Towards a Rigorous Evaluation of XAI Methods on Time Series

Udo Schlegel
University of Konstanz
Konstanz, Germany
u.schlegel@uni-konstanz.de

Hiba Arnout
Siemens CT & TU Munich
Munich, Germany
hiba.arnout@siemens.com

Mennatallah El-Assady
University of Konstanz
Konstanz, Germany
menna.el-assady@uni-konstanz.de

Daniela Oelke
Siemens CT
Munich, Germany
daniela.oelke@siemens.com

Daniel A. Keim
University of Konstanz
Konstanz, Germany
keim@uni-konstanz.de

Abstract

Explainable Artificial Intelligence (XAI) methods are typically deployed to explain and debug black-box machine learning models. However, most proposed XAI methods are black-boxes themselves and designed for images. Thus, they rely on visual interpretability to evaluate and prove explanations. In this work, we apply XAI methods previously used in the image and text-domain on time series. We present a methodology to test and evaluate various XAI methods on time series by introducing new verification techniques to incorporate the temporal dimension. We further conduct preliminary experiments to assess the quality of selected XAI method explanations with various verification methods on a range of datasets and inspecting quality metrics on it. We demonstrate that in our initial experiments, SHAP works robust for all models, but others like DeepLIFT, LRP, and Saliency Maps work better with specific architectures.

1. Introduction

Due to state-of-the-art performance of Deep Learning (DL) in many domains ranging from autonomous driving [14] to speech assistance [4] and the developing democratization of it, interpretability and explainability of such complex models captured more and more interest. Agencies such as DARPA introduced the explainable AI (XAI) initiative [11] to promote the research around interpretable Machine Learning (ML) to foster trust into models. Laws like the EU General Data Protection Regulation [7] got ratified to force companies to be able to explain the decisions of algorithms to support fairness and privacy and mitigate trust issues of users and costumers. The desiderata of ML systems (fairness, privacy, reliability, trust building [6]) led to a new selection process for models [21]. Depending on the task either interpretable models, such as decision trees [12], or new XAI methods, e.g., local interpretable model-agnostic explanations (LIME) [20], on top of trained complex models, for instance, Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN), are incorporated to guarantee the interpretability demands [10]. Due to these new methods for interpretability on a level above a model, we introduce a few new definitions. In the following, we refer, e.g., LIME as XAI method. Explainers are defined as an XAI method used on top of a model to get an XAI explanation of the decision making.

Many prominent XAI methods are tailored onto certain input types such as images, e.g., Saliency Maps [25], or text, e.g., layer-wise relevance propagation (LRP) [2]. They often benefit of their domain to explain with certain aspects, such as a heatmap on the input [22], as they can be used as an overlay by building an abstract feature importance [10]. However, for instance, videos (sequences of images) and audio have another temporal dimension which is currently omitted by XAI methods. Only limited consideration is taken into account for sequence or temporal data, e.g., on XAI method evaluation on natural language processing [1]. There is currently only limited work about XAI on time series data such as interpretable decision tress [12], calculating prototypes [8] and using attention mechanisms [13]. Dividing the hard task of video classifier explanation into time series and image tasks is not possible as there is no good time series solution. However, due to sensors getting cheaper and cheaper, more time-oriented data besides video and audio is generated, and thus, it is important first to test already prominent XAI methods and discover new ones. Analyzing time series further enables to automate more actions, e.g., heartbeat anomaly detection [3], solve new tasks, e.g., predictive maintenance [17], and predict stock, e.g., stock market forecasting [15].

2473-9944/19/$31.00 ©2019 IEEE
DOI 10.1109/ICCVW.2019.00516

To debug and optimize time series prediction models in diverse tasks, not only understanding is essential but also that the XAI explanation is correct itself [18]. Evaluating and verifying these explanations is a difficult task due to raw time series being large and hardly interpretable even by domain experts themselves, and so an evaluation by raw data and explanation inspection is not feasible. Due to this lack of connectable domain knowledge, a quantifiable approach is necessary to verify explanations. Notably, in computer vision exists some work about the evaluation of explanations [23] (e.g., set relevant pixels to zero [26]), which is also possible to use on time series. However, these methods omit temporal dependencies by assuming feature independence or only local (short-term) dependency and thus are only limited verifiable on time-oriented data. Hence, adapted or novel variants of previous methods are needed to evaluate explanations on time series.

