Selection of XAI Methods Matters: Evaluation of Feature Attribution Methods for Oculomotoric Biometric Identification

Anonymous Author(s)
Affiliation
Address
email

Abstract

Substantial advances in oculomotoric biometric identification have been made due to deep neural networks processing non-aggregated time series data that replace methods processing theoretically motivated engineered features. However, interpretability of deep neural networks is not trivial and needs to be thoroughly investigated for future eye tracking applications. Especially in medical or legal applications explanations can be required to be provided alongside predictions. In this work, we apply several attribution methods to a state of the art model for eye movement-based biometric identification. To asses the quality of the generated attributions, this work is focused on the quantitative evaluation of a range of established metrics. We find that Layer-wise relevance propagation generates the most robust and least complex attributions, while DeepLIFT attributions are the most faithful. Due to the absence of a correlation between attributions of these two methods we advocate to consider both methods for their potentially complementary attributions.

1 Introduction

Eye movements are known to reflect cognitive processes that include attentional mechanisms [1, 2]; they are therefore considered to be a *window on the mind and brain* [3]. Eye movements can serve as a basis to automatically screen for ADHD [4], dyslexia [5, 6], and autism spectrum disorder [7]. Since eye movements are also known to be highly individual [8], they can be used as a biometric characteristic [9, 10, 11].

For medical screening applications, it is imperative that a machine-learning model can be understood to detect the actual evidence of the condition of interest rather than clever-hans signals or social biases inherent in the training data. Also for biometric identification, it is highly relevant to analyze which signals the model reacts to in order to understand vulnerabilities, biases, and aspects of fairness of the model [12].

In recent years, deep neural networks that process raw gaze-velocity data have achieved dramatic performance increases compared to machine learning on engineered features. For example, recent deep neural network architectures for oculomotoric identification reduces the time for an identification and the error rate by one order of magnitude compared to identification based on engineered saccadic features [11, 13]. Engineered eye-gaze features are often derived from findings from neurophysiological research [14, 15]. Analyses of feature importance of models that rely on such features are therefore meaningful for experts. For deep neural networks, many feature attribution methods have been developed [16, 17, 18, 19]. Unfortunately the interpretability of such complex models is hard and deep neural nets are well known to be black boxes due to the complexity of non-linear activations.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

While a wide range of quantitative performance metrics exist for explainability methods, there is no consensus about the merit of each of these metrics. Metrics quantify the complexity [20, 21, 22], faithfulness [23], robustness [24] or localization [25, 26] of attributions. Image data remains the most popular application domain of these approaches. Unlike gaze data, image data naturally lends itself well to human inspection of attributions, and therefore the evaluation of explainability usually relies in no small part on the plausibility of visualizations of attributions [27, 28].

Prior work on applying feature attributions to eye gaze based convolutional neural networks simply quantifies the overall importance of input channels by channel-wise aggregation of feature attributions [4]. Nevertheless there exists no extensive analysis of feature attributions applied to eye gaze based models up to this date. As a first step towards trusted models and explanations in deep learning based eye gaze applications, in this paper we will therefore rigorously evaluate feature attribution methods with respect to complexity, faithfulness, robustness, and consistency across methods.

In this paper we will restrict ourselves to the task of oculomotoric biometrics, as this is where these models exhibit top performance and it is intuitive that good explanations will also need good predictions[29]. We will train the state-of-the-art neural network *Eye Know You Too* [13] on the publically available *JuDo1000* [30] dataset. We will apply a range of feature attribution methods to the trained models, namely *DeepLIFT* [18], *Integrated Gradients* [17] and *Layer-Wise Relevance Propagation* [16], as well as *Input X Gradient* [31]. We will further quantitatively evaluate these attributions by a variety of established metrics to assess desired properties like complexity, sensitivity, faithfulness and robustness. We will finally evaluate the agreement of attributions across different attribution methods.

The main contributions of this papers are:

- we generate and visualize commonly used attribution methods for a state-of-the-art oculomotoric biometric model on a real-world dataset,
- we present the first work to evaluate attributions of oculomotoric biometric models on several established metrics,
- we evaluate the agreement between the generated attributions across the attribution methods.

2 Materials and Methods

This section is structured as follows: Subsection 2.1 introduces biometric identification and the biometric model under investigation, whereas Subsection 2.2 lists and briefly describes the applied attribution methods. In Subsection 2.3 we give an overview of the employed attribution metrics and in Subsection 2.4 we present the dataset on which the attribution methods will be evaluated. Subsection 2.5 lays out the data preprocessing steps applied to the dataset and Subsection 2.6 describes the underlying evaluation protocol.

2.1 Eye Tracking-Based Biometric Identification

We investigate the explainability of a state-of-the-art neural network model for oculomotoric biometric identification, namely *EyeKnowYouToo*, introduced by Lohr and Komogortsev [13]. Given a known population of individuals, we investigate a *multi-class classification* setting, where the model is trained using one or more sequence(s) of eye movement recordings $\{((x_0, y_0) \ldots, (x_n, y_n)\}$ of each user, where x_i and y_i are the yaw (horizontal) and pitch (vertical) gaze eye movement velocities. At test time, an eye gaze sequence recorded on a different day is used to recognize a user. *EyeKnowYouToo* is a DenseNet-based architecture [32] that uses the raw yaw and pitch angular velocities as input. This end-to-end dilated convolutional network is trained to minimize the categorical cross-entropy along with a multi-similarity loss. We use an extended version of Dillon Lohr's PyTorch implementation licensed under CC BY-NC-SA 4.0.

2.2 Attribution Methods

Attribution methods refer to local post-hoc explainability methods that attribute positive or negative contribution values to each input feature of a specific model prediction. This facilitates interpretability of given predictions by highlighting relevant parts of the input signal. Attribution methods can be

divided into perturbation-based and backpropagation-based methods [33]. Due to the computationally expensive approach of perturbation-based methods like SHAP [] or Occlusion [] we limit this study to the evaluation of the backpropagation-based attribution methods Input x Gradient (IxG) [31], Integrated Gradients (IG) [17], DeepLIFT (DL) [18] and Layer-wise relevance propagation (LRP) [16, 34, 35]. Figure 1 presents an example of generated attributions for each of the methods introduced in this subsection. We use the Captum library [36] for its IxG, IG and DL implementations and the Zennit library [37] for the implementation of LRP rules.

IxG is an early attribution method for which relevance is computed by backpropagating the prediction gradient with a final element-wise multiplication with the actual input.

LRP computes input relevance by backpropagating the model output to its input according to a specific set of rules. Relevance of each unit is passed down to the lower units depending on the product of activations and weights of the respective layer units and connections, while keeping the total relevance in each layer constant. Although over time a range of different relevance passing rules was proposed [25], we limit ourselves to the basic original LRP-ε rule for a more concise presentation. We set $\varepsilon = 0.25std$ according to the parameter selection in Montavon et al. [38]. High ε values will result in less attributions close to zero, and vice versa for lower ε values.

