Explainable Deep Learning for Medical Time Series Data

Thomas Frick[1,3(✉)], Stefan Glüge[2], Abbas Rahimi[3], Luca Benini[3], and Thomas Brunschwiler[1]

[1] IBM Research Zurich, Smart System Integration, Zurich, Switzerland
{fri,tbr}@zurich.ibm.com
[2] Zurich University of Applied Sciences, Institute of Applied Simulation, Zurich, Switzerland
stefan.gluege@zhaw.ch
[3] ETH Zurich, Integrated Systems Laboratory, Zurich, Switzerland
abbas@iis.ee.ethz.ch, lbenini@ethz.ch

Abstract. Neural Networks are powerful classifiers. However, they are black boxes and do not provide explicit explanations for their decisions. For many applications, particularly in health care, explanations are essential for building trust in the model. In the field of computer vision, a multitude of explainability methods have been developed to analyze Neural Networks by explaining what they have learned during training and what factors influence their decisions. This work provides an overview of these explanation methods in form of a taxonomy. We adapt and benchmark the different methods to time series data. Further, we introduce quantitative explanation metrics that enable us to build an objective benchmarking framework with which we extensively rate and compare explainability methods. As a result, we show that the Grad-CAM++ algorithm outperforms all other methods. Finally, we identify the limits of existing explanation methods for specific datasets, with feature values close to zero.

Keywords: Explainable deep learning · Convolutional Neural Network · Explanation quality metric · Medical time series data

1 Introduction

Neural Networks have become the state-of-the-art to model complex problems in computer vision [1], speech recognition [2], and many other areas [3]. Given enough data, they can be trained to be exceptionally accurate classifiers, sometimes even surpassing human performance. While some classical machine learning models such as Decision Trees are inherently interpretable, Neural Networks are unfortunately black boxes when it comes to understanding why a classification decision was made. However, for many applications, e.g. in autonomous driving or in healthcare, it is of uttermost importance that the decisions of

J. Ye et al. (Eds.): MobiHealth 2020, LNICST 362, pp. 244–256, 2021.
https://doi.org/10.1007/978-3-030-70569-5_15

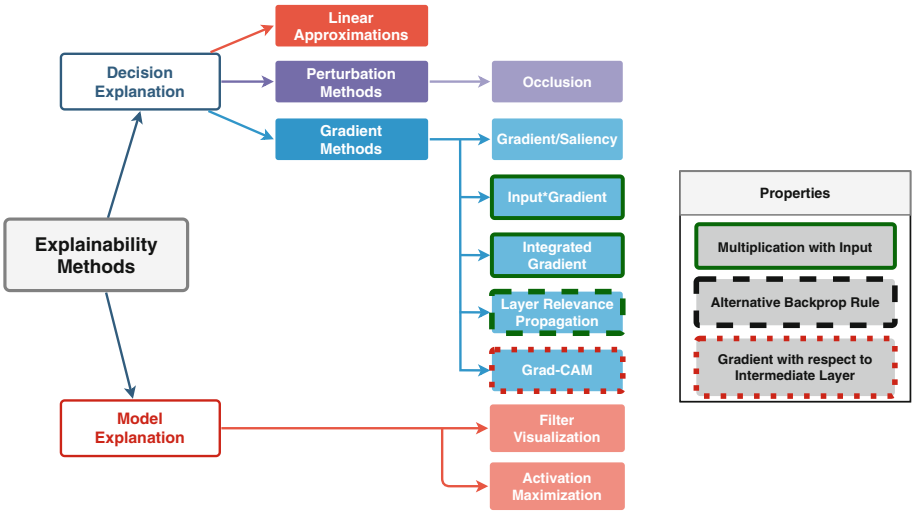

Fig. 1. Taxonomy of a few explanation methods for Neural Networks

machine learning models can be explained and therefore trusted. Decisions may have severe consequences, especially in healthcare. Therefore, medical experts need to be able to trust their models to make the right decisions for the right reasons. Additionally, Neural Network explanations could be used to identify new patterns in convoluted signals such as electroencephalography (EEG) and therefore, advance scientific knowledge.

In the last few years, there has been an effort to develop different Neural Network explanation methods [1,3–11]. Most of these methods have been proposed for computer vision tasks. They highlight the areas of the model input that are most relevant to the decision process. These so-called attributions or heat maps associate a relevance score with each individual input feature. Some algorithms generate explanations in more complex feature spaces such as the frequency domain. Ultimately, one could aim for explanations in a given expert language - meaning that the explanation from a more complex feature space is translated into terms that are familiar to the user (e.g. text book rules in medicine).

