forum acusticum 2026
8-12 September · Graz, Austria

LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATION (LIME) FOR PREDICTING LOUDNESS CLASSES FROM ELECTROENCEPHALOGRAPHY (EEG) DATA

Christian Koch[1]

[1]*Physikalisch Technische Bundesanstalt, Braunschweig, Germany*

Abstract

The utilization of artificial intelligence in critical and high-risk applications is poised to become a prevalent practice soon and the development of a methodology for testing and approving machine or deep learning models is imperative. As most experience has been accumulated in the medical field to date, this contribution employs a classification task using electroencephalography data to illustrate the preparation of reliable and effective approval and test methods.

A deep learning network was trained to assign auditory brainstem responses to three loudness classes. The generation of artificial data was achieved by utilizing audiological experimental reference data, with a local instance being defined to represent the normal hearing of a hypothetical test subject. The time series (trials) were created by employing a perturbation method. The single-layer perceptron model was selected amenable to an interpretation, with the weights indicating the data to which the network is sensitive. To validate a specified relevance parameter derived from the perceptron weights, the accuracy and a metric distance were calculated as performance measures. It has been demonstrated that the perceptron can characterize the relevance of specific data within a local environment surrounding the instance. Furthermore, conclusions have been drawn regarding the rendering of new test data.

Keywords: *electroencephalography, deep learning, loudness classification, LIME, approval methods*

Corresponding author: *christian.koch@ptb.de.*

1. Introduction

The integration of machine or deep learning models within critical infrastructure or high-risk applications is an imminent prospect. The development of a methodology for testing and approving such models, as well as the establishment of requirements for the Artificial Intelligence Act of the European Union, are imperative. This encompasses devices designed for quantitative evaluations, including measurements pertinent to legal or financial concerns, as well as analyzing data for safety or health applications.
Explainable AI methods have been extensively developed and comprise a broad spectrum of features and application possibilities. Nonetheless, these systems are highly specialized and require adaptation to suit any given application. Despite efforts to make them generally available for the systematic assessment of datasets and AI software [1], this has not yet been achieved. Furthermore, the testing methods employed for the purpose of device approval need to be applicable in a black-box manner and to be straightforward. This is to ensure that a testing body is able to efficiently handle them [2]. The key element in this process is the construction of test data, i.e. corner cases which can be used to evaluate the reliability of AI models and to identify their limitations.
To date, probably most experience with the development of test methods meeting these requirements has been gained in the medical area where several devices are already approved for a release [3]. In this contribution, medical data are used to develop a method for creating data to test a neural network (hereafter referred to as the "network-model"). A local interpretable model-agnostic explanation (LIME) procedure was adapted to a classification task for electroencephalography (EEG) data, and a particular representing model (hereafter referred to as the "explain-model") was created to support the

understanding of the decision process. Performance measures were defined, and the quality of the interpretation information from the explain-model was determined. Conclusions could be drawn where rendering the input data is possible to improve the sharpening of the test functionality.

2. Methods

2.1 Preparation of EEG data

The present study utilized artificially obtained data, which, however, is entirely based on experimental data [4]. Despite their essentially artificial nature, the data are derived from experimental sources and are representative of the characteristic behavior of audiological practice. Commencing from audiological reference points, master time series were deduced for three loudness classes: loud, medium loud, and silent. In order to obtain the EEG-voltage-time series (trials), the data was modified to represent the behavior of single test persons. The trials were processed without additional feature extraction to avoid any potential for bias. The data design is documented in an open-access repository, and the data used in this study can be downloaded.

The data were designed for a task to classify measurements into three loudness classes for an SPL of 30 dB, 50 dB, and 70 dB. For each class, 35 data sets containing 3,900 trials with 97 samples were generated, illustrating 35 test persons. The utilization of these EEG data was for the purpose of executing a classification task through the employment of a neural network. In [2] three distinct networks were designed and employed for a comparable task, albeit with wholly different data. Smooth results with a minimum computation time could be obtained from the feed-forward network (FFN) and it was used in this study as the network-model f to be tested. The training of the network was achieved by utilizing the data of 25 test persons, from which multiple training iterations were derived.