DeepLIFT [18] is a very similar backpropagation-based attribution method, but in contrast to LRP, an explicit baseline input is used for calculating activation reference points. Activation differences are then backpropagated layer by layer according to a set of rules.

Integrated Gradients (IG) [17] is somewhat different in that it computes the gradients of the model, which makes it implementation independent. It also uses an explicit baseline, which is then stepwise linearly interpolated into the actual input at hand. For each of those interpolations and for each input feature gradients are calculated, then integrated and finally multiplied with the feature difference between baseline and actual input.

One drawback of attribution methods which use an explicit baseline is the susceptibility for its choice [39]. Usually a zero or mean baseline is chosen for both of these baseline-based attribution methods, but theoretically every input which leads to a neutral output can be chosen. One drawback of using a constant baseline is the introduced low attribution bias to input values close to the baseline value.

2.3 Attribution Metrics

Apart from a qualitative visual analysis we will evaluate the generated attributions by different metrics to measure their complexity, faithfulness and robustness. Due to a lack of ground truth segmentation mask we omit the evaluation of attribution localization. We use the Quantus python package [40] for the calculation of all attribution metrics used in this paper.

Let $\mathbf{x}$ be an instance with d features, $\mathbf{f}$ our model and $\mathbf{g}$ an explanation function, where $\mathbf{g}(\mathbf{f}, \mathbf{x})$ refers to the feature attribution for the model prediction $\mathbf{f}(\mathbf{x})$. We define $\mathbf{g}(\mathbf{f}, \mathbf{x})_i$ as the attribution of the i-th input feature. The fractional contribution distribution is defined as $\mathbb{P}_{\mathbf{g}}(i) = \frac{|\mathbf{g}(\mathbf{f},\mathbf{x})_i|}{\sum_{j\in[d]}|\mathbf{g}(\mathbf{f},\mathbf{x})_j|}$, for $i \in [d]$, and the probability distribution by $\mathbb{P}_{\mathbf{g}} = \{\mathbb{P}_{\mathbf{g}}(i) \mid 0 \leq i \leq d\}$.

Complexity metrics As complexity measures we use entropy [20] (Equation 1), sparseness [21] (Equation 2) and effective complexity [22] in (Equation 3) that are defined below.

The entropy metric is defined as follows:

$$\mu_E(\mathbf{f}, \mathbf{g}, \mathbf{x}) = \mathbb{E}[-\ln(\mathbb{P}_{\mathbf{g}})] = -\sum_{k=1}^{d} \mathbb{P}_{\mathbf{g}}(k) \ln(\mathbb{P}_{\mathbf{g}}(k)) \tag{1}$$

The sparseness metric is based on the Gini coefficient and measures the dispersion between high and low attribution values. It is defined as follows:

$$\mu_S(\mathbf{f}, \mathbf{g}, \mathbf{x}) = 1 - 2\sum_{k=1}^{d} \frac{\mathbf{g}(\mathbf{f}, \mathbf{x})_k}{\|\mathbf{g}(\mathbf{f}, \mathbf{x})\|_1} \frac{d-k+0.5}{d} \tag{2}$$

Finally the effective complexity measures the amount of absolute attribution values above a certain threshold ε and is defined as follows:

$$\mu_{EC}(\mathbf{f}, \mathbf{g}, \mathbf{x}) = \frac{1}{d}\Big|\Big\{\, a \in \mathbf{g}(\mathbf{f}, \mathbf{x}) \mid a > \varepsilon \,\Big\}\Big| \tag{3}$$

Faithfulness metric To measure faithfulness we use the region perturbation metric proposed by Samek et al. [23]. In this metric we iteratively perturb the input instance $\mathbf{x}$ by non-overlapping patches which are ordered descendently by their sum of inner feature attribution values. The perturbed instance at step i is denoted as $\mathbf{x}^{(i)}_{\text{MoRF}}$.

By iterating through a number of steps N we can thus generate a mean perturbation curve for a set of instances. The underlying intuition is, that perturbing features with high-scoring attributions should lead to a steep drop in target output $\mathbf{f}(\mathbf{x})$ if the evaluated attributions are being actually faithful.

To account for the drop in target output due to model robustness properties, we create a baseline where the non-overlapping order of patches is random. The perturbed instance for step i using this random patch drawing method is denoted as $\mathbf{x}^{(i)}_{\text{Random}}$. The overall metric score is then quantified as the area between the ordered and the random perturbation curve (Equation 4).

$$\mu_{RP}(\mathbf{f}, \mathbf{g}, \mathbf{x}) = \frac{1}{N+1}\sum_{i=1}^{N} \mathbf{f}(\mathbf{x}^{(i)}_{\text{Random}}) - \mathbf{f}(\mathbf{x}^{(i)}_{\text{MoRF}}) \tag{4}$$

Robustness metrics As robustness metrics we use the local Lipschitz estimate [24] (Equation 5). This metric perturbs the complete input instance by superimposing noise and measuring the distance between between explanations generated for the perturbed and unperturbed input.

Denote the perturbation of an input instance $\mathbf{x}$ by $\hat{\mathbf{x}}$ and by $\mathcal{N}(\mathbf{x})$ x with added noise drawn from a Gaussian distribution with a mean of 0 and a standard deviation of 0.1. Reusing the notations introduced in the paragraph above, we can define the local Lipshitz estimate as follows:

$$\mu_L(\mathbf{f}, \mathbf{g}, \mathbf{x}) = \underset{\hat{\mathbf{x}} \in \mathcal{N}(\mathbf{x})}{\arg\max} \frac{\|\mathbf{g}(\mathbf{f}, \mathbf{x}) - \mathbf{g}(\mathbf{f}, \hat{\mathbf{x}})\|_2}{\|\mathbf{x} - \hat{\mathbf{x}}\|_2} \tag{5}$$

2.4 Dataset

We evaluate the model and attributions on the binocular *JuDo1000* [30] dataset, which is sampled at a frequency of 1000 Hz. The dataset consists of 150 subjects recorded in 4 sessions that are least two weeks apart from each other. During each session the participants are instructed to visually follow five randomly placed dots on the screen.The intervals between each subsequent dot presentation range between 250 ms and 1 s.

