# Bridging the Gap: Gaze Events as Interpretable Concepts to Explain Deep Neural Sequence Models

Krakowczyk, Daniel G ; Prasse, Paul ; Reich, David R ; Lapuschkin, Sebastian ; Scheffer, Tobias ; Jäger, Lena A

Abstract: Recent work in XAI for eye tracking data has evaluated the suitability of feature attribution methods to explain the output of deep neural sequence models for the task of oculomotric biometric identification. These methods provide saliency maps to highlight important input features of a specific eye gaze sequence. However, to date, its localization analysis has been lacking a quantitative approach across entire datasets. In this work, we employ established gaze event detection algorithms for fixations and saccades and quantitatively evaluate the impact of these events by determining their concept influence. Input features that belong to saccades are shown to be substantially more important than features that belong to fixations. By dissecting saccade events into sub-events, we are able to show that gaze samples that are close to the saccadic peak velocity are most influential. We further investigate the effect of event properties like saccadic amplitude or fixational dispersion on the resulting concept influence.

DOI: https://doi.org/10.1145/3588015.3588412

# Bridging the Gap: Gaze Events as Interpretable Concepts to Explain Deep Neural Sequence Models


Daniel G. Krakowczyk
daniel.krakowczyk@uni-potsdam.de
University of Potsdam
Potsdam, Germany

Paul Prasse
paul.prasse@uni-potsdam.de
University of Potsdam
Potsdam, Germany

David R. Reich
david.reich@uni-potsdam.de
University of Potsdam
Potsdam, Germany

Sebastian Lapuschkin
sebastian.lapuschkin@hhi.fraunhofer.de
Fraunhofer Heinrich Hertz Institute
Berlin, Germany

Tobias Scheffer
tobias.scheffer@uni-potsdam.de
University of Potsdam
Potsdam, Germany

Lena A. Jäger
jaeger@cl.uzh.ch
University of Zurich
Zurich, Switzerland
University of Potsdam
Potsdam, Germany



## ABSTRACT

Recent work in XAI for eye tracking data has evaluated the suitability of feature attribution methods to explain the output of deep neural sequence models for the task of oculomotric biometric identification. These methods provide saliency maps to highlight important input features of a specific eye gaze sequence. However, to date, its localization analysis has been lacking a quantitative approach across entire datasets. In this work, we employ established gaze event detection algorithms for fixations and saccades and quantitatively evaluate the impact of these events by determining their *concept influence*. Input features that belong to saccades are shown to be substantially more important than features that belong to fixations. By dissecting saccade events into sub-events, we are able to show that gaze samples that are close to the saccadic peak velocity are most influential. We further investigate the effect of event properties like saccadic amplitude or fixational dispersion on the resulting concept influence.



## CCS CONCEPTS

• **Human-centered computing** → **Scientific visualization**; • **Computing methodologies** → **Machine learning**; • **Applied computing** → **Psychology**.

## KEYWORDS

xai, explainability, concept influence, time-series, eye movements

**ACM Reference Format:**
Daniel G. Krakowczyk, Paul Prasse, David R. Reich, Sebastian Lapuschkin, Tobias Scheffer, and Lena A. Jäger. 2023. Bridging the Gap: Gaze Events as Interpretable Concepts to Explain Deep Neural Sequence Models. In *2023 Symposium on Eye Tracking Research and Applications (ETRA '23), May 30–June 02, 2023, Tubingen, Germany.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3588015.3588412





*ETRA '23, May 30–June 02, 2023, Tubingen, Germany*



ACM ISBN 979-8-4007-0150-4/23/05.
https://doi.org/10.1145/3588015.3588412


## 1 INTRODUCTION & RELATED WORK

Deep neural networks led to considerable advances in the domains of computer vision [Voulodimos et al. 2018], speech recognition [Nassif et al. 2019], time series analysis [Fawaz et al. 2019] and gaze analysis [Jäger et al. 2019]. It has been widely observed that end-to-end training of deep neural networks, that is, using unprocessed data as input to the network and let the model learn internal representations, typically outperforms approaches that use aggregated data and engineered features as model input [Krizhevsky et al. 2012]. We observe this trend also in gaze analysis, where several network architectures have been presented over the last years that improved on the state-of-the-art by using the non-aggregated gaze velocity time series as input. Breakthroughs were especially made on the task of oculomotoric biometric identification, where large improvements in performance were observed that even went along with a decrease in the required duration of input sequence length for successful identification [Jäger et al. 2019; Lohr et al. 2021; Lohr and Komogortsev 2022; Makowski et al. 2021].

The downside of these complex neural networks is their black box nature as they are generally not interpretable. This issue is particularly acute for the vast amount of potential medical applications of gaze analysis, such as the detection of autism [Alcañiz et al. 2021; Jiang and Zhao 2017], ADHD [Deng et al. 2023] or developmental language disorders [Key et al. 2020; Raatikainen et al. 2021], as medical applications typically require explainable model predictions.

To provide explanations alongside model predictions, local post-hoc feature attribution methods have been developed to provide saliency maps of the relevance of each input feature [Bach et al. 2015; Shrikumar et al. 2016; Sundararajan et al. 2017]. This way we can visualize the positive and negative impact of the input and point to the most important parts of the input that led to a specific model decision. These methods are local in the sense that they are computed for a single data instance, and they are post-hoc as they are applied to an existing trained model [Molnar 2022].