2.2 LIME procedure

A successful strategy for explaining the decision process of a neural network f is the distillation of knowledge to a simpler and, thus, interpretable model g [5]. A number of strategies for managing the inherent information reduction are conceivable. One such strategy is to restrict g to a locally limited area. Local Interpretable Model-agnostic Explanations (LIME) is a known method that implements this principle [6]. The present study employs this methodology for the purpose of analyzing the classification of electroencephalography data, with the scope of the study being extended to encompass the processing of time series data [7]. The initial step in the process is the selection of an instance x which consists of a special but common individual EEG response. This is followed by the design of a perturbed dataset with a defined proximity to the selected instance. The perturbed times series (trials) p are represented by a binary vector z, which contains information about the changes made (see Figure 1). These vectors constitute the input into the explain-model $g(z)$, whereas the trials p are used as the input into the network-model $a_n = f(p)$.

The application of the EEG data was conducted in accordance with the master EEG data, which functioned as the instance. From these voltages, perturbed trials were generated. The perturbation could be realized through alterations in each sample of the time series. However, to circumvent the substantial computation times required in this case, clusters of samples were grouped into sections that were subsequently treated as a block. The instance time series was divided into $N_{TS} = 8$ trial sections (TS) following audiological arguments: the six waves [8] and a starting and finishing part. The intersection points are the local minima in the master, i.e. the instance time series.

As illustrated in Figure 1, the perturbation was realized through the implementation of three distinct modification methods, each of which was applied to a complete trail section TS. Basic idea was to reduce the information within the TS to a distinct extent implementing a varying proximity to the instance:

1. Cutting a wave by replacing the voltages with linear values, connecting the intersection points of the corresponding section,
2. Replacement of voltage values by a high-amplitude white noise,
3. Setting the voltage values to zero.

As illustrated in Figure 1, the proximity between the perturbed trials and the instance varies depending on the modification method. The degree of locality is a pivotal parameter, with the potential to influence the test capabilities of the procedures.

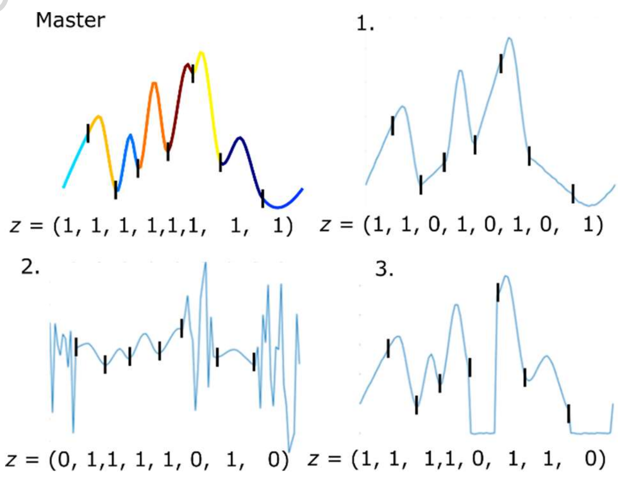

Figure 1. Master for class "loud" (upper left, the colors highlight the TS in accordance with Figure 2) and impact of modification methods on the trials (plot numbering see text), the black lines mark the TS borders, z is the representing vector giving zero if the TS is changed, otherwise one.

Every trial is represented by a binary vector z (see Figure 1) of length N_{TS} encoding the information whether the trial section is modified (zero) or not (one). These vectors are the input into the explain-model. It needs to be understandable to humans and a perceptron with a linear activation function was chosen:

$$a_{\mathrm{e}} = g(z) = \boldsymbol{W}\boldsymbol{z}' + b \qquad (1)$$

The Matrix $\boldsymbol{W}$ contains the weightings of the model consisting of a row vector of length N_{TS}. As for the network-model, the output a_{e} of the perceptron reflects the probability that one perturbed trial $\boldsymbol{z}$ belongs to a defined class. The larger the elements w_i of the matrix $\boldsymbol{W}$ the higher the probability. Since w_i relates to the i^{th} trial section it is a direct measure of the relevance of this trial section to the decision process of the network-model; the bias b is constant for all trial sections. This is the core of the interpretability of the explain-model.
During the training process of the explain-model the output values a_{n} of the network-model were used as targets and the loss function was defined as