2.5 Data Preprocessing

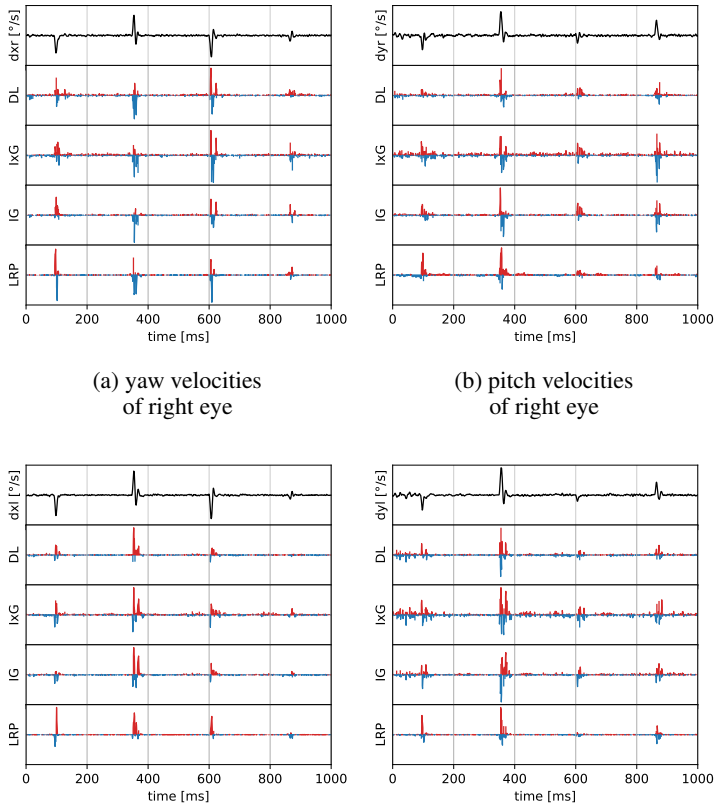
We base our data preprocessing pipeline on the method proposed by Lohr & Komogortsev [13]. We first transform positional data into velocity data by applying the Savitzky-Golay differentiation filter [41] with a window size of 7 and an order of 2. We create non-overlapping subsequences with a rolling window approach where we use a window size of 1 second (1000 samples @ 1000 Hz). We exclude all subsequences which need padding or which include more than 50% of missing values and clamp all velocities to $\pm 1000\ ^\circ/s$. Further, we apply z-score normalization and finally replace all missing values with 0.

2.6 Evaluation Protocol

In order to evaluate the introduced attribution methods for this specific task and dataset we apply the following protocol: We split the JuDo1000 dataset by a leave-one-session-out scheme, where each fold includes a complete session as a test set and the remaining three sessions as the training set. This results in a 4-fold cross validation protocol to which we adhere for the complete evaluation pipeline. Model accuracy as well as attribution metrics are evaluated on the test set of each fold only.

We take the predicted class as the target class to create all attributions. We normalize attribution values by the maximum absolute attribution value of the respective instance.

We evaluate every setting on a AMD EPYC 7742 CPU and a NVIDIA DGX A100 GPU. We train all models using the PyTorch [42] library utilizing the NVIDIA CUDA platform. We implement the model evaluation framework using the scikit-learn [43] machine learning package. The code can be found online.[1]



(a) yaw velocities of right eye

(b) pitch velocities of right eye

(c) yaw velocities of left eye

(d) pitch velocities of left eye

Figure 1: Attributions generated by the employed attribution methods (see Subsection 2.2) for a single example instance. Each subfigure represents one of the four input velocity channels (from upper left to lower right: a) yaw right eye, b) pitch right eye, c) yaw left eye, d) pitch right eye). The respective channel signals are plotted as a continuous black line in the first row of each subfigure with y-axis scale from -1000 to 1000 $^\circ/s$. The remaining rows depict the generated feature attributions for the method labeled at the y-axis. Red represents positive attributions and blue represents negative attributions. All attributions are normalized in the range between 0 and 1.

3 Results

Literature has shown, that we need high performing model to get reliable explanations [29]. To test the performance of our model, we followed the evaluation protocol described in Section 2.6, where we train our model to identify 150 different identities in a multi-class setting. From our experiments we can conclude, that our model has a high accuracy of $92.5 \pm 0.6\%$, yielding state-of-the-art performance [13]. The remaining section is structured as follows: Section 3.1 evaluates the different attribution methods in a qualitative manner, where the attribution methods are evaluated in a

[1] `anonymousURL`

quantitative manner in Sections 3.2-3.4. In Section 3.5 we evaluate the agreement of the different attribution methods.

3.1 Qualitative Attribution Analysis

We start the evaluation of attributions by a qualitative visual analysis to put the subsequent quantitative metrics into perspective. Due to the lack of space we unfortunately have to limit the presentation to the single instance given in Figure 1. We refer the interested reader to a selection of additional instances in Appendix A to get a more complete impression of the generated feature attributions. Apart from the cases where we specifically point to a feature of the given example instance the general observations hold true for the vast amount of the instances of the dataset.

We first start with a short inspection of the input signal. There exist absolute velocity peaks at the time steps at roughly 100 ms, 350 ms, 600 ms and 850 ms. Although the peaks are reasonably time-aligned across channels, they are sized different depending on the respective input channel. This stems from differences in yaw and pitch velocity of the underlying eye movement event but also from deviations between both eyes. We further observe some low velocity oscillations above the noise floor in the pitch channels of both eyes during the first 100 ms.

When observing the generated attributions we first note the tendency of very high absolute attribution values being time-aligned with the previously identified high velocity eye movements. Still, depending on the attribution method, lower mid attribution values are clearly present in-between these high velocity eye movements. Only LRP exhibits close to zero attribution values at almost all input features not in the temporal vicinity of the identified velocity peaks. We additionally note that the three other attribution methods attribute lower mid attribution values to the previously identified low velocity oscillations during the first 100 ms of both pitch velocity channels, especially from the left eye.

Regarding the consistency of top attribution values among the attribution methods, we observe that DL, IxG and IG attribute the yaw velocity of the left eye at about 350 ms as the most important input feature, while LRP identifies the time step at about 150 ms in the same channel as the most important feature.

Another aspect that we note is the seeming ambivalence between positive and negative attributions. Positive attributions express a positive influence of a specific input feature towards the target class, and vice versa with negative attributions. We can observe some consistency regarding this among the attribution methods, for example we see positive attribution during the rise and fall of the velocity peak at around 600 ms (left eye yaw velocity channel in Figure 1c) and negative attribution during the peak itself. We notice this ambivalence also across channels, for example the LRP attributions during the velocity peak at about 100 ms have opposing signs for the rise and fall of the input velocity profile. Taking the mean of attribution values across input channels for each time step would mitigate this issue though, as the positive attributions of each channel outweigh the negative ones at this time step.

3.2 Attribution Complexity

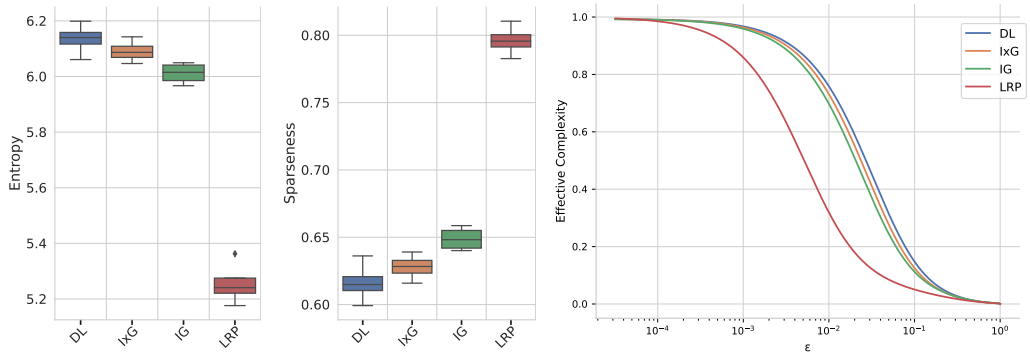
We present the results of the quantitative complexity metrics regarding entropy μ_E, sparseness μ_S and effective complexity μ_{EC} in Figure 2.