In this study, our goal is to adapt and benchmark Neural Network explainability methods for medical time series data (Sect. 2). So far, researchers have mainly compared explanation methods on image data, using a qualitative assessments of the resulting explanations by humans. There have been some theoretical attempts at defining characteristics of good explanations [12,13]. In contrast, we introduce and apply quantitative explanation metrics (Sect. 3). This allows us to rate and compare the different methods objectively (Sect. 4). Additionally, the metric demonstrates the strength and limitations of the various approaches.

2 Taxonomy of Explanation Methods for Neural Networks

Machine Learning models can be divided into inherently interpretable models (e.g. Decision Trees) and black box models (e.g. Neural Networks) that use formulations too complex to be interpreted by humans. In an attempt to offer explanations for decisions made by such models, a multitude of methods have been proposed to analyze the model after training (*a posteriori* explanations) [1, 3,4,6–11,14,15]. These algorithms visualize what a Neural Network has learned and how classification decisions are made. Most of these methods have been developed in the context of computer vision (image data). However, the lack of quantitative evaluation metrics does not allow an objective benchmarking.

The existing *a posteriori* explainability methods for Neural Networks can be classified into multiple categories according to their underlying algorithms and properties (see Fig. 1). At high level, they can be split into two types of explanations:

- **Model Explanations** are visualizations of the patterns and concepts that the network has learned during training.
- **Decision Explanations** highlight the most relevant parts of a given model input that led to a specific classification decision.

2.1 Model Explanations

Two kinds of model explanations are most relevant: *Filter Visualization* [7] and *Activation Maximization* [8]. The former depicts the kernel weights of the first few layers of Convolutional Neural Networks (CNN), indicating the patterns each filter is most susceptible to. The latter is based on the idea that each neuron is looking for a particular pattern in the input. If this pattern is present, the activation of the neuron is high, otherwise the activation is low or zero. This is achieved by optimizing the input to the network with the objective of maximizing the activation of a particular neuron. In the case of CNNs, there is the choice between optimizing the activation of a neuron, a layer, a channel, a class logit, or a class probability.

2.2 Decision Explanations

Decision explanations explain a specific classification decision, given a particular input sample, by attributing relevance scores to single pixels (for images) or time steps (for time series). The higher the relevance, the more impact a pixel or time step has on the classification decision. It is important to note that the terms saliency map, attribution map, heat map and relevance scores are used interchangeably to describe the same concept.

Decision explanations can be further divided into methods that linearly approximate the model, methods that use the internal gradient flow within the network, and perturbation based methods that alter the input while observing the change in output probabilities to calculate attribution:

1. **Linear approximation methods** (e.g. LIME [9]) construct a linear proxy model that serves as an approximation of a black box model by probing the output behavior around a given input sample.
2. **Perturbation-based methods** (e.g. Occlusion [16]) calculate attributions by removing or altering the input while observing how the classification probabilities change. The higher the change, the more relevant is the part of the input that has been altered.
3. **Gradient-based methods** (e.g. Saliency Maps [14], Gradient*Input [11] [6], Integrated Gradients [17], Epsilon Layer Relevance Propagation (Epsilon-LRP) [18], Grad-CAM [10], Grad-CAM++ [19]) make use of the partial derivative of the logits of the output class with respect to the input or with respect to the output of an intermediate layer as a measure of sensitivity, and thereby as a measure of attribution. Because gradient-based methods are computed with a single forward and backward pass, they are typically faster than occlusion or linear approximation methods. Additionally, Gradient-based methods can be classified using the following three characteristics as show in Fig. 1:
   - **Backpropagation:** Some methods (Epsilon-LRP) change the distribution of the gradients during backpropagation, to improve certain properties of the attribution map.
   - **Gradient:** Some methods either use the gradient with respect to the input (Gradient, Gradient*Input, Integrated Gradient, Epsilon-LRP) or the gradient with respect to an intermediate layer (Grad-CAM, Gad-CAM++).
   - **Multiplication with the Input:** Some methods (Gradient*Input, Integrated Gradient, Epsilon-LRP) use a multiplication with the input in their calculation of attribution.

