Time Series Model Attribution Visualizations as Explanations

Udo Schlegel*
University of Konstanz

Daniel A. Keim†
University of Konstanz

ABSTRACT

Attributions are a common local explanation technique for deep learning models on single samples as they are easily extractable and demonstrate the relevance of input values. In many cases, heatmaps visualize such attributions for samples, for instance, on images. However, heatmaps are not always the ideal visualization to explain certain model decisions for other data types. In this review, we focus on attribution visualizations for time series. We collect attribution heatmap visualizations and some alternatives, discuss the advantages as well as disadvantages and give a short position towards future opportunities for attributions and explanations for time series.

Index Terms: Human-centered computing—Human computer interaction (HCI); Computing methodologies—Artificial intelligence

1 INTRODUCTION

Explainable AI (XAI) introduces techniques as well as algorithms to open black-box models and support understanding, debugging, as well as refining of complex models [35]. Such techniques are essential to handle the growing democratization of deep learning models [10] and their state-of-the-art performance in an also increasing amount of research fields. Nevertheless, critical application fields like healthcare or criminal justice need explanations to allow the usage of black-box models [22]. Interpretable models are good baselines and alternatives in such critical cases. However, as state-of-the-art performances are often only achieved by black-boxes, easily understandable explanations can overcome a few limitations. Such explanations can help to combine peak performances with understandable decisions of complex models to tackle critical applications [21]. In most cases, XAI techniques, which open such deep learning black-box models, are compromised of local explanation using so-called attribution techniques [8]. Such attribution methods are developed to explain specific input samples for complex models, some model-agnostic [14, 21] and a large amount model-specific [4, 32]. Thus, various attribution techniques exist, and deciding which method to use is a rather tedious task as an evaluation is often either time-consuming with user preferences or computation heavy with automatic approaches [24].

The evaluation of XAI explanations is often divided into two categories: qualitative and quantitative evaluations [15]. The qualitative evaluation focuses on the human aspect and incorporates human understanding of explanations into the evaluation [21]. Quantitative evaluation is often done automatically, focusing on essential parts of the decisions in contrast to the input samples [11]. A current trend focuses on the evaluation of attributions using automatic methods for images [11], text [2], and time series [25] through perturbation analysis. In such an evaluation analysis, the most important regions or parts of the input for the attribution are perturbed to a non-information holding value [42]. However, e.g., for time series such non-information values are not easily defined [25]. Such automatic evaluations are often preferred as easily understandable visualizations for attributions are difficult to create for a general audience. The current trend, e.g., in computer vision, presents heatmaps on the image input to visualize the attributions and relevances of corresponding pixels [8]. For instance, such heatmap visualizations can be transferred to time series but are rather hard to interpret even for experts [24]. Thus, the question arises, how can we support time series attribution visualizations to create understandable and robust explanations for changing demands and user groups.

This review collects and introduces attribution technique concepts for time series and presents visualization techniques for time series attributions. We focus on approaches using heatmaps and collect related papers to present examples. We highlight prominent approaches and discuss the advantages and disadvantages of such heatmap visualizations. Further, we present first alternatives to such heatmaps for time series incorporating line plots. At last, we introduce our position towards attribution visualization for different users and improved explanations as well as future opportunities for explanations on time series.

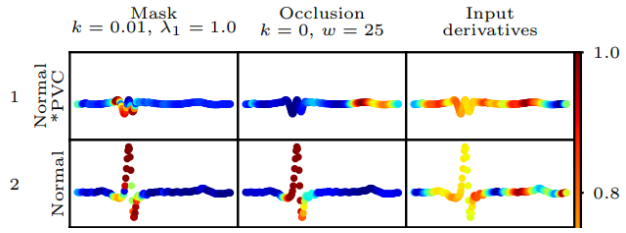
2 ATTRIBUTIONS ON TIME SERIES

Attributions are local explanations of various XAI techniques and help to explain model decisions for single samples. Global explanations present the overall decision-making of models and are hard to achieve for complex models such as deep neural networks. In contrast, local explanations show only the decision-making for a single or a limited amount of input samples. Attributions are a particular form of local explanations, which demonstrate the importance of input variables of a sample based on the predicted output of a model. Thus, such attributions show how a model attributes output predictions to input features based on a single sample.

