Proceedings on Automation in Medical Engineering
Proc AUTOMED, 2026, Vol. 3, No 1, Article ID 2511, 2 Pages

Infinite Science Publishing

Local Explanations for Classification of Ventilation Data by Neural Networks

Tim Bogumil[1,2*], Ulrike Engeln [1,2], and Sibylle Schupp [2]

[1] *WEINMANN Emergency Medical Technology GmbH, Hamburg, Germany*
[2] *Institute for Software Systems, Hamburg University of Technology, Hamburg, Germany*
[*] *Corresponding author email: t.bogumil@weinmann-emt.de*

Abstract: Neural networks (NNs) have great potential to improve individualization of medicine, e.g., through analysis of signals. However, they are generally not interpretable. Understanding NN decisions is crucial, especially in safety-critical domains such as medicine. This work presents a new method to provide local explanations for classifications of signals. Our method extends the Sig-LIME explanation method from one-dimensional signals to multidimensional signals by introducing new perturbation techniques. We evaluate the proposed method on a NN that classifies the positive end-expiratory pressure (PEEP) applied by a ventilator. The evaluation shows that the generated explanations are plausible, stable and concise.

I. Introduction

Adaption of the therapy to the individual patient is essential in modern medicine. One possibility to achieve individualization is the application of data-driven approaches such as neural networks (NNs) to make personalized therapy decisions, for example through classification of patient data. However, NNs generally lack interpretability, which is crucial in medical applications, where lives depend on the classification result. Medical decisions are often based on multiple signals, e.g., ECG leads or ventilation curves. Thus, we require techniques that explain signal-based classifications of NNs. A possible way of explanation is to indicate which sequences of the signal are decisive for the classification result. Such explanations improve interpretability, and by that create trust and make classifying neural networks applicable in the medical domain. In this work, we present a new method to provide local explanations for classifications of multidimensional signals made by NNs. We extend Sig-LIME, an explanation technique for classification of one-dimensional signals to multidimensional input signals. For evaluation, we demonstrate the proposed method on a NN that classifies the positive end-expiratory pressure (PEEP) level based on two-dimensional ventilation curves, i.e. pressure and flow signals.

II. Explainable AI (LIME and Sig-LIME)

Local interpretable model-agnostic explanations (LIME) [1] is a popular technique to explain blackbox classifiers such as neural networks. The general idea is to not explain the overall behaviour of the classifier but rather explain a specific given decision through a surrogate model as illustrated in Fig. 1. Fig. 2 shows the workflow of LIME. First, a perturbed dataset is created by perturbing the original model input for which the decisions shall be explained, to obtain artificial input data close to the original one. The artificial input data is then labeled with the classifier that shall be explained. This corresponds to the small circles/crosses in Fig. 1. Then, a simple surrogate model is trained on the artificial dataset. It locally describes the model behaviour, in Fig. 1 through a linear classifier. An explanation is obtained by analyzing the surrogate model. For this work we use a Random Forest as a surrogate model.

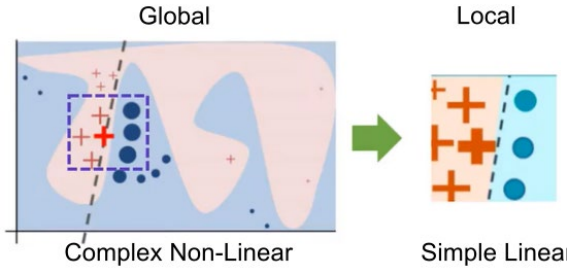

Figure 1: The classification of the bold red cross should be explained. Left: the original, complex decision function of the NN as the blue area, the simpler decision function of the surrogate model as the dashed line. Right: the surrogate model for the bold red cross with the linear decision boundary. [1]

Sig-LIME [2] is an extension of LIME to signals. It improves LIME for signal input by adapting the data perturbation step. Artificial input signals are created by dividing the original signal into overlapping segments and applying white Gaussian noise to one segment, while keeping the remaining original signal. The noise power is decided by a given signal-to-noise ratio. The trained surrogate model is a Random Forest. To generate the explanation, the feature importances of the samples from the signal are calculated as Gini importance and important features are highlighted. This strategy demonstrated to be suitable for explaining classifiers that use signals as input.

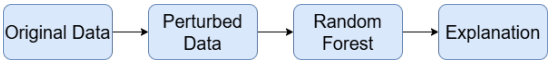

Figure 2: LIME workflow

III. Sig-LIME for multiple input signals

We enhance Sig-LIME by adapting the data perturbation step to an approach for classifiers with multidimensional signal input. We propose a synchronous noise method that applies noise to all signal dimensions in the same time segment. The applied noise in a given perturbation segment is individually calculated for each dimension based on the signal-to-noise ratio.