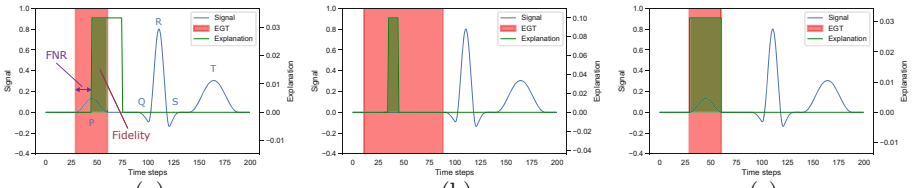

Fig. 2. Synthetic heartbeats (ECG) depicted as P, Q, R, S, T-complex with corresponding Explanation Ground Truth (EGT) and example explanations resulting in: (a) a Fidelity Score of 0.5 and a False Negative Rate of 0.5 for a normal sinus rhythm, (b) a Fidelity Score of 1.0 and a False Negative Rate of 0.8 for atrial fibrillation, (c) a Fidelity Score of 1.0 and a False Negative Rate of 0.0 for a normal sinus rhythm.

3 Explanation Ground Truth and Quality Metrics

To the best of our knowledge, no method exists to rate different explanation methods quantitatively with respect to a ground truth explanation. Most publications make use of qualitative comparisons. Alvarez-Melis et al. [20] proposed a quantitative metric to assess how *faithful* the generated explanation is with respect to the model. The basic idea is to observe the change in the model's predictions while removing pixels or time steps and correlating the attribution score of the explanation method with that change. This faithfulness metric is not a measure of how well the attributions correlate with a ground truth explanation of the input, but only indicates how *faithfully* the explanation represents what the model bases its decision on.

In order to asses the quality of generated explanations, we have to annotate the data with a "true" explanation, also called the **Explanation Ground Truth** (EGT). We can then assess individual explanations relative to this ground truth. It contains all pixels or time steps that provide distinctive information for the relevant class. The EGT does not contain features which are not distinct for the specific class and could appear in other classes as well. Figure 2 depicts an example of the EGT for a synthetic heartbeat as recorded by a ECG.

The P-Wave (first peak of the heartbeat waveform) is one of the distinctive factors to distinguish a sinus rhythm from atrial fibrillation. The patient is suffering from atrial fibrillation if the P-Wave is missing. Therefore, for the sinus rhythm class, we place the EGT on all time steps containing the P-Wave (see Fig. 2a). For the atrial fibrillation class, we define the EGT on all time steps where the P-Wave could be located if it would be present (see Fig. 2b).

We constructed our own synthetic time series datasets with distinct features, indicative for a specific class, to benchmark the explanation methods on controlled cases. Furthermore, this approach allowed us to identify strengths and limitations of the methods for time series data with different characteristics.

In order to benchmark an explanation, we defined **Scores** that allow to perform a quantitative comparison. First, the generated explanation $A = \{A_0, ..., A_N\}$ needs to be normalized to a total area of 1, for N time steps.

$$\bar{A}_i = \frac{A_i}{\sum_i A_i} \tag{1}$$

We then propose two new metrics, which in combination characterize the generated explanation under investigation: the Fidelity Score and the False Negative Rate.

– **Fidelity Score:** This metric measures how much of the explanation appears inside the actual "true" explanation. We define it as the total sum of all attribution values (area under the curve) inside the EGT. Due to the normalized total area, this corresponds to the precision metric in classification.

$$S_{Fidelity} = \sum_{i \in \mathcal{Z}} \bar{A}_i \qquad \text{where } i \in \mathcal{Z} \text{ if } EGT_i = 1 \tag{2}$$

– **False Negative Rate (FNR):** This metric measures the completeness vs. narrowness of an explanation. We define it as the number of non-relevant pixels or time steps inside the EGT divided by the total number of pixels or time steps in the span of the EGT. Since the explanation values are almost never exactly zero, we use a threshold ϵ below which we consider the pixel or time step to have no relevance. This threshold is chosen to be equal to the explanation values of a perfectly distributed explanation (given a signal of length N): $\epsilon = 1/N$.

$$S_{\text{FNR}} = \frac{\sum_{i \in \mathcal{Z}} 1_{\{\bar{A}_i \leq \epsilon\}}}{\sum_i \text{EGT}_i} \qquad \text{where } i \in \mathcal{Z} \text{ if } \text{EGT}_i = 1 \tag{3}$$