However, in the context of end-to-end trained neural networks without explicit feature engineering we also have an input space that is much less interpretable than engineered features. That leads to *feature attribution methods* coming short in interpretability as

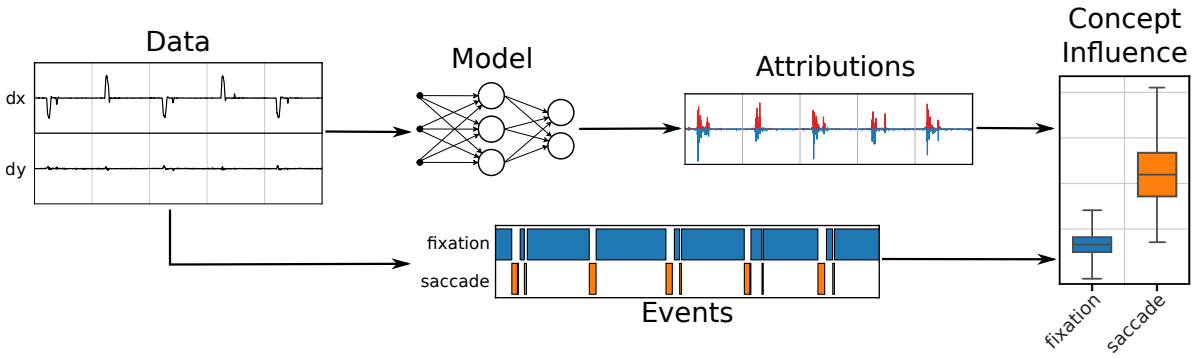



**Figure 1: The overall process for evaluating the concept influences of distinct types of gaze events.**

saliency maps lack semantic concepts in complex input spaces. Moreover, global insights across whole datasets are desired to explain the intertwined system of data and trained model. To overcome this limitation, we can evaluate the impact of the occurence of particular semantic concepts in the input signal, a method named *concept influence* [Theiner et al. 2022].

Alternative methods to evaluate the impact of semantic concepts on the model output include the inpainting or masking of specific concepts and quantifying the change of output of the neural network [Williford et al. 2020]. Moreover, internal representations of neural networks can be harnessed to generate concept activation vectors [Kim et al. 2017].

Gaze events like fixations and saccades are main descriptive concepts of eye tracking research [Holmqvist et al. 2011] and thus are natural candidates for semantic concepts in gaze signals. By evaluating the *concept influence* of these gaze concepts across entire datasets we generate insights on both model and data. Prior eye gaze related research either uses descriptive input features on machine learning models [Rigas et al. 2016] or is limited to statistical analysis [Holland and Komogortsev 2013; Rigas et al. 2018] to gain insights on the impact of semantic concepts. First studies on the interpretability of models in the eye tracking domain are carried out by Kumar et al. [2020] and feature a PCA analysis of the embedding layer of a neural network. Further, Krakowczyk et al. [2022] evaluate several feature attribution methods which are applied in the context of oculomotoric biometric identification.

This work in turn puts forward the following contributions for evaluating deep neural sequence models:

- Evaluation of *concept influence* of the basic event types saccades and fixations;
- Dissection of saccades into sub-events and evaluation of their *concept influences*;
- Investigation of the relationship between event properties and the resulting *concept influence*.

## 2 PROBLEM SETTING

We investigate the explainability of models which get eye gaze velocity time-series data as input and which are trained in a biometric identification setting. We choose this specific task for its most striking performance benefits over traditional feature engineering approaches.

Given eye gaze time-series data $X \subset \mathbb{R}^{N \times D \times L}$ with $N$ instances, $D$ channels and a sequence length of $L$, and a one-hot coded participant labeling $Y \in \{0,1\}^{N \times K}$ with $K$ labels, we can train a biometric model $\phi : \mathbb{R}^{D \times L} \rightarrow \mathbb{R}^{K}$ in both a multiclass and a metric learning setting to output the presumed identity of a recorded participant [Lohr and Komogortsev 2022]. We can further create a local post-hoc feature attribution function $f(\mathbf{x}, \phi)$ which attributes a relevance to each input feature of any instance $\mathbf{x}$ in regard to the output of the actual model $\phi$ [Krakowczyk et al. 2022].

For image-data these feature attributions are often called pixel-wise explanations [Bach et al. 2015], and although referred to as *explanations*, they can nevertheless lack interpretability as single pixels are not inherently interpretable by default [Theiner et al. 2022]. This problem can intensify with multi-channel time-series data, as the input space is potentially visually less interpretable.

By measuring the overlap of the highest attributed input features and an interpretable segmentation that refers to a specific *concept*, we can compute the *concept influence* of this concept [Theiner et al. 2022]. Whereas in the image-domain, segmentations are sets of pixels that are associated with detected objects like houses, streets, persons or the sky, in the time-series domain, segmentations refer to events delimited by their on- and offsets. In the case of eye tracking data, gaze events suggest themselves as interpretable concepts for our study, as their function and underlying neuro-biological processes have been extensively researched over the past decades [Engbert and Kliegl 2003; Martinez-Conde et al. 2004, 2006; Rayner et al. 2004; Rayner and Pollatsek 1983], and there also exist a vast amount of methods to automatically detect them [Andersson et al. 2016; Startsev and Zemblys 2022].