For all three complexity metrics we observe that LRP sets itself apart from the other evaluated methods. LRP attributions consistently exhibit less entropy and are more sparse. We further notice distinctively less attribution values between 0.0001 and 0.2 than it is the case for the other methods.

Regarding the mean values for each attribution method we observe the exact same rank order across all three complexity metrics. In regard to this study LRP attributions are the least complex, followed by IG, IxG and finally DL. We note though that the variance across folds is higher for LRP and DL than the other two attribution methods.

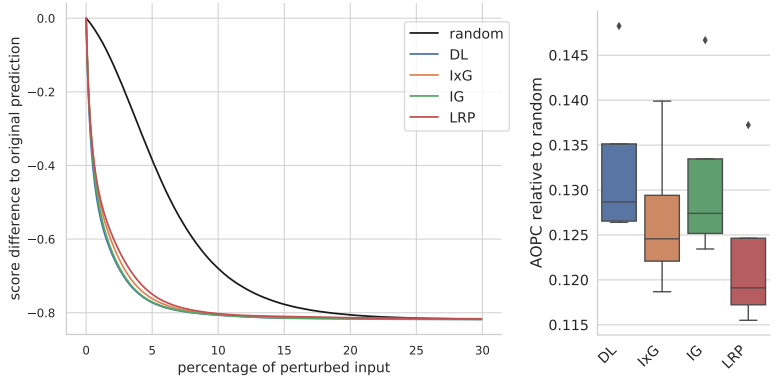
3.3 Attribution Faithfulness

We continue our quantitative attribution evaluation with the faithfulness measure of the region perturbation metric μ_{RP}. As described in the respective metric paragraph in Subsection 2.3 we increasingly perturb the input instance on non-overlapping patches and measure the model output difference for the respective target output.



(a) Entropy μ_E (b) Sparseness μ_S (c) Effective Complexity μ_{EC}

Figure 2: Attribution complexity metrics. A) Entropy (the lower the better), B) sparseness (the higher the better), C) effective complexity (the lower the curve the better). See Section 2.3 for metric definitions.



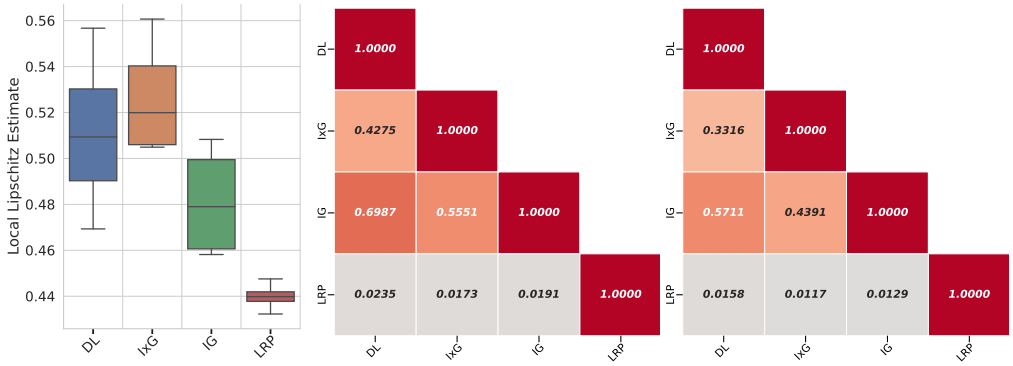
(a) Mean pertubation curves (b) Boxplot for the AOPCs (μ_{RP})

Figure 3: Region perturbation. a) Mean perturbation curves for the employed attribution methods. The mean random perturbation curve is plotted as a continuous black line. The greater the area between the random perturbation curve and the attribution perturbation curve (AOPC relative to random) the better. b) Boxplot for the AOPCs relative to the random perturbation curve. The higher the better. See Section 2.3 for metric definitions.

Figure 3a depicts the mean perturbation curves for each attribution method together with the random perturbation curve. We observe a very similar curve shape across all attribution methods, with LRP being slightly less steep and DL and closely next IG being steeper than IxG. This is in concordance with the boxplot in Figure 3b. The perturbation curves of each method converge to the random perturbation curve after about 23 % of perturbed input. However, the boxplot exposes for all methods a relatively high variance and except for IxG one better performing outlier fold. We note that the mean values for all attribution methods span an interval that is less than 1.5 times the interquartile range.

3.4 Attribution Robustness

We evaluate the attribution methods on the robustness metric μ_L, which measures the difference in attributions on noise superimposition across the whole input. Figure 4a depicts the respective metric results. We observe the lowest attribution difference and thus highest attribution robustness for LRP.



(a) Local Lipschitz estimate μ_L (b) Spearman correlation (c) Kendall correlation

Figure 4: Attribution robustness and correlation metrics. a) Local Lipschitz estimate (the lower the better). b) Spearman's ρ coefficient. c) Kendall's τ coefficient. All correlation coefficients are significant with a p-value below $\alpha = 0.001$. See Section 2.3 for metric definitions.

Moreover LRP also exhibits by far the lowest variance across folds. IG is the second most robust method followed by DL and IxG.

3.5 Agreement Across Attribution Methods

We finally evaluate the agreement across attribution methods by correlation analysis using Spearman's ρ coefficient and Kendall's τ coefficient. The corresponding correlation matrices are depicted in Figure 4b and 4c.

We note that there is next to no attribution correlation between LRP and the other three methods. The highest correlation is between DL and IG, followed by the correlation between IxG and IG and IxG and DL. Kendall correlations are consistently lower than Spearman correlations while ranks are nevertheless preserved between both correlation methods.

4 Discussion

We quantitatively evaluated the four attribution methods DeepLIFT (DL), Input x Gradient (IxG), Integrated Gradients (IG) and Layer-wise Relevance Propagation (LRP) for complexity, faithfulness and robustness on the real world dataset JuDo1000 and the biometric model Eye Know You Too.

Although we identified LRP to create the least complex and most robust attributions, their faithfulness was slightly lacking in relation to the other methods. This seems like being a trade-off between complexity and robustness on one hand, and faithfulness on the other. Less complex attributions will probably miss some important relevance relative to more complex attributions. And yet we can also imagine scenarios in which attributions are that complex that they include features that were not actually important for the model decision.

The challenge which has to be tackled for each single eye tracking application each time again, is therefore the assessment which of these aspects attract higher priority. This also depends on the recipient of the explanations. Clinical experts will be for example potentially more able to interpret complex attributions than lay persons.

Moreover it will be interesting to see if we can tune some of these attribution methods in such a way, that an optimal trade-off between certain metrics can be found. Especially LRP with the scalar ϵ parameter as well as its additional layer rules seems a promising candidate for such an undertaking. Further work will also have to be required in assessing the model influence on the resulting attributions.