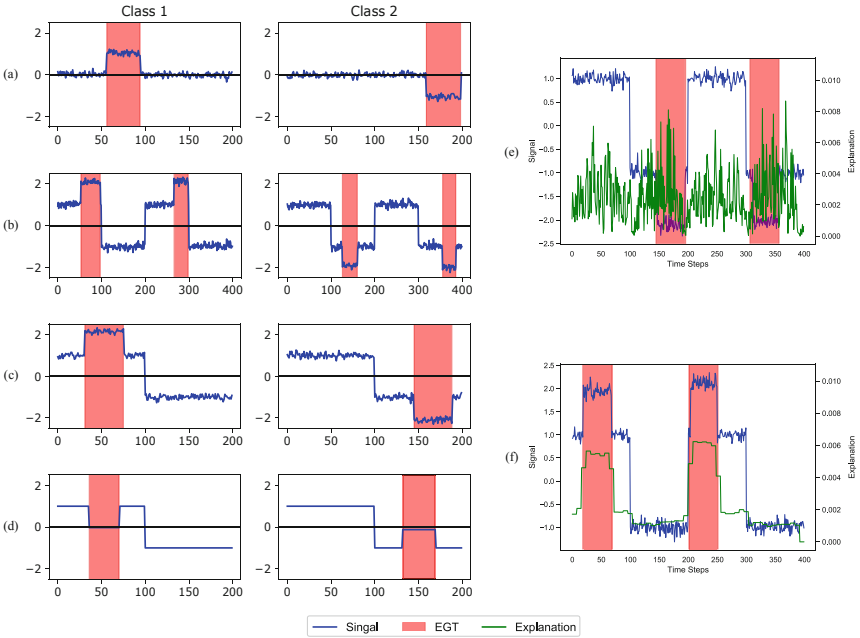

Fig. 3. Synthetic datasets (blue) generated for the benchmarking of the explanation methods: (a) Dataset Noise, (b) Background Dataset Noise, (c) Background Wide Dataset Noise, (d) Background Dataset Inverse; Example Explanations (green) for the Background Dataset Noise as resulted from: (e) Epsilon-LRP, (f) Grad-CAM++. EGT's are depicted in red for all cases. (Color figure online)

4 Experiments and Results

To evaluate our explanation score and the explanation methods, four synthetic datasets were designed and a Neural Network architecture was defined. Detailed results of the experiments are reported and discussed in this chapter. Each chosen explanation method was tested and analyzed. Finally, we show the limitations of existing explanation methods when applied to a carefully constructed dataset.

4.1 Datasets

The four synthetic datasets consist of uni-variate time series signals, with the following characteristics:

1. All generated datasets only contain clearly separable classes (1000 samples per class). There is no ambiguity about the class membership.
2. Each class must be characterizable by features present in a specific region of the signal. We call these regions Explanation Ground Truth (EGT) and use them to qualitatively measure how close an attribution map is to the "true" explanation.

The following four synthetic datasets were used for the experiments (see Fig. 3):

1. **Dataset Noise:** This dataset places a square wave signal within a class specific region. Class one is represented by a negative square wave, while class two is represented by a positive one. Additionally, white noise was added to increase the classification problem complexity.
2. **Background Dataset Noise:** This dataset augments the Dataset Noise by adding a constant background signal which forces the network to not only recognize a non zero region, but also to identify the class specific signal on top of another constant signal.
3. **Background Wide Dataset Noise:** This dataset considers two periods of the Background Dataset Noise. In contrast to the Background Dataset, there are two locations which simultaneously contain the class identifying signal, testing the network's ability to base its decision on multiple relevant regions.
4. **Background Dataset Inverse:** This dataset inverts the square wave such that, in the relevant region, the signal goes from the background value to zero. The explanation method is thus forced to cope with a zero signal where the explanation value should be high. Thus, no noise was added.

4.2 Models

For our experiments, we use a LeNet [4] and VGG [15] inspired Convolutional Neural Network (CNN): two convolutional layers and a MaxPooling layer make up a block that is repeated three times. This is followed by a single fully connected output layer.