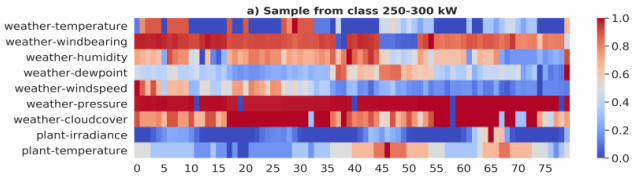
Attributions can essentially help to understand single decisions of a model towards one sample. With different approaches, attribution methods generate a relevance score for every input variable of a sample using the input model as a base. Collecting such relevance for the specific input then creates the attribution vector in the end. In some cases, these attributions show the sensitivity towards specific input variable changes, e.g., occlusion [42]. In contrast, in other approaches, additive attributions get calculated to give each feature value an additive score, e.g., SHAP [14]. There are many different approaches to generate such an attribution vector with other intentions and strategies. Nevertheless, each of these techniques assigns a score towards each input value into the model called attribution.

In general attributions can be generated for every input of every model as they show the importance of the feature on the selected sample regardless of the input dimensionality. However, depending on the data type, other aspects are essential for attributions. Image data has, in many cases, neighborhood importance for the pixels and regions with higher attributions. On the other hand, tabular data does not incorporate neighborhoods as the features can be reordered without further challenges in many cases. Correlations in tabular data are often not generated by the neighborhood, while pixels in images are highly correlated in the surrounding area. Time series, in contrast, also form neighborhoods in the time direction while also adding correlations over features and time. Revealing such correlations is not trivial and requires good attribution techniques with a way to communicate the explanation. Thus, attributions can be calculated for complex time series models but are often difficult to interpret without visualizations or further abstract representations.

*e-mail: u.schlegel@uni-konstanz.de
†e-mail: keim@uni-konstanz.de



(a) Van der Westhuizen and Lasenby [39] visualize their feature importance of an LSTM model trained on healthcare data with the jet color scale (blue to green to red).



(b) Assaf and Schumann [3] visualize attributions for multi-variate time series as a heatmap using the attribution technique Grad-CAM with a color scale blue (small) to red (high).

Figure 1: Van der Westhuizen and Lasenby [39] and Assaf and Schumann [3] visualizations for multi-variate time series attributions as different heatmap approaches, directly on the line plot and as an abstract heatmap on the time points as rectangles. Van der Westhuizen and Lasenby [39] visualize the feature attributions for ECG data with various features and an LSTM model. Assaf and Schumann [3] present the feature attribution of Grad-CAM on different CNNs on energy consumption data on the individual time points and over time.

A growing number of techniques can generate attributions, often categorized into gradient, structure, and surrogate techniques [23]. These categories are based on the method they incorporate to extract attributions and are independent of the explanations. Some of these methods can help to generate easier to understand explanations as they already implement some aggregations to enable abstractions, e.g., showing an attribution through addition on a baseline for the input dimensions [14]. Most others operate directly on the input to the model and present the raw attributions, which in some cases are not easily interpretable [12] or unreliable [13].

Gradient methods such as Saliency [32] and GuidedBackpropagation [36] use the gradient of the model to propagate the importance of specific output neurons back to the input neurons to generate attributions. As gradients are fast to compute for single samples in most cases, gradient methods are thus a good starting point for attributions in general. However, as gradients can be noisy due to the shattered gradients problem [5], gradient ensemble techniques such as SmoothGrad [34] try to overcome these issues by adding noise to the input to improve attributions by smoothing the gradients over the noisy inputs. Others such as Integrated Gradients [38] incorporate a baseline input and slowly changing such a baseline to the selected input capturing the gradients and calculating the integral of the collected gradients to generate the attributions.