We have further shown that LRP has close to no correlation with the other attribution methods when it comes to the two employed rank correlation coefficients Spearman's ρ and Kendall's τ. We

can explain some of this due to the lower prevalence of lower attribution values above a certain epsilon in LRP. As a great amount of the input features are attributed low non-zero attribution values using the other methods, this will result in a pronounced decline of correlation if the rank order is not met. Although the non-correlation can seem as an issue at first, it can also be beneficial for some applications to have uncorrelated attributions due to the prospect of highlighting features that complement each other. From the correlations we conclude that LRP can complement DL or IG well enough. Due to the slightly higher faithfulness of DL we would advocate for an ensemble of LRP and DL.

Regarding the qualitative analysis of the attribution methods we found a tendency for time-alignment of high velocity peaks and high absolute attribution values for all four attribution methods. This is in concordance with previous literature in which the predictive quality of high velocity eye movement statistics is shown to be high [44, 45, 46]. Nevertheless we can also identify low velocity regions which don't correspond to common eye movement types that still exhibit lower to mid attribution values across DL, IxG and IG. We therefore expect that applying the discussed attribution methods to current and future eye tracking applications will further enhance the still ongoing visual analysis in the field and can potentially help in discovering of new types of eye movement features.

We further identified some limitations of this work. First and most obvious, qualitative visual analysis can only analyze a small subset of the data due to limited human resources. This leaves room for undiscovered phenomena and the criticism of cherry picking example instances. On the one hand qualitative visual analysis by human experts cannot be completely replaced when evaluating attributions, as humans will be the recipients of explanations and there is no way generelly predict human preference on such a broad research problem.

On the other hand, some aspects of the undertaken qualitative analysis can be performed computationally, especially the time-alignment analysis between high absolute velocities and attribution values. Moreover, feature attributions by themselves lack interpretability, especially in tasks where model exhibit better-than-human performance due to the complexity of the input space. As future work we therefore propose to employ eye movement detection algorithms to quantify attribution localizations in regard to these human interpretable features which can be extracted by computational models.

Last but not least, we identified attribution ambivalence for all four attribution methods, where positive and negative attribution values are in close vicinity. This currently leads to issues in interpretability, as will be hard to explain these contradictory explanations. Issues in interpretability of these attributions are especially severe for models with subpar prediction performance, as provision of explanations usually biases the recipient towards accepting the decision. Further analysis will be needed to correctly assess this issue together with its root cause.

5 Conclusion

We have quantitatively evaluated the introduced attribution methods in regard to complexity, faithfulness and robustness. While Layer-wise relevance propagation exhibits low complexity and high robustness, attributions generated by DeepLIFT are most faithful. Due to the non-correlation of both methods we advocate for considering both methods for their potentially complementary attributions.

Although we identify similarities across attribution methods through visual analysis and quantitive metrics, we also clearly see differences and can conclude that the selection of the respective attribution method will have decisive influence on conclusions derived from such data.

This work therefore is the starting point and a possible baseline for a line of future research in the eye tracking community. We see future work regarding the tuning of attribution methods and models for achieving better metric results, and improving on human interpretability of attributions through existing eye movement concepts in the psychological literature. We propose that future work on models in eye tracking research increasingly includes measures of explainability to their model evaluation protocols to facilitate and assess the usefulness of the systems in real world problems.

References

[1] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognitive Psychology*, vol. 12, no. 4, pp. 441–480, 1976.

[2] J. M. Henderson, "Human gaze control during real-world scene perception," *Trends in Cognitive Sciences*, vol. 7, no. 11, pp. 498–504, 2003.

[3] R. P. G. van Gompel, M. H. Fischer, W. S. Murray, and R. L. Hill, eds., *Eye movements: A window on mind and brain*. Elsevier, 2007.

[4] S. Deng, P. Prasse, D. R. Reich, S. Dziemian, M. Stegenwallner-Schütz, D. Krakowczyk, S. Makowski, N. Langer, T. Scheffer, and L. A. Jäger, "Detection of ADHD based on eye movements during natural viewing," 2022.

[5] M. Nilsson Benfatto, G. Öqvist Seimyr, J. Ygge, T. Pansell, A. Rydberg, and C. Jacobson, "Screening for dyslexia using eye tracking during reading," *PloS one*, vol. 11, no. 12, p. e0165508, 2016.

[6] P. Raatikainen, J. Hautala, O. Loberg, T. Kärkkäinen, P. Leppänen, and P. Nieminen, "Detection of developmental dyslexia with machine learning using eye movement data," *Array*, vol. 12, p. 100087, 2021.

[7] M. Alcañiz, I. A. Chicchi-Giglioli, L. A. Carrasco-Ribelles, J. Marín-Morales, M. E. Minissi, G. Teruel-García, M. Sirera, and L. Abad, "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*, vol. 15, no. 1, pp. 131–145, 2022.

[8] D. Noton and L. Stark, "Scanpaths in eye movements during pattern perception," *Science*, vol. 171, no. 3968, pp. 308–311, 1971.

[9] P. Kasprowski and J. Ober, "Eye movements in biometrics," in *International Workshop on Biometric Authentication*, pp. 248–258, 2004.

[10] R. Bednarik, T. Kinnunen, A. Mihaila, and P. Fränti, "Eye-movements as a biometric," in *SCIA 2005*, pp. 780–789, 2005.

[11] S. Makowski, P. Prasse, D. R. Reich, D. Krakowczyk, L. A. Jäger, and T. Scheffer, "Deepeye-dentificationlive: Oculomotoric biometric identification and presentation-attack detection using deep neural networks," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 506–518, 2021.

[12] P. Prasse, D. R. Reich, S. Makowski, L. A. Jäger, and T. Scheffer, "Fairness in oculomotoric biometric identification," in *2022 Symposium on Eye Tracking Research and Applications*, ETRA '22, (New York, NY, USA), Association for Computing Machinery, 2022.

[13] D. Lohr and O. V. Komogortsev, "Eye know you too: Toward viable end-to-end eye movement biometrics for user authentication," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2022.

[14] R. Schleicher, N. Galley, S. Briest, and L. Galley, "Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired?," *Ergonomics*, vol. 51, no. 7, pp. 982–1010, 2008.

[15] I. Rigas, L. Friedman, and O. Komogortsev, "Study of an extensive set of eye movement features: Extraction methods and statistical analysis," *Journal of Eye Movement Research*, vol. 11, no. 1, 2018.

[16] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, pp. 1–46, 07 2015.

[17] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, PMLR, 06–11 Aug 2017.

[18] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153, PMLR, 06–11 Aug 2017.

[19] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *CoRR*, vol. abs/1706.03825, 2017.

[20] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and aggregating feature-based model explanations," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (C. Bessiere, ed.), pp. 3016–3022, International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[21] P. Chalasani, J. Chen, A. R. Chowdhury, S. Jha, and X. Wu, "Concise explanations of neural networks using adversarial training," in *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, JMLR.org, 2020.

[22] A.-p. Nguyen and M. R. Martínez, "On quantitative aspects of model interpretability," 2020.