When investigating which parts of the input sequence are most relevant for the output of deep neural sequence models, we can make use of the established concepts of distinct gaze events and evaluate which ones of these exhibit the highest influence on the model output.

## 3 MATERIALS AND METHODS

This section is structured alongside Figure 1, where we illustrate the overall evaluation process. We start out by presenting the used

datasets in Subsection 3.1 and subsequently describe our data preprocessing steps in Subsection 3.2. We continue by introducing the employed algorithms for gaze event detection in Subsection 3.3 and detail the method for dissecting a saccadic event into sub-events in Subsection 3.4. Subsection 3.5 briefly covers the biometric task and model under investigation, whereas Subsection 3.6 gives an overview of the employed attribution methods. We delineate the term *concept influence* in Subsection 3.7. The overall evaluation protocol is detailed in Subsection 3.8.

## 3.1 Dataset

We make use of the three publicly available datasets *GazeBase* [Griffith et al. 2021], *JuDo1000* [Makowski et al. 2020] and the *Potsdam Textbook Corpus (PoTeC)* [Jäger et al. 2021]. All datasets are recorded at a sampling rate of 1000 Hz. Only *JuDo1000* contains binocular recordings. We have reduced the GazeBase dataset to the first 4 rounds where most subjects participated in. We use all of the available stimuli for evaluation. Table 1 in the Appendix provides a brief summary of dataset properties.

## 3.2 Data Preprocessing

Based on the preprocessing pipeline of Lohr and Komogortsev [2022], we apply the Savitzky-Golay differentiation filter [Savitzky and Golay 1964] with a window size of 7 and an order of 2 to transform positional data into gaze velocity data. We construct subsequences by a non-overlapping rolling window with a window size of 5 s for *GazeBase* and 1 s for *JuDo1000* and *PoTeC*. We clamp velocities to ±1000 °/s and exclude all subsequences which would need padding or comprise more than 50% missing values. We finally apply z-score normalization and replace all missing values with 0. We leverage the pymovements package for these preprocessing steps [Krakowczyk et al. 2023].

## 3.3 Gaze Event Detection Algorithms

We limit our study to the investigation of fixations and saccades as these two event types are the main ones investigated in the literature. Further candidates would have been post-saccadic oscillations, smooth pursuit and blinks, but also micromovements like drift and tremor. Note that we do not distinguish between saccades and microsaccades in this work and include microsaccades in the set of saccades. For binocular data we solely use the right eye for event detection.

We use distinct detection algorithms for fixations and saccades as suggested by Andersson et al. [Andersson et al. 2016]. We use the I-VT algorithm [Salvucci and Goldberg 2000] to detect fixations and the algorithm of Engbert and Kliegl [2003] to detect saccades. Table 2 in the Appendix lists all parameters used in the event detection process. Although the I-VT algorithm originally just uses a single parameter for its fixation velocity threshold, we make use of an additional minimum fixation duration and maximum fixation dispersion threshold to avoid misclassifications. Fixations that exceed these values will be simply excluded from evaluation.

The main parameter of the employed saccade detection algorithm is the threshold factor which is multiplied with the adaptively determined noise threshold. We further exclude saccades from evaluation which exceed the valid ranges for saccade duration and peak velocity stated in Table 2 in the Appendix.

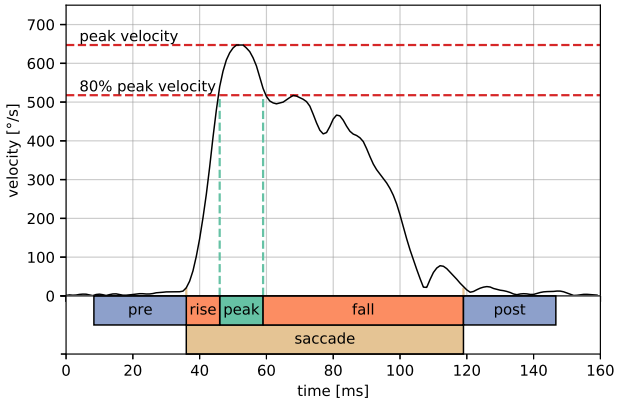



**Figure 2: Illustration of the saccade event dissection algorithm. The saccade event, depicted as the brown horizontal bar at the bottom, is detected by the saccade detection algorithm [Engbert and Kliegl 2003]. All samples in a saccade that reach at least 80% of its peak velocity are associated with the peak phase (green bar). The rise phase lasts from saccade onset to peak onset, the fall phase lasts from peak offset to saccade offset (orange bars). Finally we add a pre and post phase at the beginning and the end of a saccade with a duration of 1/3 of the total saccade duration (blue bars).**

## 3.4 Event Dissection for Saccades

Looking at a typical velocity profile of saccadic eye movements, we can identify distinct phases which are illustrated in Figure 2. We see a rapid increase in velocity at the beginning and a velocity decline at the end of a saccade, with a short phase in-between where the gaze velocity is near saccade peak velocity. We call these sub-events the rise, fall and peak phases. The peak phase is defined as the time steps where the velocity is at least 80% of the peak velocity of the respective saccade event. We further chain an additional event to each before the beginning (pre-phase) and after the end (post-phase) of a saccade, with a duration set to be 1/3 of the associated saccade. Due to the neglible occurence of samples which are below 80% of the peak velocity but are between two local peaks greater than 80% we disregard these samples in our analysis. This way we get a total of 5 granular events out of a single saccade event.