$$L = \|a_{\mathrm{n}} - a_{\mathrm{e}}\| = \sum_{p,z} (a_{\mathrm{n}}(p) - a_{\mathrm{e}}(z))^2 \qquad (2)$$

which was minimised using an Adam-optimisation procedure. Separate training processes were made for every class (number of classes: $N_{\mathrm{c}} = 3$) and modification method.
To evaluate the efficacy of the LIME procedure two alternative tools were applied. First, saliency maps [9] were calculated and the gradient of the probability to belong to a class was calculated for every class and modification. Secondly, the performance was analysed to evaluate whether higher weights from the explain-model are really related to an improvement of the classification process. Particularly useful for this purpose are variables deduced from the confusion matrix. Since a single class is addressed in one LIME-process only accuracy A could be applied. It was calculated for the incident x and for every specific perturbed time series $\boldsymbol{z}_i$, and the loss of accuracy LA = $A(x)$ - $A(p_i)$ was used as performance measure.
To accompany this by a substantial different measure a metric distance was defined in the space opened up by the probabilities $0 \leq P_c \leq 1$, where c means all classes involved. The network-model delivers N_{c} probabilities p_c for every disturbed trial and the lower the distance to the target the better is the performance of the network:

$$MD = \sqrt{\sum_{c=1}^{N_{\mathrm{c}}} |p_c - t_c|^2} \qquad (3)$$

where t_c are the target probabilities which are equal to one for every class, respectively.
For LIME procedure 32 000 perturbed trials were generated and up to $N_{\mathrm{Comb}} = 7$ TS could be modified at once (randomly set). Thus, 254 combinations were possible. The training of the explain-model started with random values for $\boldsymbol{W}$ and 5 cycles were calculated and the w_i were averaged.

The present study sought to ascertain the correlation between weights and performance. To this end, performance metrics were calculated for time series in which one single TS was modified. This method offered the potential to provide a direct indication of the relevance of the addressed TS. Regrettably, the modification proved to be negligible, resulting in unreliable outcomes. Consequently, the performance metrics were determined for a comprehensive run encompassing all 254 combinations, with the number of trials per combination being randomly set. A comparison was made between these values and the sum of the respective involved normalized weightings, which is referred as the relevance parameter RP:

$$RP = \sum_{\wedge_i\{z_i=0\}}^{N_{\mathrm{Comb}}} \frac{w_i}{\sum_j^{N_{\mathrm{TS}}} w_j} \qquad (4)$$

3. Results

Saliency maps and the output of the LIME procedure were used to analyze the classification process. The results of a representative training run are depicted in Figure 2 for class 1 (loud).