In this work, we show the practical use of various XAI methods on time series and present the first evaluation of selected methods on a variety of real-world benchmark datasets. Further, we introduce two sequence verification methods and a methodology to evaluate and check XAI explanations on time series automatically. In preliminary experiments, we show the results of our verification techniques for the selected XAI techniques and their results.

2. Time Series Explanations

XAI methods have their main application field in computer vision due to the state-of-the-art success of black-box DL models in object recognition and detection [19] and the visual interpretability of the input [18]. However, a need for explainability is desired in other domains to either understand the decision making or to improve the models' performance by debugging failures. Thus, the domain of time series prediction has a high demand for XAI methods.

A classification dataset with univariate time series data D consists of n samples with classes $c_1, c_2, c_3, ..., c_k$ from a label (multiple classes) with k different classes. A sample t of D consists of m time points $t = (t_0, t_1, t_2, ..., t_m)$. E.g., an anomaly detection dataset has only two classes (anomaly, e.g., c_2, and normal, e.g., c_1). In the following, the generally considered explanation of most XAI methods is the local feature importance. Time points get converted to features to introduce a workaround to use XAI methods on time series. The local feature importance produces a relevance r_i for each time point t_i. Afterward a tuple (t_i, r_i) can be build or more general for the time series vector $t = (t_0, t_1, t_2, ..., t_m)$ a relevance vector can be generated as $r = (r_0, r_1, r_2, ..., r_m)$.

A model m trained on a subset X from D with labels Y can be formalized to $m(x) = y$ with $x \in X$ and $y \in Y$. The model m learns based on the provided data X, Y to predict an unseen subset X_{new}. In the case of time series, x is a sample like $t = (t_0, t_1, t_2, ..., t_m)$ with m time points. If then an XAI method xai is incorporated to explain the decisions of such a model, another layer on top of it is created. An explanation can then be formalized as $xai(x, m) = exp$ with exp being the resulting explanation. With time series, the explanation exp is a relevance $r = (r_0, r_1, r_2, ..., r_m)$ vector for m time points.

Similar to the saliency masks on images, a heatmap can be created based on the relevance produced by XAI methods. It is possible to create a visualization with this heatmap enriching a line plot of the original time series. Together with domain knowledge, an expert can inspect the produced explanation visualizations to verify the result qualitatively. Figure 1. shows an example of relevance heatmaps on time series. However, as these heatmaps are hard to interpret and a significant challenge to scale to large datasets or long time series, automated verification needs to be applied.

3. Evaluating Time Series Explanations

There are various options on how to evaluate and verify XAI explanations automatically. In computer vision, a common method consists of a perturbation analysis [26]. This analysis method substitutes a few pixels (e.g., exchange to zero) of an image according to their importance (most or least relevant pixels). However, because, e.g., a zero could be an indicator for an anomaly in a time series task, the methodology of evaluation of XAI methods for time series needs specialized heuristics. We present two novel methods suited explicitly for time series by taking the sequence property of the time-oriented data into account.

3.1. Perturbation on time series

At first, a perturbation analysis presents preliminary comparison baselines. The evaluation is based on the assumption that if relevant features (time points) get changed, the performance of an accurate model should decrease massively. If random time points of the data get changed, the performance should either stagnate or decrease.

Perturbation Analysis – The assumption follows the time series $t = (t_0, t_1, t_2, ..., t_m)$ and the relevance produced by the XAI method as $r = (r_0, r_1, r_2, ..., r_m)$ to get a worse result of the quality metric qm for the classifier if combined. A time point t_i gets gets changed if r_i is larger than a certain threshold e, e.g. the 90th percentile of r. Due to XAI methods have problems with some time-series samples, the threshold leads to only changing a small number of time points. In the case of time series, the time point t_i is set to zero or the inverse $(max_{t_i} - t_i)$ and leads to the new time series samples t^{zero} and $t^{inverse}$.

Perturbation Verification – To verify the assumption, a random relevance $r_r = (r_0, r_1, r_2, ..., r_m)$ is used for the same procedure. The number of changed time points, amount of r_i larger than the threshold e, is the same as in