[23] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2015.

[24] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," 2018.

[25] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, "Towards best practice in explaining neural network decisions with lrp," *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2020.

[26] J. Theiner, E. Müller-Budack, and R. Ewerth, "Interpretable semantic photo geolocation," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1474–1484, 2022.

[27] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[28] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[29] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, *The (Un)reliability of Saliency Methods*, pp. 267–280. Cham: Springer International Publishing, 2019.

[30] S. Makowski, L. A. Jäger, P. Prasse, and T. Scheffer, "Judo1000 eye tracking data set.," 2020.

[31] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," 2016.

[32] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016.

[33] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018.

[34] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[35] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[36] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," 2020.

[37] C. J. Anders, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, "Software for dataset-wide xai: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy," *CoRR*, vol. abs/2106.13200, 2021.

[38] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, *Layer-Wise Relevance Propagation: An Overview*, pp. 193–209. Cham: Springer International Publishing, 2019.

[39] P. Sturmfels, S. Lundberg, and S.-I. Lee, "Visualizing the impact of feature attribution baselines," *Distill*, vol. 5, jan 2020.

[40] A. Hedström, L. Weber, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, "Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations," 2022.

[41] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), pp. 8024–8035, Curran Associates, Inc., 2019.

[43] 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, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[44] C. Holland and O. V. Komogortsev, "Biometric identification via eye movement scanpaths in reading," in *2011 International Joint Conference on Biometrics (IJCB)*, pp. 1–8, 2011.

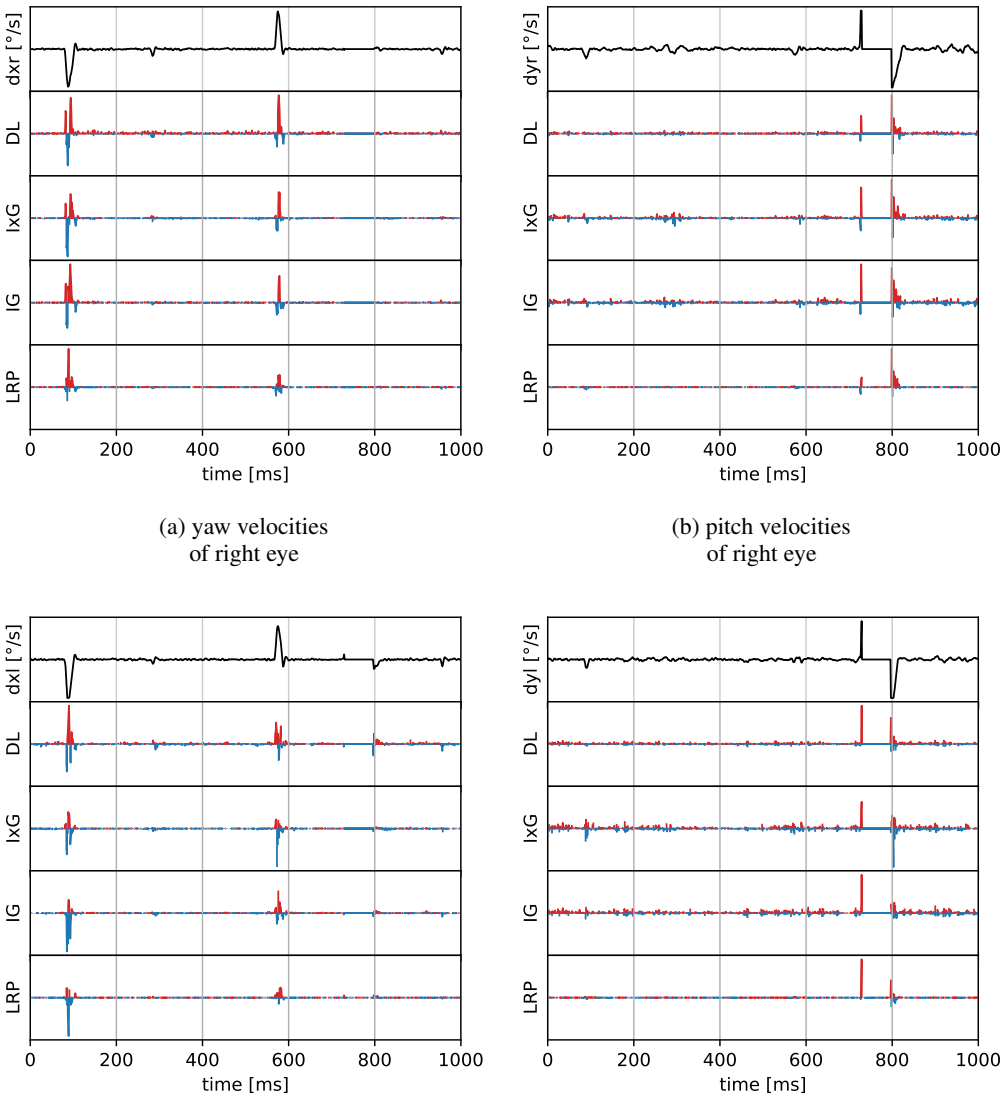
[45] C. D. Holland and O. V. Komogortsev, "Complex eye movement pattern biometrics: Analyzing fixations and saccades," in *2013 International Conference on Biometrics (ICB)*, pp. 1–8, 2013.

[46] I. Rigas, O. Komogortsev, and R. Shadmehr, "Biometric recognition via eye movements: Saccadic vigor and acceleration cues," *ACM Transactions on Applied Perception (TAP)*, vol. 13, no. 2, pp. 1–21, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discussed the limitations in Section 4.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discussed the potential societal impacts in Section 4. Wrong Explanations could falsely generate trust in model predictions. On the other hand could also possible observed biases in models used in negative ways (e.g. fairness issue as addressed in [12].
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We submitted the code as supplemental material and will open source the code and add the URL to the final version of this paper.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We addressed the specificatoin in Section 2.6.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use the Quantus package to evaluate all attributions [40] and used the publically available JuDo100 dataset [30]. We use the Captum library [36] for its DeepLIFT, IG and SmoothGrad implementations and the Zennit library [37] for the implementation of LRP rules and composites. We use a modified implementation of Dillon Lohr's Eye Know You Too model.
   (b) Did you mention the license of the assets? [No] We included the links/references to the used packages and will open source our implementation. Dillon Lohr's Eye Know You Too implementation is licensed under "Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License"
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The source code is added as supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All used packages/datasets are open source.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We investigated the identification using eye tracking data. The data we used (JuDo1000), does not contain any personal information of the individuals in the data. So we assume, that the persons are not identifiable in the wild. We introduced the dataset in Section 2.4.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional Attribution Examples