## 3.5 Biometric Model

We investigate the explainability of *EyeKnowYouToo*, a state-of-the-art neural network model for oculomotoric biometric identification developed by Lohr and Komogortsev [2022]. This model is a convolutional network that uses multi-channel sequences of yaw (horizontal) and pitch (vertical) angular gaze velocities as input and that is end-to-end trained to minimize a weighted sum of categorical cross-entropy and multi-similarity loss. Instead of using the output of the embedding layer for comparison with existing embeddings of an enrollment database as in the original biometric system [Lohr and Komogortsev 2022], we make use of the nodes of

the classification layer as targets for calculating attributions of each inference. We modified Dillon Lohr's model implementation [Lohr and Komogortsev 2022] in order to facilitate our application of attribution methods. Changes relate to the naming and grouping and layers while the overall model architecture and behavior during training is left the same as in the original.

We restrict our study to this single model as it exhibits the best performance on the given biometric task while also being smallest in the number of model parameters. Although a comparison of explainability metrics between state-of-the-art biometric models is interesting, such a comparison unfortunately cannot be in scope of our study.

### 3.6 Attribution Methods

Feature attribution methods attribute relevance to each input feature, such that saliency maps can be generated to visualize the positive and negative impact of the input for a specific model prediction [Bach et al. 2015; Shrikumar et al. 2016; Sundararajan et al. 2017].

Based on the findings of Krakowczyk et al. [2022], we limit this study to the three best performing methods: DeepLIFT (DL) [Shrikumar et al. 2017], Integrated Gradients (IG) [Sundararajan et al. 2017] and Layer-wise Relevance Propagation (LRP) [Bach et al. 2015; Montavon et al. 2017, 2018]. We use the Zennit library [Anders et al. 2021] for the implementation of LRP rules and the Captum library [Kokhlikyan et al. 2020] for its DeepLIFT and IG implementations.

All three methods are backpropagation-based in the way that they propagate the relevance of the model output back to each input feature [Ancona et al. 2017]. DeepLIFT and IG additionally require a baseline reference input which is desired to generate neutral model output and supposed to have low relevance across all input features. We set this baseline to zero in concordance with predominant usage [Sturmfels et al. 2020].

IG attributes relevance by computing the gradients of a model with respect to each input feature. Input features are step-wise linearly interpolated from the reference baseline into the given input instance. The integral of the gradients along this interpolatation path is multiplied by the difference between reference baseline and given input instance [Sundararajan et al. 2017].

LRP attributions are computed by backpropagating the model output layer by layer. Depending on the product of activations and weights of the incoming connections, relevance of each unit is passed down to the preceding layer. We limit this study to the vanilla LRP-$\varepsilon$ rule [Kohlbrenner et al. 2020] and set $\varepsilon = 0.25$ [Montavon et al. 2019].

DeepLIFT [Shrikumar et al. 2017] is similar to the former layer-wise backpropagation method but uses the reference baseline to calculate activation reference points for each unit. Activation differences to the reference points are then backpropagated as relevance.

### 3.7 Concept Influence

As stated in the problem setting in Section 2, the drawback of pixel-wise feature attributions is that pixels are usually not inherently interpretable on their own. When we ask which parts of the input have the biggest impact on the output of the model under investigation, we usually expect the explanation to be given in interpretable high-level concepts instead of a simple saliency map. We further do not want to be limited to local attributions computed for individual data instances only, but aim at global inferences about the dataset and model as a whole.

The concept influence method proposed by Theiner et al. [2022] which is originally developed for image data, tackles this issue by quantifying the overlap between given concepts and the highest attributions. To this end, each concept is represented as a binary segmentation $S \in \{0,1\}^L$, where $S_i = 1$ encodes the presence of the respective concept at the specific step $i$ in the sequence of length $L$. We refer to the segmentation $S$ as the *concept segmentation*, with its size $|S|$ being defined as $\sum_{i=1}^{L} S_i$ given the sequence length $L$.

We further create a second segmentation $T \in \{0,1\}^L$ from the top-$k$ highest feature attribution values, after squashing multi-channel feature attributions to a single channel by taking the step-wise maximum. As in the original work by Theiner et al. [2022] we set $k$ to be 2 % of the input size (20 for an $L$ of 1000, 100 for an $L$ of 5000).

To measure the influence of a specific concept, we take the size of the intersection of the concept segmentation $S$ and the top-$k$ attribution segmentation $T$. This intermediate result is called the top-$k$ intersection [Theiner et al. 2022]. To account for the fact that the size of the concept segmentation has an impact on the resulting intersection, we perform a weighting by dividing the top-k intersection by the size of the segmentation $|S|$ relative to the sequence length $L$ as defined in Equation 1.

$$c = \frac{L}{|S|} \frac{1}{k} \sum_{i=1}^{L} S_i \wedge T_i \tag{1}$$

The resulting value $c$ is the *concept influence* for the respective concept and ranges between 0 and $L \;/\; |S|$. A concept influence above 1 is regarded as highly influential [Theiner et al. 2022].