Structure methods such as LRP [4] and DeepTaylorDecomposition [17] use the learned network weights and biases to propagate a score from the output to the input using specific weighting rules for each layer type. Through such an auxiliary score, the challenges with computed gradients can be mitigated. However, other challenges arise, such as the selection of rules for the layers with, e.g., LRP [18] or stable references for, e.g., DeepLIFT [29]. By providing a rule of thumb for the rule selection for the layers [16] and the references of the inputs, valuable attributions can be generated with both approaches. Further, as the score is often straightforward to calculate even for deeper and wider models, these techniques are also relatively fast, enabling an easy comparison if the gradients are shattered, or other problems emerged with the gradient methods.

Surrogate and sampling methods such as LIME [21] and SHAP [14] first collect perturbed instances of the input sample and, based on these newly created data points, train an interpretable model or a game-theoretical model to generate the attribution scores. One of the most considerable problems of these methods is the sampling, as better perturbed instances produce better interpretable models and thus better attributions [26]. Also, more perturbed samples lead to better attributions, but generating more perturbed instances takes more time as the model has to be applied to them for the prediction [21]. However, as these methods are model-agnostic and thus applicable to every input model and data, they are highly desired and needed by the industry for production-ready models.

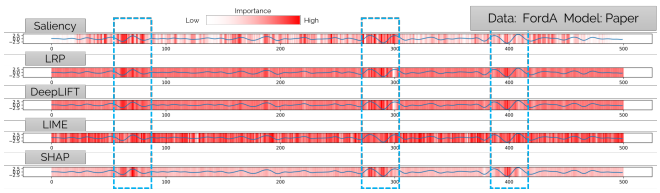
Evaluation is essential for all attribution methods as the amount of possible applicable techniques is quite large, and the fidelity needs to be guaranteed. In many cases, these methods are evaluated using either qualitative [21] or quantitative [11, 24] methods. Schlegel et al. [24] present a quantitative evaluation with a fidelity perturbation analysis to find the most promising attribution techniques for time series, which focus on how accurate the explanations capture the model (trustworthiness). Fig. 2a shows the comparison of a few evaluated approaches on uni-variate time series with a heatmap line plot visualization. They perturb (exchange) time points with high relevance to a non-informative value. As a non-informative value is challenging to define in time series, they use zero, the inverse, and the mean as such a value [24]. These fidelity perturbation approaches enable selecting the most fitting technique for the data type, and model architecture users can apply to time series [25]. However, after detecting a suitable attribution technique, understandable communication as an explanation to users is still necessary as attributions hold relevances for every input variable for the given model.

3 Heatmaps as Basis for Attribution Explanations

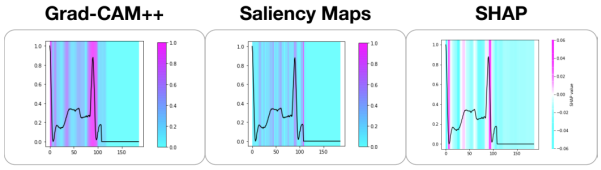
Attribution heatmaps are often used as explanation visualizations for images for attribution techniques. Such a visualization technique presents the relevance value of the corresponding input variable on top of it. Such an approach works quite nicely for images with a heatmap value right on top of the original pixel, and through opacity, it is still possible to combine both variables. However, for instance, time series often use line plots for a baseline visualization, and such a combination of line plots and relevances leads to further challenges.

Van der Westhuizen and Lasenby [39] present how to show activations and attributions for LSTM networks on time series with saliency and occlusion. They remove the line of a line plot and extend the size of the circles corresponding to the data with a color gradient at every time point to show the relevances of the attribution as seen in Fig. 1a. Through such a relatively small and direct approach to a heatmap on time series, challenges occur early. In their examples, it is hard to identify relevant time points due to the color gradient and, in some cases, overplotting. Especially without interaction methods, such a visualization leads to a distorted perception as later ones overdraw early rendered points. However, for small-scale time series with a dynamic size of the circles, the visualization can lead to the first findings into the model.