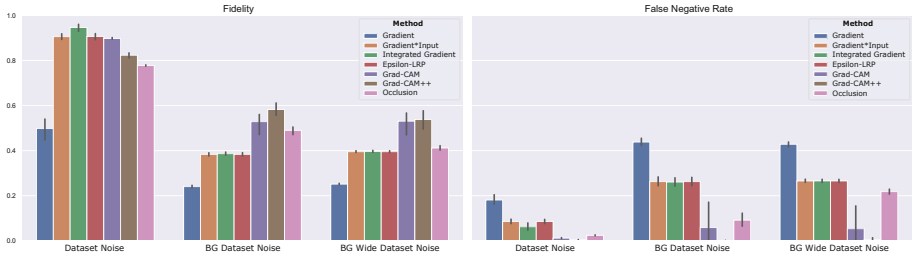

Fig. 4. Performance experiment results for three different datasets showing Grad-CAM++ outperforming all other methods.

4.3 Explanation Method Experiments

In our experiments, we investigated how well each method performs on a network that has been trained to perfect test accuracy. We then focused on the investigation of the convergence of explanation methods during training of the neural network.

Comparison of Converged Explanation Methods: In these experiments, we investigated the quality of the generated explanations using the proposed metrics. For every explanation we calculated the Fidelity Score and the FNR. As shown in Fig. 4, we observed that we can split the methods into four groups:

1. **Gradient** shows the lowest Fidelity Score and the highest FNR. This method clearly performs worse than any of the other explanation algorithms.
2. **Gradient*Input, Integrated Gradients and Epsilon-LRP** form a group of similarly performing methods. We attribute this to the shared property of these three methods: multiplication with the input resulting in attribution maps that are largely dominated by the input signal. Additionally, Ancona et al. [21] show that Gradient*Input and Epsilon-LRP are equivalent if the model under investigation exclusively uses ReLU activation functions. We therefore chose Epsilon-LRP as a representative for this group of the remaining experiments ($\epsilon = 10^{-7}$).
3. **Grad-CAM and Grad-CAM++** build a third group, outperforming all other methods in terms of Fidelity-Score and FNR. The improved Grad-CAM++ method slightly outperformed the original Grad-CAM algorithm. We used Grad-CAM++ as a representative of this group for the remaining experiments.
4. **Occlusion** performed slightly better than the second group but still worse than the third.

We note that Grad-CAM, Grad-CAM++ and Occlusion perform well, but show some artefacts, which result from the up-sampling of the embedding's attribution map, which in general is not aligned with changes in the signal. Therefore, we observe an attribution value that is an average of the attributions of either

