29] T. Schlosser, F. Beuth, T. Meyer, A. S. Kumar, G. Stolze, O. Furashova, K. Engelmann, D. Kowerko, Visual acuity prediction on real-life patient data using a machine learning based multistage system, Scientific Reports 14 (2024) 5532.
[30] M. A. Morid, O. R. L. Sheng, J. Dunbar, Time series prediction using deep learning methods in healthcare, 2022. arXiv:2108.13461.
[31] L. R. Medsker, L. Jain, Recurrent neural networks, Design and Applications 5 (2001) 2.
[32] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, nature 323 (1986) 533–536.
[33] S. Hochreiter, Long short-term memory, Neural Computation MIT-Press (1997).
[34] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. URL: https://arxiv.org/abs/1412.3555. arXiv:1412.3555.
[35] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), Ieee, 2016, pp. 565–571.
[36] C. B. Barber, D. P. Dobkin, H. Huhdanpaa, The quickhull algorithm for convex hulls, ACM Trans. Math. Softw. 22 (1996) 469–483. URL: https://doi.org/10.1145/235815.235821. doi:10.1145/235815.235821.
[37] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional lstm networks, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., volume 4, IEEE, 2005, pp. 2047–2052.
[38] A. Bhattacharya, J. Ooge, G. Stiglic, K. Verbert, 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, 2023, pp. 204–219.
[39] A. Zijdenbos, B. Dawant, R. Margolin, A. Palmer, Morphometric analysis of white matter lesions in mr images: method and validation, IEEE Transactions on Medical Imaging 13 (1994) 716–724. doi:10.1109/42.363096.

A. Low fidelity prototype

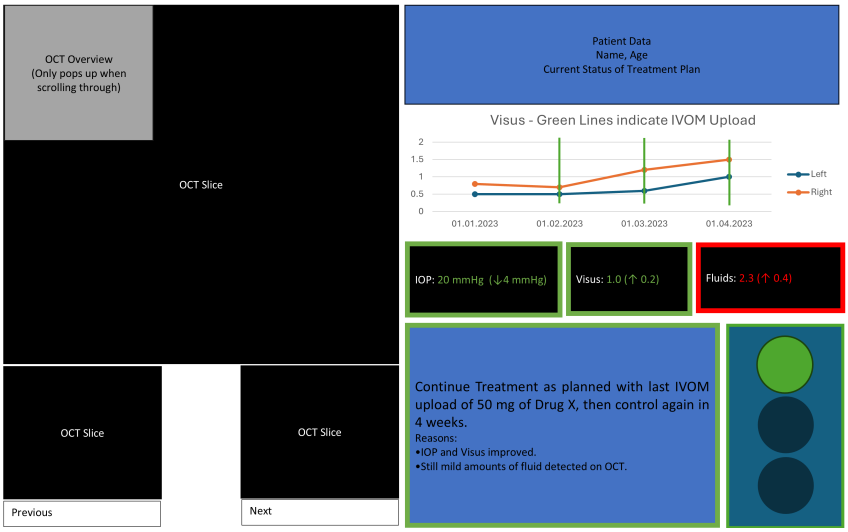


Figure 4: Low fidelity prototype developed after a workflow assessment interview with one doctor and an expert.