### 3.8 Evaluation Protocol

We train model weights and subsequently generate attributions by applying a k-fold cross-validation protocol which splits data into training and test sets. We use different schemes for each dataset to split data into folds: leave-one-round-out for *GazeBase* ($k = 4$), leave-one-session-out for *JuDo1000* ($k = 4$), and leave-one-text-out for *PoTeC* ($k = 12$). We compute attributions on the test set and set the predicted class as the target class for relevance computation.

We implement the general evaluation framework using scikit-learn [Pedregosa et al. 2011] und use the attribution metric implementations from Quantus [Hedström et al. 2023]. Our hardware setup comprises an AMD EPYC 7742 CPU and a NVIDIA DGX A100 GPU. The code can be found online.[1]

## 4 EXPERIMENTAL RESULTS

We present our experimental results in the following section. All attributions are evaluated on the model described in Section 3.5. Figure 1 in the Appendix reports accuracies of at least 90% for each used dataset. We begin by putting forward the concept influence

[1] https://github.com/aeye-lab/etra-2023-bridging-the-gap

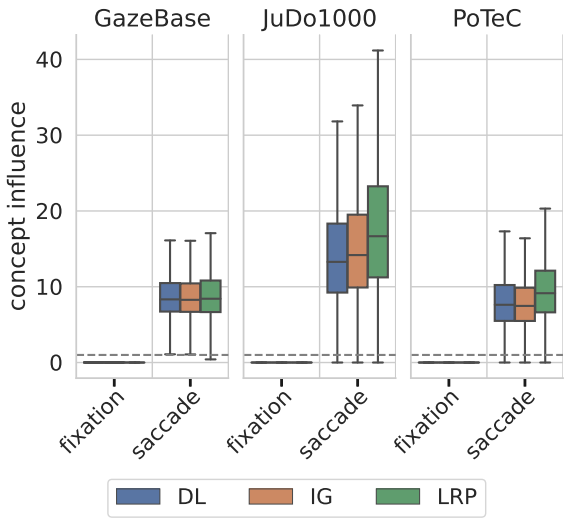



**Figure 3: Concept influences for saccades and fixations**

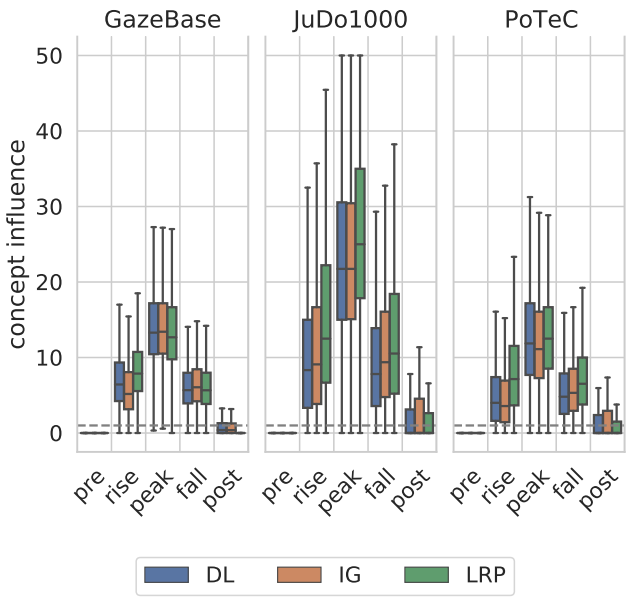



**Figure 4: Concept influences for saccade sub-events.**

evaluation of the basic event types in Section 4.1, whereas Section 4.2 deals with the concept influences of saccade sub-events. We bin fixations and saccades according to several properties and show our results in Section 4.3.

## 4.1 Concept Influence of Fixations and Saccades

Out of the detected saccades and fixations from Section 3.3 we create concept segmentations for each event type. The distribution of segmentation sizes is depicted in Figure 2 of the Appendix. Across all three datasets we observe a much greater segmentation size for fixations than for saccades.

The resulting concept influences in Figure 3 demonstrate very high values for saccades instead. In contrast, fixations rarely exceed a concept influence of 0.1 which assesses their influence on model prediction to be very limited. This holds true across all datasets and attribution methods. We can conclude that according to the chosen evaluation method, saccades have a much bigger concept influence and thus their velocity profiles contain more information with respect to the problem setting of biometric identification than it is the case for fixations.

## 4.2 Saccade Sub-Event Dissection

We further present the results for the event dissection experiment where we dissect saccades into sub-events as shown in Figure 4. The distribution of the sub-event segmentation sizes can be found in Figure 3 of the Appendix. Across all datasets and attribution methods, we observe the highest concept influence for samples belonging to the peak phase of the saccadic profile. The concept influences of rise and fall phases are both about half as high as the peak phase. We note close to no concept influence for the pre phase, but a moderately influential post phase which can be associated with the occurence of post-saccadic oscillations.