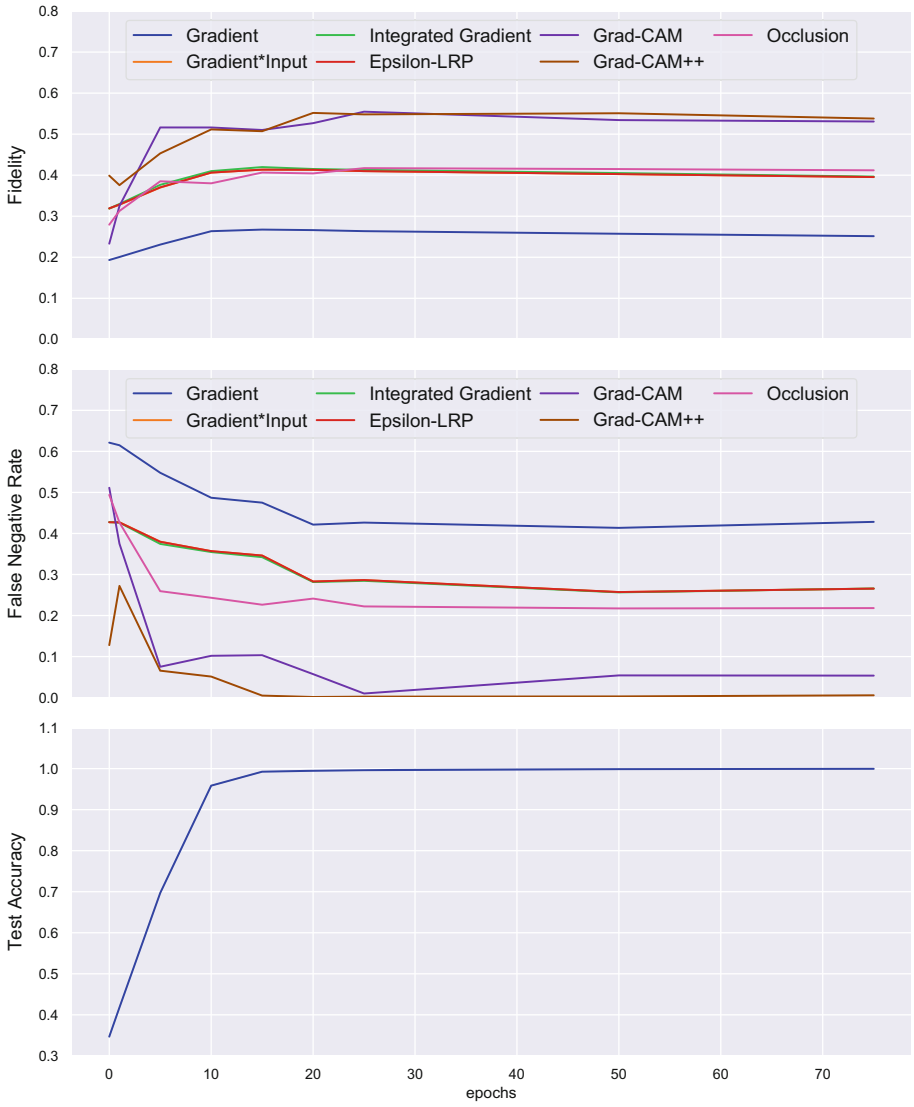

Fig. 5. Behavior of explanation methods during the training process of the Network. All methods converge to a steady state during training.

side of the input signal change. Figure 3e and 3f show two samples of generated explanations. We observe that methods which generate explanations using a partial derivative with respect to the input produce more noisy explanations as opposed to the Grad-CAM++ method. In summary, Grad-CAM++ outperformed all other methods with regard to Fidelity Score and FNR.

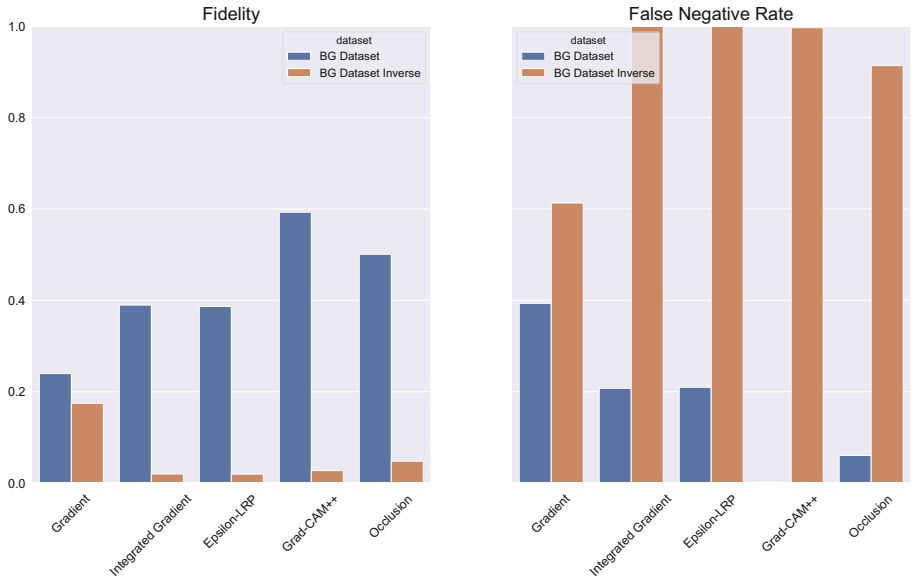

Fig. 6. Performance collapse of existing explanation methods on the inverse dataset due to an explicit or implicit multiplication with the input which is close to zero within the EGT.

Comparison of Explanation Methods During Training. In the next step, we investigated the evolution of explanations and their convergence during training of the Neural Network. We evaluated the performance of the explanation methods at different stages during training.

Figure 5 shows the change of the Fidelity Score and FNR during training. All methods improved the Fidelity Score and the FNR while the network was learning. Once the network converges to a state where the accuracy does not improve anymore (approximately after 20 epochs), the metrics converge to a steady state.

4.4 Limitations of Explanation Methods

We observed that the nature of the dataset influences the performance of the explanation methods. This is especially pronounced for methods that make use of a multiplication with the input signal. We push this to an extreme by constructing a dataset (Background Dataset Inverse 3(d)) for which these methods completely fail: The samples are constructed in such a way that the input signal is close to zero for time steps which belong to the EGT.