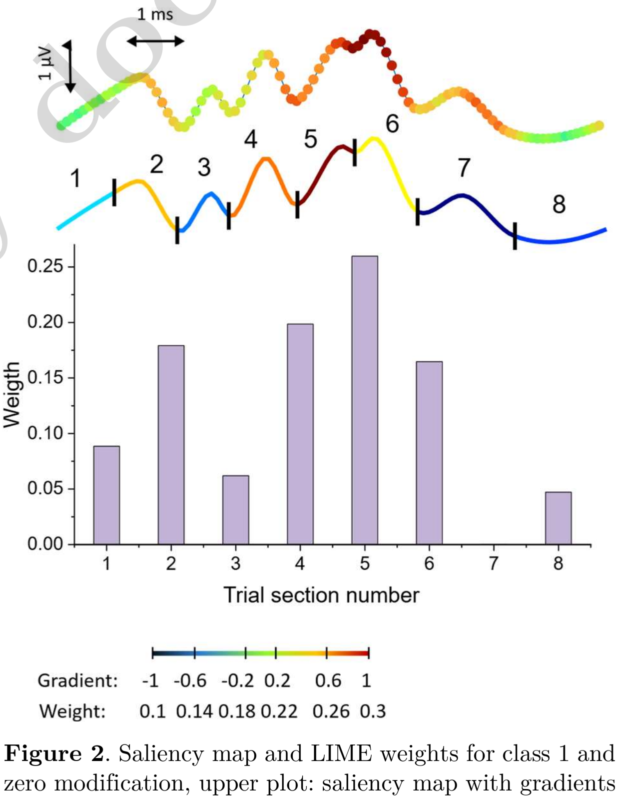

Figure 2. Saliency map and LIME weights for class 1 and zero modification, upper plot: saliency map with gradients colour-coded following the legend at the bottom of figure; medium plot: weights for zero modifications, the black perpendicular lines mark the TS borders, the TS are identified by the black numbers; lower plot: column-plot of LIME weights for every section.

The gradients in the saliency map are generally correlated with the values of the weights of the LIME procedure for the corresponding trial sections. It can be

seen that the importance of the trial section (TSxx) for the decision of the network is directly related to the gradients; that is, the greater the gradients, the greater the weights and the more significant the trial section.
This information is frequently characterised by relativity; for instance, the weight of TS2 exceeds that of TS1 (see Figure 2), and the gradients are elevated in TS2 relative to TS1 in the saliency map. It was determined that wave IV (TS5) and wave V (TS6) exhibited the highest weight values, thereby indicating their paramount significance in network decision-making. This finding aligns with the established principles of audiological expertise, as discussed in the discussion section.
As illustrated in Figure 3, the relationships between the weight values of individual TS are consistent with the gradient relations of the saliency map for class 3. Nevertheless, the gradients are observed to be of a smaller magnitude than those seen in class 1. Despite the negative nature of the data, the gradients of the saliency map correspond to high weight values for TS5 and TS6, thereby demonstrating the importance of these trial sections. It appears that pronounced negative gradients also constitute a vital component of the network decision-making process.

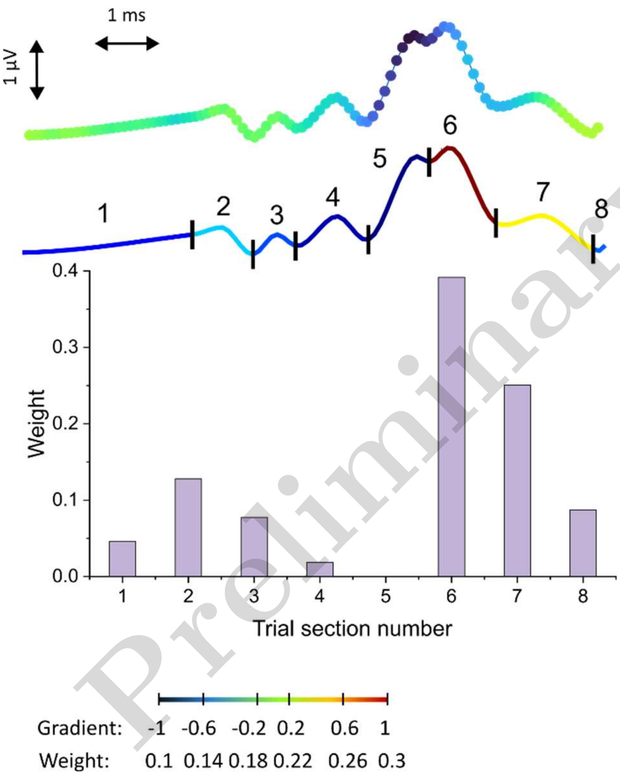

Figure 3. Saliency map and LIME weights for class 3 and zero modification, structure of figure as in Figure 2.

To analyse the quality of classification, the performance measures are depicted in dependence on the relevance parameter in Figure 4 for the zero modification. A quite clear relation was found, the higher the relevance parameter, the better the performance. The coefficient of determination R^2 varies between 0.6 and 0.7 showing satisfying correlation.
In order to assess the effectiveness of the various modification methods, the correlations between performance measures and relevance parameters were calculated for all three methods, as illustrated in Figure 4. Table 1 and 2 summarises all values, and a wide spreading was identified. It is evident that the majority of smooth and reasonable values occur for the zero modification.