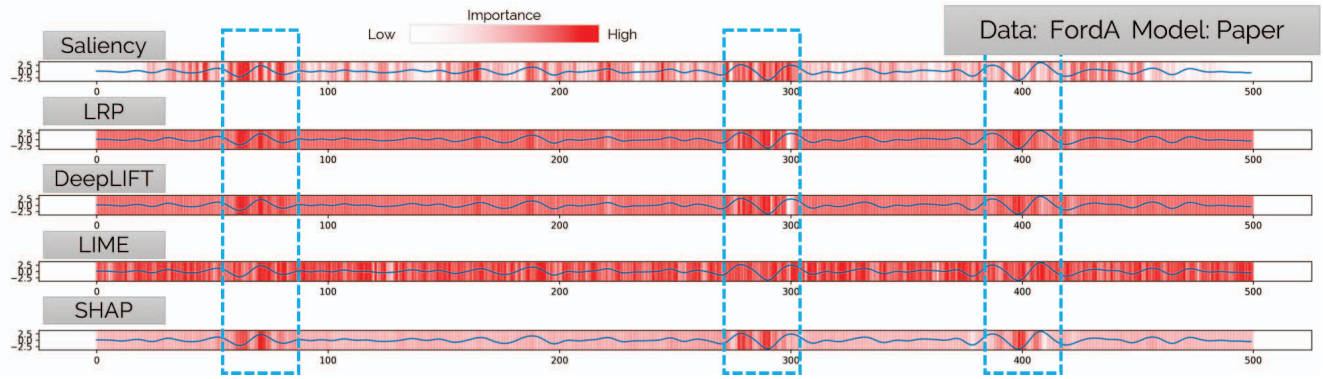

Figure 1. Relevance Heatmaps on an exemplary time series of the FordA dataset using a ResNet paper model. XAI methods shown with their relevance heatmaps are Saliency Maps, LRP, DeepLIFT, LIME, and SHAP. The blue rectangles display controversial parts of the time series for the XAI methods with red marking high importance for the classification which are, e.g., set to zero by verification methods.

the case before to set the same prerequisites for the classifier. This technique creates new time series like the perturbation analysis such as t_r^{zero} and $t_r^{inverse}$. The assumption to verify the model and the XAI method with the random relevance method follows the schema that the quality metric qm shows e.g. $qm(t) \geq qm(t_r^{zero}) > qm(t^{zero})$ for a model that maximizes qm.

3.2. Sequence Evaluation

To verify that the model and the XAI method also includes time series features such as slopes or minima, we present two novel sequence-dependent methods. If the assumptions of the perturbation analysis hold, there is still a lack of evaluation of trends or patterns in the time series. E.g., for the classification, a decrease to zero could be significant, but the perturbation sets the zero to the max as it is essential for the model and so the classification should get worse. However, if a model learns the general pattern and generalizes good enough to overcome this change, the testing is useless. Thus to take the inter-dependency of time points into account, a closer look onto the time points itself is crucial. We propose two new techniques to test and evaluate XAI methods incorporating this hypothesis.

Swap Time Points – The first additional method again takes the time series $t = (t_0, t_1, t_2, ..., t_m)$ and the relevance for it $r = (r_0, r_1, r_2, ..., r_m)$. However, it takes the time points with the relevance over the threshold as the starting point for further changes of the time series. So, $r_i > e$ describes the start point to extract the sub-sequence $t_{sub} = (t_i, t_{i+1}, ..., t_{i+n_s})$ with length n_s. The sub-sequence then gets reversed to $t_{sub} = (t_{i+n_s}, ..., t_{i+1}, t_i)$ and inserted back into the time series. Further, in another experiment, the sub-sequence gets set to zero to test the method. Also, like in the perturbation analysis, the same procedure is done with a random time point positions to verify the time points relevance again.

Mean Time Points – Same as the first additional method, the second one also takes into account the time series $t = (t_0, t_1, t_2, ..., t_m)$ and the relevance for it $r = (r_0, r_1, r_2, ..., r_m)$. Also, it takes the time points with the relevance over the threshold as the starting point for further changes of the time series. However, instead of swapping the time points, the mean μ of the sub-sequence $t_{sub} = (t_i, t_{i+1}, ..., t_{i+n_s})$ is taken to exchange the whole sub-sequence to $t_{sub} = (\mu_{t_{sub}}, \mu_{t_{sub}}, ..., \mu_{t_{sub}})$ and inserted back into the time series. Further, in another experiment, the sub-sequence gets set to zero to test the method. Also, like in the perturbation analysis, the same procedure is done with a random time point positions to verify the time points relevance again.

3.3. Methodology

The methodology to verify an XAI method is conducted in three stages (model training and evaluation, model explanation creation, explanation evaluation, and verification).

1. In the first step, a model learns the training data. Afterward, the trained model predicts the test data and a quality measure (e.g., accuracy) calculates the performance of the result.