Figure 6 depicts the performance drop of methods that contain a multiplication with the input (i.e. Gradient*Input, Integrated Gradients, Epsilon-LRP) for the "Background Dataset Inverse" (Fig. 3(d)). The occlusion method also failed, since it replaces the original values of the input with zeros - replacing zeros with

zeros does not perturb the input, thus no change in output probabilities can be measured. Surprisingly, Grad-CAM also failed, even though it had been superior to the other methods. We explain this as follows: Grad-CAM incorporates a multiplication with the embedding to calculate the attribution for each spatial location. If all of the inputs in the receptive field of a convolutional layer are zero, the output map of that layer will also be zero. This propagates through the network to the embedding. Therefore, Grad-CAM also indirectly contains a multiplication with zero, which makes the method fail similarly to the other algorithms.

5 Conclusion

We introduced two quantitative metrics to benchmark a Neural Network explanation quality: the Fidelity Score and the False Negative Rate. The Fidelity Score measures the overlap of the generated explanation with the Explanation Ground Truth while the False Negative Rate measures the number of time steps of the Explanation Ground Truth not covered by the generated explanation.

Using these two metrics in combination with our specifically crafted synthetic datasets, we investigated the performance of various explanation methods and concluded that the Grad-CAM++ algorithm outperforms all other methods (Saliency, Gradient*Input, Integrated Gradient, Layer Relevance Propagation and Occlusion).

Additionally, we demonstrated that existing explanation methods suffer from a performance collapse for input data with values close to zero within the Explanation Ground Truth.

5.1 Future Work

In Future work, the benchmarked explainability methods should be applied to actual ECG data. Additionally, the explanation methods described in this work produce explanations that indicate which locations of the input are responsible for the network's decision. However, location is not the only factor that influences a classification decision. There can also be frequency factors - the structure of an object - that are essential for discerning two classes. Existing explanation methods are not capable of communicating a reliance on frequency components (structural features). To fully explain a network's decision, an explanation algorithm should be developed that can visualize dependencies on a combination of location and frequency.

Acknowledgment. We would like to thank Anirban Das from the Rensselaer Polytechnic Institute and Adam Invankay from IBM Research Zurich for their valuable discussions and feedback on our work.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional Neural Networks. Technical report. http://code.google.com/p/cuda-convnet/

2. Chiu, C.C., et al.: State-of-the-art speech recognition with sequence-to-sequence models, December 2017. http://arxiv.org/abs/1712.01769
3. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning, May 2015. https://doi.org/10.1038/nature14539
4. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
5. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining Explanations: An Overview of Interpretability of Machine Learning. Technical report (2019)
6. Kindermans, P.J., Schütt, K., Müller, K.R., Dähne, S.: Investigating the influence of noise and distractors on the interpretation of neural networks, November 2016. http://arxiv.org/abs/1611.07270
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012). http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
8. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill **2**(11), e7 (2017). https://doi.org/10.23915/distill.00007. https://distill.pub/2017/feature-visualization
9. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": explaining the predictions of any classifier. CoRR abs/1602.04938 (2016). http://arxiv.org/abs/1602.04938
10. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from Deep Networks via gradient-based localization, October 2016. http://arxiv.org/abs/1610.02391
11. Shrikumar, A., Greenside, P., Kundaje, A.: Learning Important Features Through Propagating Activation Differences. Technical report. http://goo.gl/qKb7pL
12. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. CoRR abs/1812.04608 (2018). http://arxiv.org/abs/1812.04608
13. Holzinger, A., Carrington, A.M., Müller, H.: Measuring the quality of explanations: the system causability scale (SCS). comparing human and machine explanations. CoRR abs/1912.09024 (2019). http://arxiv.org/abs/1912.09024
14. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps, December 2013. http://arxiv.org/abs/1312.6034
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
16. Zeiler, M.D., Fergus, R.: Visualizing and Understanding Convolutional Networks. Technical report
17. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic Attribution for Deep Networks. Technical report (2017)
18. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015). https://doi.org/10.1371/journal.pone.0130140
19. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. CoRR abs/1710.11063 (2017). http://arxiv.org/abs/1710.11063

20. Alvarez-Melis, D., Jaakkola, T.S.: Towards robust interpretability with self-explaining Neural Networks, June 2018. http://arxiv.org/abs/1806.07538
21. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for Deep Neural Networks, November 2017. http://arxiv.org/abs/1711.06104