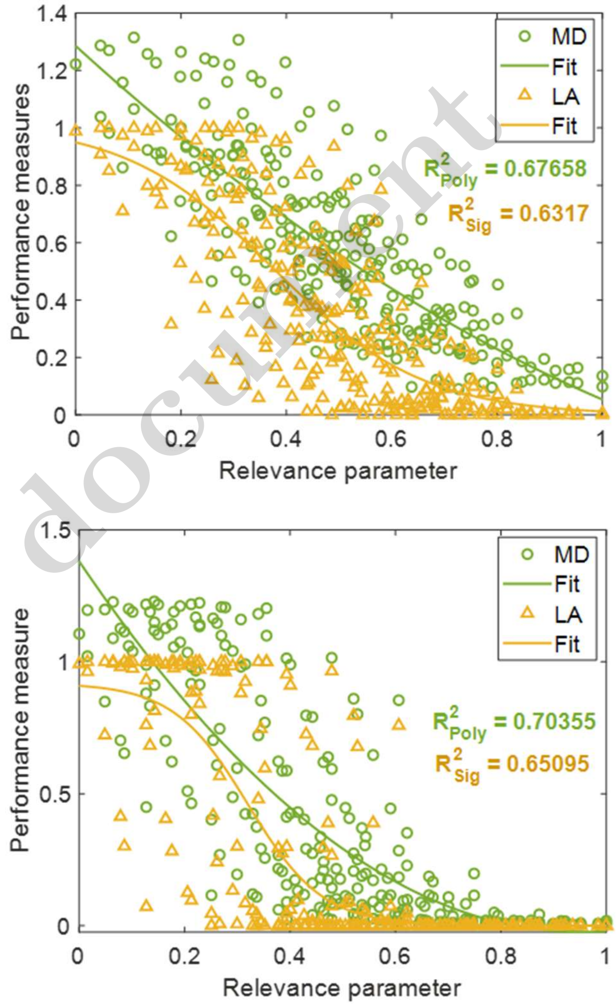

Figure 4. Plot of performance measures metric distance (MD) and loss of accuracy (LA) in dependence on the relevance parameter *RP*; upper plot: class 1, zero modification; lower plot: class 3, zero modification.

4. Discussion

A LIME method was developed for finding explanation routes for a loudness classification task realized by a neural network. A perceptron was used as an explanatory model, and weights were calculated for a dedicated splitting of the trial time series into sections. The values of the weights correspond to the information obtained from the gradients of a saliency map. It became clear that the weights with high importance for the network decision can be separated from the low-importance ones. A relevance parameter was defined by summing up all the weights involved. There is a clear correlation between the performance of the classification, as defined by a metric measure, and the accuracy.

Table 1. R^2 values for all classes and modifications for the metric distance (MD), the colours represent the numbers from blue (low values), green (medium values) to brown and finally yellow (highest values).

MD	Class 1	Class 2	Class 3
cut	0.66	0.92	0.72
noise	0.57	0.37	0.62
zero	0.68	0.80	0.70

Table 2. R^2 values for all classes and modifications for loss of accuracy (LA), the colours represent the numbers from blue (low values), green (medium values) to brown and finally yellow (highest values).

LA	Class 1	Class 2	Class 3
cut	0.40	0.42	0.12
noise	0.66	0.53	0.63
zero	0.63	0.69	0.65