2. In the next step, a selected XAI method creates explanations for every sample of the test data. Based on the time point relevance by the explanations, the test data gets changed by the evaluation and verification methods mentioned before.

3. Then, in the last step, each of these newly created test sets gets predicted by the model, and the quality measure is calculated for the comparison.

If the XAI method produces correct explanations, the assumptions $qm(t) \geq qm(t_r^c) > qm(t^c)$ with qm as the quality measure, t the original time series, t_r^c the random changed, and t^c the relevant changed time series, holds.

CNN	Zero	Inverse	Swap	Mean	**RNN**	Zero	Inverse	Swap	Mean	**Paper**	Zero	Inverse	Swap	Mean
Saliency	0.24	**0.45**	0.39	0.34	Saliency	**0.29**	**0.42**	**0.23**	0.22	Saliency	0.06	0.08	0.07	0.07
LRP	**0.44**	0.39	**0.41**	**0.41**	LRP	0.21	0.21	0.14	0.13	LRP	**0.29**	0.29	**0.29**	0.34
DeepLIFT	**0.48**	**0.45**	**0.40**	**0.39**	DeepLIFT	0.00	0.00	0.00	0.00	DeepLIFT	**0.29**	**0.30**	**0.29**	**0.35**
LIME	0.16	0.32	0.17	0.17	LIME	0.10	0.21	0.06	0.07	LIME	0.02	0.06	0.04	0.02
SHAP	0.25	**0.46**	0.33	0.29	SHAP	**0.26**	**0.35**	**0.23**	**0.23**	SHAP	**0.29**	**0.40**	**0.31**	**0.38**
Random	*0.17*	***0.45***	*0.15*	*0.10*	*Random*	*0.13*	*0.23*	*0.03*	*0.03*	*Random*	*0.13*	*0.21*	*0.07*	*0.04*

Table 1. Results table with the averaged changed accuracy of the different models over all datasets. Change to test accuracy is calculated by normalizing the base accuracy to the one from the changed data.

4. Discussion

The discussion divides into three parts. At first, the datasets and employed models are addressed to help to reproduce the experiments. Afterward, the selected XAI methods are introduced in short, giving an overview. Lastly, we discuss the preliminary evaluation results.

4.1. Datasets & Models

Nine datasets of the UCR Time Series Classification Archive [5] and a ECG hearbeat dataset [9] are included in a real-world focused preliminary experiment. These ten datasets, namely FordA, FordB, ElectricDevices, MelbournePedestrian, ChlorineConcentration, Earthquakes, NonInvasiveFetalECGThorax1, NonInvasiveFetalECGThorax2, Strawberry [5], and Physionet's MIT-BIH Arrhythmia [9], consist of two different tasks, binary and multi-class prediction. Primarily, binary classification, e.g., for anomaly detection, is a critical use case for time-series predictions to tackle applications like predictive maintenance or heartbeat categorization.

During the experiments, two different architectures (CNN and RNN) are used as baseline models. If available, the architecture provided by the dataset paper is also incorporated. The considered CNN consists of a 1D convolution layer with kernel and channel size of three. Afterward, a dense layer with 100 neurons learns the classification for a specific problem. The considered RNN consists of an LSTM layer with 100 neurons and again a dense layer with 100 neurons for the classifier. Both networks train each dataset individually for 50 epochs. The paper models consist of ResNet-based architectures and also 50 epochs.

4.2. XAI methods

The experiment is conducted with the five most prominent XAI methods (LIME [20], LRP [2], DeepLIFT [24], Saliency Maps [25], SHAP [16]). LIME employs a so-called surrogate model to explain the decision of an ML model. Thru sampling data points around an example to be explained, it learns a linear model to extract local feature importance for the prediction of the more complex model. By propagating the gradients through the network, Saliency Maps and DeepLIFT build a heatmap as feature importance. LRP propagates a relevance score backward through the underlying model to specify feature importance. SHAP employs shapely values and game theory to find the best fitting feature to gain the most for the prediction.

4.3. Results

Our preliminary results, see Table 1., show that DeepLIFT and LRP have the largest overall quality metric decrease in CNNS for the perturbation and sequence analysis, which shows the working local feature importance. Saliency Maps and SHAP outperform the others in RNNs by showing quality metric decreases, which is somewhat unexpected but shows a need for further exploration of RNNs with XAI methods. In more advanced ResNet architectures, SHAP produces the best results. However, also DeepLIFT and LRP show good results, which again shows the practical local feature importance. LIME shows terrible results in all cases, most likely due to large dimensionality and the employed linear classifier. Further, the results show the desiderata for the sequence verification methods as the random perturbation of time points has a significant quality metric decrease. Our proposed sequence verification methods present more clearly that the assumption $qm(t) \geq qm(t^{mean}_{random}) > qm(t^{mean})$ with qm as the quality measure, t the time series, t^{mean}_{random} and t^{mean} the changed time series holds.