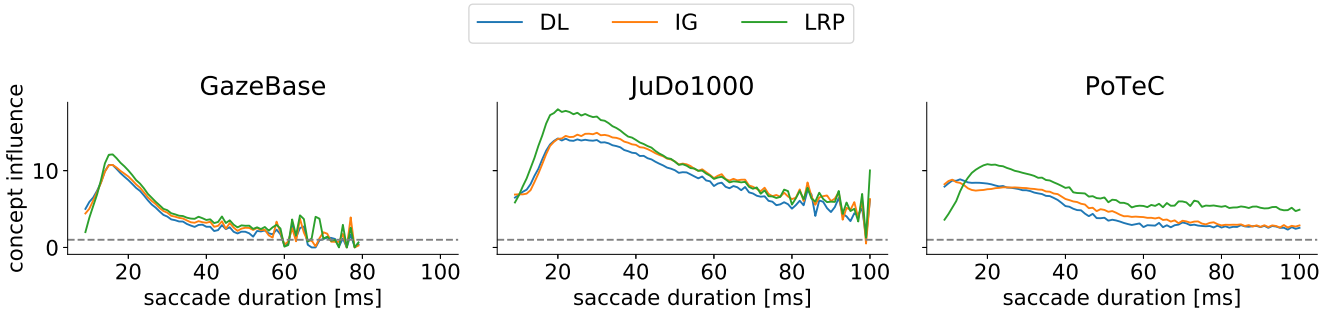
## 4.3 Event Property Binning

In this subsection we study the effect of different event properties on the resulting concept influence. We select two properties for each event type: we study the duration and amplitude of saccades and the dispersion and velocity standard deviation of fixations. The distribution of the segmentation sizes across these properties can be found in Figure 4 of the Appendix.

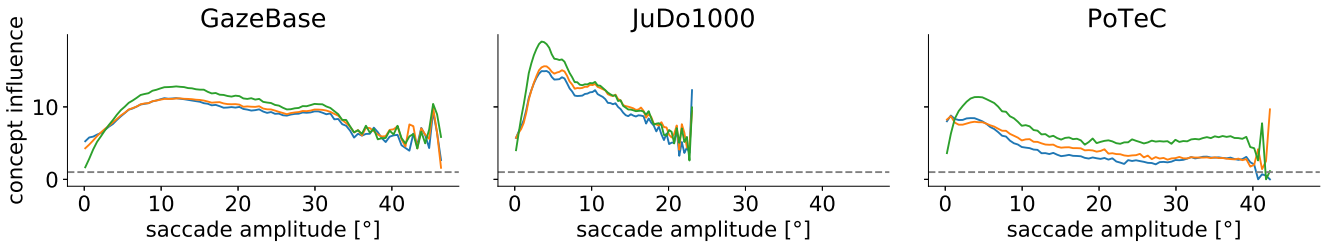
Starting with the duration of saccades in Figure 5(a), we observe the highest concept influences for saccades with a duration of about 20 ms. We further observe a concept influence peak for saccade amplitudes below 10° on the JuDo1000 and PoTeC datasets in Figure 5(b). Regarding the GazeBase dataset the peak is much less pronounced and is slightly higher at about 10°. Continuing with properties of fixation events, we mostly see flat curves and depending on the data set we spot rare outliers on the upper bounds. In the case of fixation dispersion in Figure 5(c), we observe concept influences above 1 solely on high dispersion outliers of the PoTeC dataset. In the case of the standard deviation of velocities during fixations presented in Figure 5(d), we observe a rise in concept influence on high standard deviations, but concept influences above 1 are solely reached on the JuDo1000 dataset.

# 5 DISCUSSION & CONCLUSION

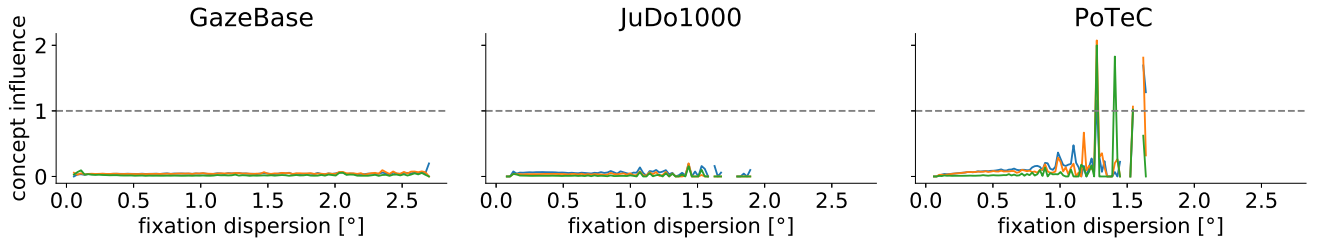
We have demonstrated the feasibility of evaluating the *concept influence* of gaze event types to gain insights on which parts of a gaze sequence is most relevant for the classification process of a state-of-the-art biometric model. We observed high concept influences for saccades with the peak phase of a saccade event to be especially influential. In contrast, fixations exhibit neglible concept influences with the exception of fixations with a high dispersion or a high standard deviation in velocity during fixation.