The NN that we analyze in this work consists of convolutional layers and gated recurrent units. It classifies the PEEP of a ventilation to either "good", "too low", or "too high". For this purpose, roughly 45000 ventilation curves are simulated by the Pulse Physiology Engine [3]. The duration of the simulations is set to one breathing cycle. The patient settings employ lower-inflection-points ranging from -20 to 10 cmH2O, upper-inflection-points ranging from 20 to 45 cmH2O and a variety of resistances (3-30 cmH20/(L/s)) and compliances (10-90 ml/cmH20) for the respiratory system. 40000 simulated curves were used for training. One breathing cycle is taken as model input. The data is labeled based on the lower- and upper inflection points of the lung compliance [4]:

- If the PEEP is below the lower inflection point, it is considered "too low".
- If the maximal ventilation pressure exceeds the upper inflection point, it is "too high".
- If the ventilation remains between both inflection points, the PEEP is considered "good".

IV. Evaluation and discussion

For evaluation of the proposed method, we evaluate stability, concision, and correctness of explanations for the NN. The perturbation dataset is generated with multiple perturbation segment sizes ranging from 4 to 240 samples and multiple signal-to-noise ratios to obtain a variety of roughly 5000 perturbed signals. We evaluate our explanation method on 9 different input signals, three from each class. One exemplary explanation is shown in Fig. 3.

Stability of an explanation signifies that explanations do not differ between multiple runs, despite the random choices in training of the surrogate model. It is crucial for meaningful, trustworthy explanations. To evaluate stability, we generate 10 explanations for each input and evaluate their correlation in terms of the resulting feature importance. The results show that explanations are stable; the average correlation coefficient is 0.83.

Concision of explanations, i.e., indicating few short segments of the signal as important, is important for helpful, understandable explanations. An explanation that marks the whole signal as important would not improve interpretability of the NN. We evaluate concision by determining the number and length of segments in the signals with feature importance above average. As visualized in the example explanation from Fig. 3, we observe that the explanations are concise.

Correctness cannot be trivially assessed on the provided example because we rely on an NN for which we do not know the underlying decision process. However, we can evaluate the plausibility of the explanations based on the context. For explanations that distinguish between "good" and "too low", we expect the beginning of the inspiration to be important and for those distinguishing between "too high" and "good", we expect the end of inspiration to be decisive. Manual evaluation shows that about 60 % of the explanations are correct according to our definition.

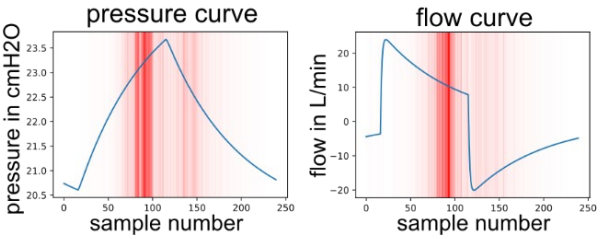

Figure 3: Explanation of a model decision. The intensity of red indicates the importance of the samples. Samples with higher intensity are more important. Only the end of the inspiration and the beginning of the expiration are marked as important, which aligns with our definition of correctness for this class.

In general, the results show that through synchronous perturbation of all signal dimensions, stable and concise explanations for the behavior of an NN classifier can be generated. A threat to validity of the evaluation is the evaluated NN, since the underlying decision process is not known. Future work should focus on more sophisticated evaluation of the correctness and concision of explanations by explaining models with known behaviour. Further, because of the synchronous application of noise in all dimensions of the signal, explanations focus on important timeframes instead of individual features. A possible continuation of this work is the evaluation of further perturbation methods, e.g., individual application of noise in each dimension.

V. Conclusions

In this work, we enhanced Sig-LIME for explanations of multidimensional signal classifications through synchronous application of noise in all dimensions. The evaluation shows that stable and concise explanations for multidimensional signal classification of NNs can be achieved, which is an important step toward their application in the medical domain.

Author's Statement

This work was funded by Weinmann Emergency Medical Technology GmbH and supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project no. 513623283 as part of the Research Training Group CAUSE.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier* KDD 2016, pages 1135–1144, New York, USA.

[2] T. A. A. Abdullah et al.. *Sig-lime: A signal-based enhancement of lime explanation technique* IEEE Access, 12:52641–52658, 2024.

[3] A. Bray et al.. *Pulse Physiology Engine: an Open-Source Software Platform for Computational Modeling of Human Medical Simulation.* SN Comprehensive Clinical Medicine. 2019.

[4] W. Oczenski, H. Andel, and A. Werba. *Atmen- Atemhilfen: Atemphysiologie und Beatmungstechnik*. Georg Thieme Verlag KG, Stuttgart, 11. Auflage, 2023