5. Conclusion and Future Work

Our methodology and verification methods show that XAI methods, proposed for images and text, work on time series data by specifying a relevance to time points. The methods also demonstrate that the models take the temporal aspect into account in some cases. In our experiment, we find that SHAP works robust for all models, but others like DeepLIFT, LRP, and Saliency Maps work better for specific architectures. LIME performs worst most likely because of the large dimensionality by converting time to features. However, we also conclude that a demand is given to introduce more suitable XAI methods on time series to guarantee a better human understanding in the process of XAI. As seen by the hard to interpret visual saliency masks (heatmaps) on time series, a need for a more abstract representation is necessary and increases the importance for more sophisticated visual XAI methods on time series.

References

[1] L. Arras, A. Osman, K.-R. Mller, and W. Samek. Evaluating Recurrent Neural Network Explanations. In *8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2019.

[2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Mller, and W. Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, July 2015.

[3] M. C. Chuah and F. Fu. ECG anomaly detection via time series analysis. In *International Symposium on Parallel and Distributed Processing and Applications*, pages 123–135. Springer, 2007.

[4] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2011.

[5] H. A. Dau, E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. *The UCR Time Series Classification Archive*. Oct. 2018.

[6] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. *A Roadmap for a Rigorous Science of Interpretability*, (Ml):1–13, 2017.

[7] European Union. *European General Data Protection Regulation*. `https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en`, 2018.

[8] A. H. Gee, D. Garcia-Olano, J. Ghosh, and D. Paydarfar. Explaining Deep Classification of Time-Series Data with Learned Prototypes. *arXiv preprint arXiv:1904.08935*, pages 1–16, 2019.

[9] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101 23:E215–20, 2000.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A Survey Of Methods For Explaining Black Box Models. *ACM Computing Surveys*, 51(5):93:1–93:42, 2018.

[11] Gunning, D. Explainable Artificial Intelligence (XAI) DARPA-BAA-16-53. Technical report, Defense Advanced Research Projects Agency (DARPA), 2016.

[12] B. Hidasi and C. Gspr-Papanek. ShiftTree: An Interpretable Model-Based Approach for Time Series Classification. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 48–64, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[13] E.-Y. Hsu, C.-L. Liu, and V. S. Tseng. Multivariate Time Series Early Classification with Interpretability Using Deep Learning and Attention Mechanism. In Q. Yang, Z.-H. Zhou, Z. Gong, M.-L. Zhang, and S.-J. Huang, editors, *Advances in Knowledge Discovery and Data Mining*, pages 541–553, Cham, 2019. Springer International Publishing.

[14] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, and others. An empirical evaluation of deep learning on highway driving. *arXiv preprint arXiv:1504.01716*, 2015.

[15] K.-j. Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1-2):307–319, 2003.

[16] S. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 16, pages 426–430, May 2017.

[17] R. K. Mobley. *An Introduction to Predictive Maintenance*. Plant Engineering. Butterworth-Heinemann, Burlington, second edi edition, 2002.

[18] S. Mohseni, N. Zarei, and E. D. Ragan. A Survey of Evaluation Methods and Measures for Interpretable Machine Learning. *arXiv preprint arXiv:1811.11839*, 2018.

[19] J. Redmon and A. Farhadi. YOLOv3: An Incremental Improvement. *ArXiv*, abs/1804.02767, 2018.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?". In *International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, New York, New York, USA, 2016. ACM Press.

[21] C. Rudin. Please Stop Explaining Black Box Models for High Stakes Decisions. *arXiv preprint arXiv:1811.10154*, Nov. 2018.

[22] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. R. Mller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

[23] W. Samek, T. Wiegand, and K.-R. Mller. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv preprint arXiv:1708.08296*, abs/1708.08296, 2017.

[24] A. Shrikumar, P. Greenside, and A. Kundaje. Learning Important Features Through Propagating Activation Differences. *International Conference on Machine Learning*, 2017.

[25] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv preprint arXiv:1312.6034*, pages 1–8, 2013.

[26] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *Computer Vision*, volume 8689, pages 818–833. 2014.