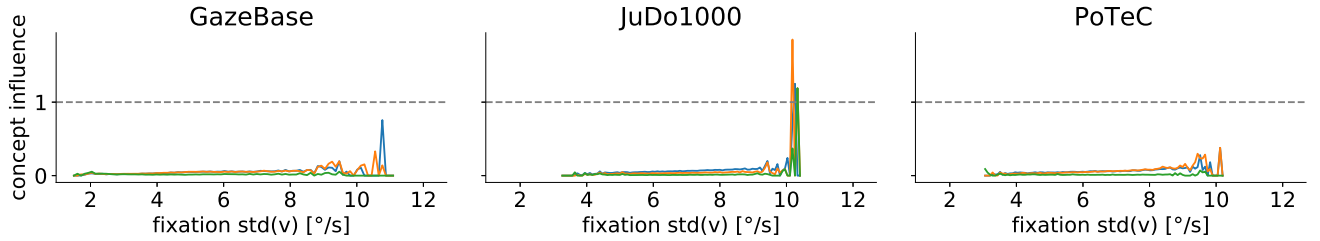

(a) Property binning by saccade duration.




(b) Property binning by saccade amplitude.




(c) Property binning by fixation dispersion.




(d) Property binning by fixation velocity standard deviation.

**Figure 5: Results for the event property binning experiment. The dashed gray horizontal line represents a concept influence of 1.**

Although the specific results of this study are very much constrained to the oculomotoric biometric setting, this work serves as a frame work for further research on the explainability of deep neural sequence models that consume gaze time-series data. This way we can harness the best of both worlds: top performance from neural networks and interpretable insights from descriptive concepts.

## ACKNOWLEDGMENTS

This work was partially funded by the German Federal Ministry of Education and Research under grant 01IS20043.

## REFERENCES

Mariano Alcañiz, Irene Alice Chicchi-Giglioli, Lucía A. Carrasco-Ribelles, Javier Marín-Morales, Maria Eleonora Minissi, Gonzalo Teruel-García, Marian Sirera, and Luis Abad. 2021. Eye gaze as a biomarker in the recognition of autism spectrum disorder

using virtual reality and machine learning: A proof of concept for diagnosis. *Autism Research* 15, 1 (nov 2021), 131–145. https://doi.org/10.1002/aur.2636

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2017. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. *arXiv preprint arXiv:1711.06104* (2017).

Christopher J. Anders, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. 2021. Software for Dataset-wide XAI: From Local Explanations to Global Insights with Zennit, CoRelAy, and ViRelAy. *arXiv preprint arXiv:2106.13200* (2021).

Richard Andersson, Linnea Larsson, Kenneth Holmqvist, Martin Stridh, and Marcus Nyström. 2016. One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms. *Behavior Research Methods* 49, 2 (2016), 616–637.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* 10, 7 (2015), 1–46.

Shuwen Deng, Paul Prasse, David R. Reich, Sabine Dziemian, Maja Stegenwallner-Schütz, Daniel Krakowczyk, Silvia Makowski, Nicolas Langer, Tobias Scheffer, and Lena A. Jäger. 2023. Detection Of ADHD Based On Eye Movements During Natural Viewing. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part VI* (Grenoble, France). Springer-Verlag, Berlin, Heidelberg, 403–418. https://doi.org/10.1007/978-3-031-26422-1_25

R. Engbert and R. Kliegl. 2003. Microsaccades uncover the orientation of covert attention. *Vision Research* 43 (2003), 1035–1045.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery* 33, 4 (mar 2019), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Henry Griffith, Dillon Lohr, Evgeny Abdulin, and Oleg Komogortsev. 2021. GazeBase, a large-scale, multi-stimulus, longitudinal eye movement dataset. *Scientific Data* 8 (2021).

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. 2023. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research* 24, 34 (2023), 1–11. http://jmlr.org/papers/v24/22-0142.html

Corey D. Holland and Oleg V. Komogortsev. 2013. Complex eye movement pattern biometrics: Analyzing fixations and saccades. In *2013 International Conference on Biometrics (IJCB)*. 1–8.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures.* Oxford University Press.

Lena A. Jäger, Silvia Makowski, Paul Prasse, Sascha Liehr, Maximilian Seidler, and Tobias Scheffer. 2019. Deep Eyedentification: Biometric Identification Using Micro-Movements of the Eye. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II* (Würzburg, Germany). Springer-Verlag, Berlin, Heidelberg, 299–314. https://doi.org/10.1007/978-3-030-46147-8_18

Ming Jiang and Qi Zhao. 2017. Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 3267–3276.

Lena A. Jäger, Thomas Kern, and Patrick Haller. 2021. Potsdam Textbook Corpus (PoTeC): Eye tracking data from experts and non-experts reading scientific texts. DOI: 10.17605/OSF.IO/DN5HP.

Alexandra P. Key, Courtney E. Venker, and Micheal P. Sandbank. 2020. Psychophysiological and Eye-Tracking Markers of Speech and Language Processing in Neurodevelopmental Disorders: New Options for Difficult-to-Test Populations. *American Journal on Intellectual and Developmental Disabilities* 125, 6 (nov 2020), 465–474. https://doi.org/10.1352/1944-7558-125.6.465

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2017. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). (2017). https://doi.org/10.48550/ARXIV.1711.11279

Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. 2020. Towards Best Practice in Explaining Neural Network Decisions with LRP. *International Joint Conference on Neural Networks* (2020), 1–7.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896* (2020).