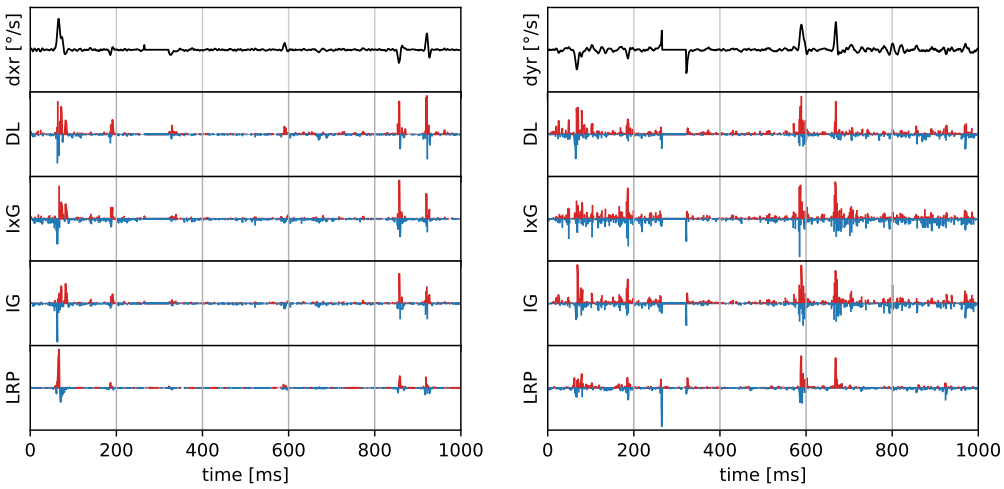
(a) yaw velocities of right eye

(b) pitch velocities of right eye

(c) yaw velocities of left eye

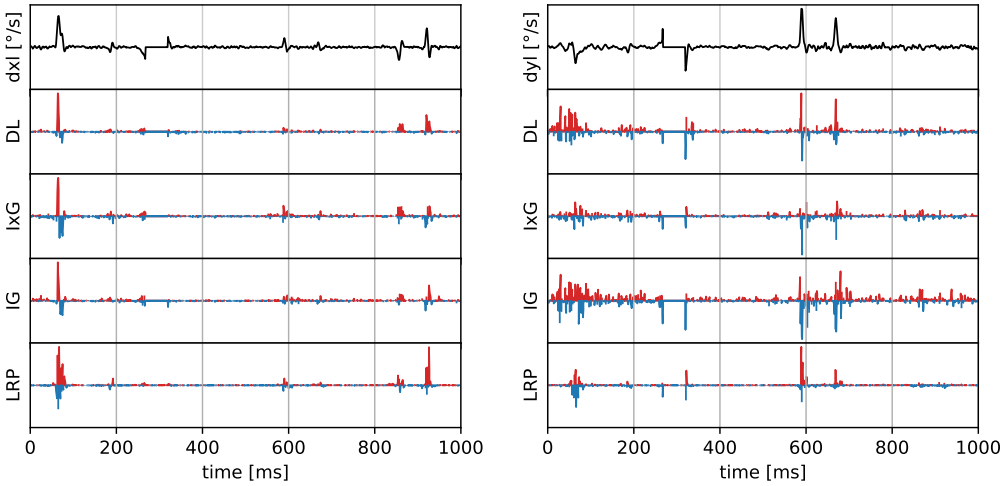
(d) pitch velocities of left eye

Figure 5: Attributions generated for a single example instance (id = 156) out of the JuDo1000 dataset. See caption in Figure1 for a complete description.



(a) yaw velocities
of right eye

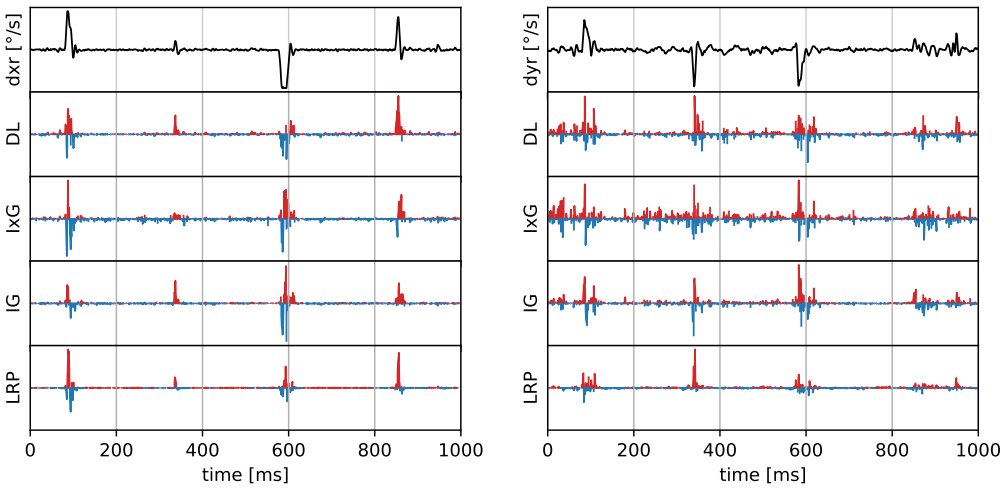
(b) pitch velocities
of right eye



(c) yaw velocities
of left eye

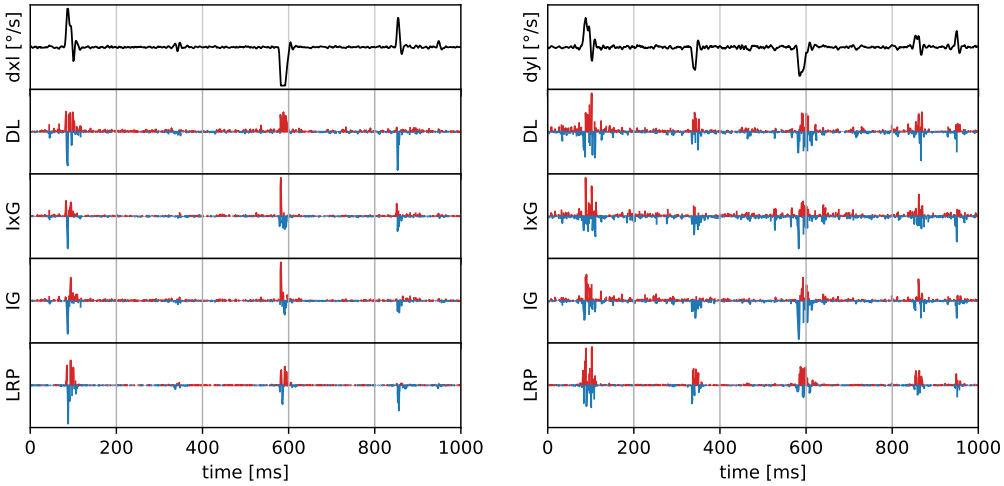
(d) pitch velocities
of left eye

Figure 6: Attributions generated for a single example instance (id = 1578) out of the JuDo1000 dataset. See caption in Figure1 for a complete description.