It is evident that strong negative gradients in the saliency map for one class are often accompanied by strong positive gradients in the maps of other classes. For instance, this can be observed in TS5 in Figures 2 and 3. It appears that these TSs contain vital information. Despite their negative characteristics, they align with high weights in the LIME analysis, even for class three. For a low number of classes this seems to be quite natural behavior. In cases of more complex classification tasks, it can be challenging to identify such pairing gradients, and a LIME analysis is the superior explanation method.
Wave V is the most significant element in this regard, as it serves as a foundation for locating and interpreting ABR records. Picton et al. explore the implications of this TS for the estimation of hearing thresholds from ABR data [10]. Fobel et al. investigated the performance of novel chirp signals by examining the properties of wave V [11]. Elberling also defined wave V as 'most important part' during their analysis of variations of latency and amplitude of wave V for investigating the improvement of their sweep stimuli [12]. This finding aligns with the results of the present study, which indicate that TS4 and TS5 have the greatest impact on the LIME process's explanatory model. It is evident that they dominate the relevance parameter deduced, and a clear correlation can be seen with performance measures.
The values of the model weights provide valuable insights into the relevance of a TS, as well as offering a performance indicator for the classification. The proposed relevance parameter has been shown to correlate with performance measures defined and can be used to quantify the quality of the classification process as a whole. This is of particular interest in the context of optimizing the learning process or for the purpose of approving trained networks [2].
Several modification methods were applied to test the influence of trial sectioning on performance and to determine the appropriate amount of proximity during the perturbation process. The most meaningful results are obtained by setting the whole section to zero. This appears to be a reasonable approach, as the information is reduced in a very precise manner, enabling clear identification of any differences in performance. The zero-modification realizes the lowest proximity, and it may be concluded that a minimum spreading during the perturbation is necessary to ensure a reliable and effective LIME procedure.

5. Conclusions

In conclusion, it can be summarized that a LIME procedure can be applied to EEG data in a way that is beneficial to the evaluation of an AI-based classification process. A single-layer perceptron was able to characterize the relevance of particular data within a local environment to the instance. It was also possible to draw conclusions on how to generate new test data. The methods outlined in this paper offer a valuable contribution to the solution of the problem of approving the application of AI software in critical processes, such as medical applications or type approval [2].

6. Acknowledgments

Final language editing was carried out with DeepL.

References

[1] D. Schwabe, K. Becker, M. Seyferth, A. Klaß, and T. Schaeffter, "The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review," *NPJ digital medicine*, vol. 7, no. 1, p. 203, 2024, doi: 10.1038/s41746-024-01196-4.

[2] C. Koch, "Intuitive, phenomenological approach to test methods for approval purposes of trained

models using the example of predicting loudness classes from EEG-data," *Meas. Sci. Technol.*, vol. 36, no. 3, p. 36010, 2025, doi: 10.1088/1361-6501/adb5a8.

[3] U. J. Muehlematter, P. Daniore, and K. N. Vokinger, "Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis," *The Lancet. Digital health*, vol. 3, no. 3, e195-e203, 2021, doi: 10.1016/S2589-7500(20)30292-2.

[4] C. Koch, "Design of customisable audible brainstem response data for training and testing of audiological AI-software," Physikalisch-Technische Bundesanstalt (PTB), Braunschweig, PTB-OAR, 2026, doi: 10.7795/710.20260210.

[5] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," 2015. [Online]. Available: https://arxiv.org/pdf/1503.02531

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," Feb. 2016. [Online]. Available: http://arxiv.org/pdf/1602.04938

[7] I. Neves *et al.,* "Interpretable heartbeat classification using local model-agnostic explanations on ECGs," *Computers in biology and medicine*, vol. 133, p. 104393, 2021, doi: 10.1016/j.compbiomed.2021.104393.

[8] T. W. Picton, *Human Auditory Evoked Potentials*: Plural Publishing, 2011.

[9] G. Ras, N. Xie, M. van Gerven, and D. Doran, "Explainable Deep Learning: A Field Guide for the Uninitiated," *jair*, vol. 73, pp. 329–397, 2022, doi: 10.1613/jair.1.13200.

[10] T. W. Picton, D. R. Stapells, and K. B. Campbell, "Auditory evoked potentials from the human cochlea and brainstem," *The Journal of otolaryngology. Supplement*, vol. 9, pp. 1–41, 1981.

[11] O. Fobel and T. Dau, "Searching for the optimal stimulus eliciting auditory brainstem responses in humans," *The Journal of the Acoustical Society of America*, vol. 116, 4 Pt 1, pp. 2213–2222, 2004, doi: 10.1121/1.1787523.

[12] C. Elberling, J. Callø, and M. Don, "Evaluating auditory brainstem responses to different chirp stimuli at three levels of stimulation," *The Journal of the Acoustical Society of America*, vol. 128, p. 215, 2010.