Daniel Krakowczyk, David Robert Reich, Paul Prasse, Sebastian Lapuschkin, Lena Ann Jäger, and Tobias Scheffer. 2022. Selection of XAI Methods Matters: Evaluation of Feature Attribution Methods for Oculomotoric Biometric Identification. In *NeuRIPS 2022 Workshop on Gaze Meets ML*. https://openreview.net/forum?id=GOLdDAP2AtI

Daniel G. Krakowczyk, David R. Reich, Jakob Chwastek, Deborah N. Jakobi, Paul Prasse, Assunta Süss, Oleksii Turuta, Paweł Kasprowski, and Lena A. Jäger. 2023. pymovements: A Python Package for Processing Eye Movement Data. In *2023 Symposium on Eye Tracking Research and Applications* (Tubingen, Germany) *(ETRA '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3588015.3590134

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (Eds.), Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Ayush Kumar, Prantik Howlader, Rafael Garcia, Daniel Weiskopf, and Klaus Mueller. 2020. Challenges in Interpretability of Neural Networks for Eye Movement Data. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Short Papers)*. Association for Computing Machinery, New York, NY, USA, Article 12, 5 pages. https://doi.org/10.1145/3379156.3391361

Dillon Lohr, Henry Griffith, and Oleg V Komogortsev. 2021. Eye Know You: Metric Learning for End-to-end Biometric Authentication Using Eye Movements from a Longitudinal Dataset. https://doi.org/10.48550/ARXIV.2104.10489

Dillon Lohr and Oleg V Komogortsev. 2022. Eye Know You Too: Toward Viable End-to-End Eye Movement Biometrics for User Authentication. *IEEE Transactions on Information Forensics and Security* 17 (2022), 3151–3164.

Silvia Makowski, Lena A. Jäger, Paul Prasse, and Tobias Scheffer. 2020. JuDo1000 Eye Tracking Data Set. DOI: 10.17605/OSF.IO/5ZPVK. , 10 pages.

Silvia Makowski, Paul Prasse, David R. Reich, Daniel Krakowczyk, Lena A. Jäger, and Tobias Scheffer. 2021. DeepEyedentificationLive: Oculomotoric Biometric Identification and Presentation-Attack Detection Using Deep Neural Networks. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 4 (2021), 506–518.

Susana Martinez-Conde, Stephen L. Macknik, and David H. Hubel. 2004. The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience* 5 (2004), 229–240.

Susana Martinez-Conde, Stephen L. Macknik, Xoana G. Troncoso, and Thomas A. Dyar. 2006. Microsaccades Counteract Visual Fading during Fixation. *Neuron* 49 (2006), 297–305.

Christoph Molnar. 2022. *Interpretable Machine Learning* (2 ed.). https://christophm.github.io/interpretable-ml-book

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. *Layer-Wise Relevance Propagation: An Overview*. Springer International Publishing, 193–209.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition* 65 (2017), 211–222.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73 (2018), 1–15.

Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. 2019. Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access* 7 (2019), 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

Peter Raatikainen, Jarkko Hautala, Otto Loberg, Tommi Kärkkäinen, Paavo Leppänen, and Paavo Nieminen. 2021. Detection of developmental dyslexia with machine learning using eye movement data. *Array* 12 (2021), 100087.

Keith Rayner, Jane Ashby, Alexander Pollatsek, and Erik D. Reichle. 2004. The Effects of Frequency and Predictability on Eye Fixations in Reading: Implications for the E-Z Reader Model. *Journal of Experimental Psychology: Human Perception and Performance* 30, 4 (2004), 720–732. https://doi.org/10.1037/0096-1523.30.4.720

Keith Rayner and Alexander Pollatsek. 1983. Is visual information integrated across saccades? *Perception & Psychophysics* 34, 1 (jan 1983), 39–48. https://doi.org/10.3758/bf03205894

Ioannis Rigas, Lee Friedman, and Oleg Komogortsev. 2018. Study of an extensive set of eye movement features: Extraction methods and statistical analysis. *Journal of Eye Movement Research* 11, 1 (2018).

Ioannis Rigas, Oleg Komogortsev, and Reza Shadmehr. 2016. Biometric recognition via eye movements: Saccadic vigor and acceleration cues. *ACM Transactions on Applied Perception* 13, 2 (2016), 1–21.

Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying Fixations and Saccades in Eye-Tracking Protocols. In *ETRA '00* (Palm Beach Gardens, Florida, USA) *(ETRA '00)*. Association for Computing Machinery, New York, NY, USA, 71–78. https://doi.org/10.1145/355017.355028

Abraham. Savitzky and M. J. E. Golay. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36, 8 (1964), 1627–1639.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3145–3153.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. *arXiv preprint arXiv:1605.01713* (2016).

Mikhail Startsev and Raimondas Zemblys. 2022. Evaluating Eye Movement Event Detection: A Review of the State of the Art. *Behavior Research Methods* (jun 2022). https://doi.org/10.3758/s13428-021-01763-7

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. 2020. Visualizing the Impact of Feature Attribution Baselines. *Distill* 5, 1 (2020).

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3319–3328.

Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. 2022. Interpretable Semantic Photo Geolocation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 750–760.

Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, and Diego Andina. 2018. Deep Learning for Computer Vision: A Brief Review. *Intell. Neuroscience* 2018 (jan 2018), 13 pages. https://doi.org/10.1155/2018/7068349

Jonathan R. Williford, Brandon B. May, and Jeffrey Byrne. 2020. Explainable Face Recognition. https://doi.org/10.48550/ARXIV.2008.00916