(a) yaw velocities
of right eye

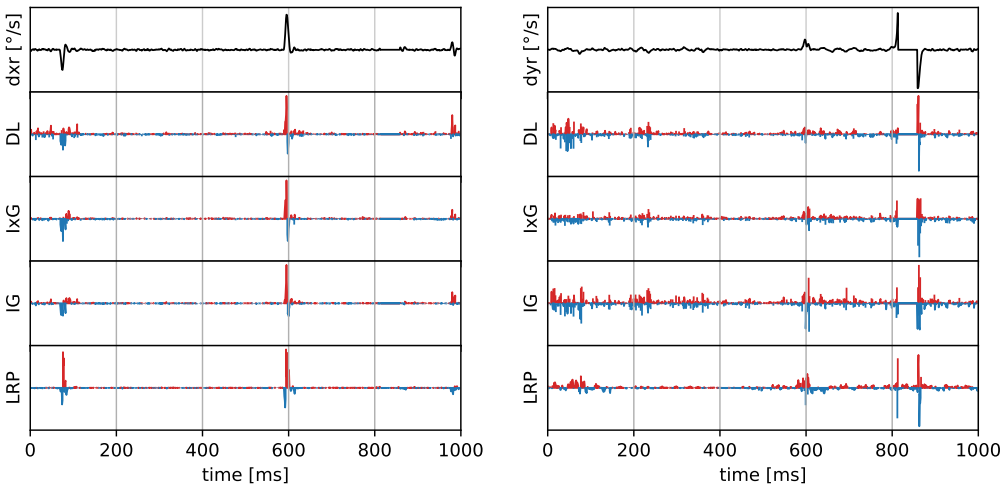
(b) pitch velocities
of right eye



(c) yaw velocities
of left eye

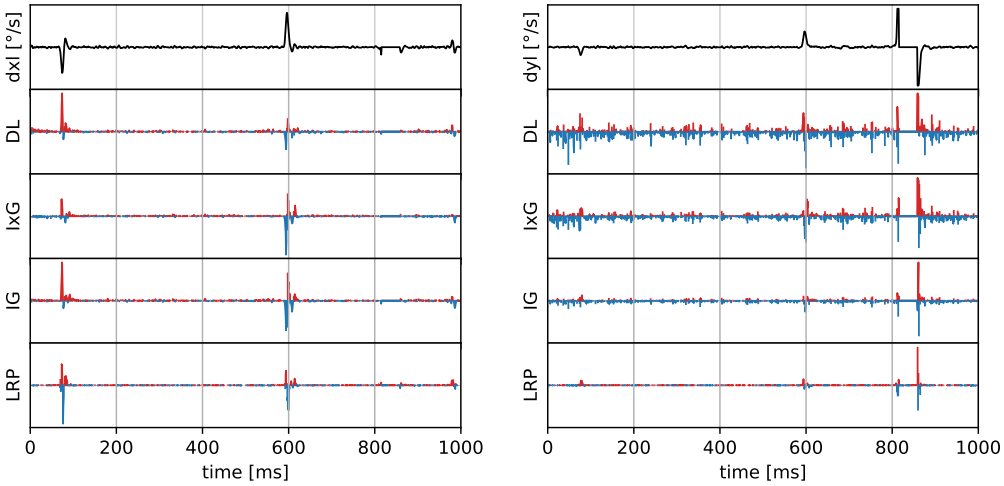
(d) pitch velocities
of left eye

Figure 7: Attributions generated for a single example instance (id = 6797) out of the JuDo1000 dataset. See caption in Figure1 for a complete description.



(a) yaw velocities
of right eye

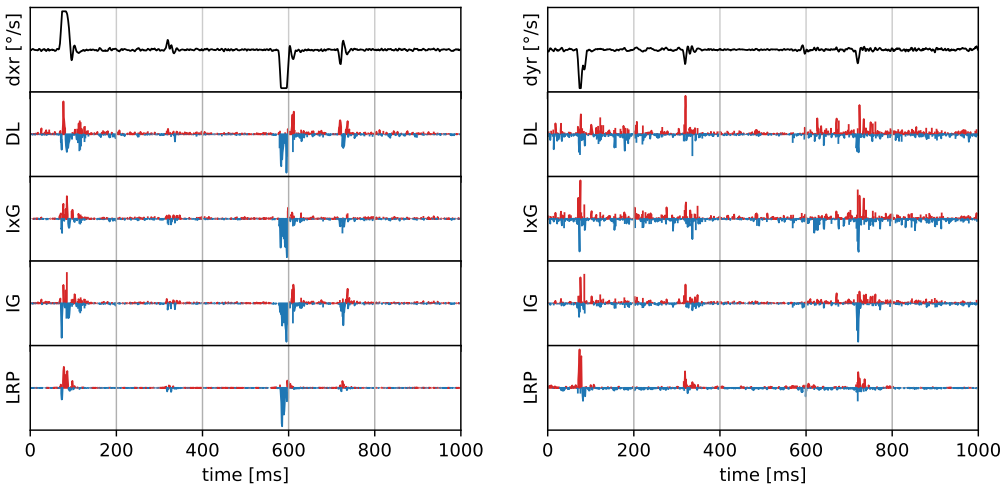
(b) pitch velocities
of right eye



(c) yaw velocities
of left eye

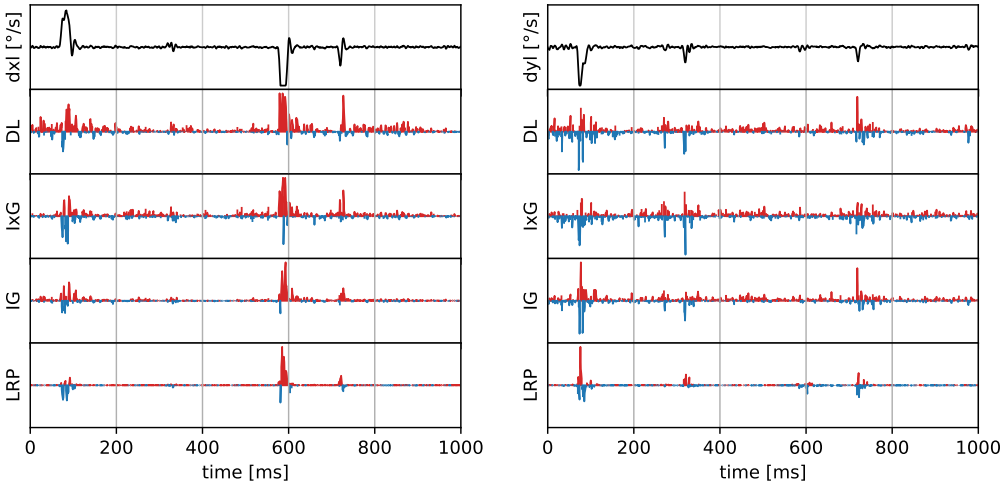
(d) pitch velocities
of left eye

Figure 8: Attributions generated for a single example instance (id = 16066) out of the JuDo1000 dataset. See caption in Figure1 for a complete description.



(a) yaw velocities
of right eye

(b) pitch velocities
of right eye



(c) yaw velocities
of left eye

(d) pitch velocities
of left eye

Figure 9: Attributions generated for a single example instance (id = 19239) out of the JuDo1000 dataset. See caption in Figure1 for a complete description.