Siddiqui et al. [31] also use a line plot as a baseline visualization. They extend the line with a gradient towards relevant time points to show the attribution. So, high attributions lead to a particular color (red) for specific parts of the line plot. Thus, the overplotting issue and the generally distorted perception can be mitigated. But, in some cases with bad gradients and coloring, the line plot itself can be rather hard to see, for instance, by having a gradient from



(a) Schlegel et al. [24] visualize attribution techniques for uni-variate time series as a heatmap from low (white) to high (red) in the background of a line plot.



(b) Jeyakumar et al. [12] create visualizations similar to Schlegel et al. [24] with attributions in the background of the line plot ranging from low (cyan) to high (purple) values.

Figure 2: Schlegel et al. [24] and Jeyakumar et al. [12] visualizations for time series attributions focusing on multiple attribution techniques using heatmaps in the background to highlight the relevance. Schlegel et al. [24] show the differences of attribution techniques and conduct a quantitative evaluation which of these attribution techniques are applicable and perform best on time series. Jeyakumar et al. [12] carry out a user study to compare the heatmap visualization of attribution techniques with other explanations such as explanation by example.

white to red or distract the attention to mitigate intervals.

Assaf and Schumann [3] remove the underlying data and show the attribution heatmap as a dense-pixel visualization by presenting each time point as a rectangle with the relevance score as the color as seen in Fig. 1b. By removing the time series data itself and only focusing on the attribution, they can investigate patterns and explain complex model decisions on time series. However, due to limiting their attributions to Grad-CAM [27], some found patterns are challenging to understand as they are even hard to explain with domain knowledge in the time domain by experts.

Viton et al. [40] take the same concept as Assaf and Schumann [3] with another color scale going from light blue to orange and black as the zero value dividing these. They exchange Grad-CAM [27] with some algorithm similar to the CAM approach but focusing more heavily on their neural network architecture. Thus, they improve the attributions of their architecture. However, the visualization still limits the insights into the model itself and is in some cases due to the selected color scale harder to read than the others.

Schlegel et al. [24] combine both approaches by visualizing a line plot of the time series with the attribution heatmap in the background as rectangles and the color of it as the relevance of the attribution. Fig. 2a shows their approaches on various attribution techniques to highlight the contrasting explanations of the different methods. Their approach tries to include the time series and attribution data, making the visualization harder to understand in general. Especially, non-exerts need more time to understand the visualization as the line plots are relatively small while the heatmap is rather large. However, as the idea of the visualization is more about showing the differences of the attribution techniques, the focus shifts in general to experts.

Jeyakumar et al. [12] and Raghunath et al. [20] use a similar approach to Schlegel et al. [24] by presenting the attributions as a gradient behind the line plot (Fig. 2b). Jeyakumar et al. [12] exchange the color gradient towards a more distinct scale from light blue to purple to highlight relevant regions. Further, they conduct a user study to find out if such attribution visualizations help to understand the data or if nearest neighbor examples are better explanations. Their visualization and also the one from Schlegel et al. [24] demonstrate the problems of some of the attribution techniques as high relevance can be next to irrelevant time points. Such challenges are hard to grasp for non-experts and experts alike as we expect the model to not learn specific time points by heart but some temporal patterns. In some cases, these results are not consistent throughout the attribution techniques, which shows the problems of the attribution techniques on time series [24].

Heatmaps have some advantages and disadvantages for users in general. For geospatial data, heatmaps help to encode more information for geo positions. Thus, they help to present more information in an easy-to-understand way for most already intelligible environments (geo-maps or images in general). However, there is a need for easier-to-understand visualizations, especially with time series and XAI explanations. Showing the attention or focus of a black-box model helps identify specific characteristics but does not guarantee an understandable explanation for the model's decision-making. In some cases in XAI and heatmap explanations, such explanations can be fooled and changed to some misguided or just wrong results [33, 37]. Thus, more abstract explanations are needed.

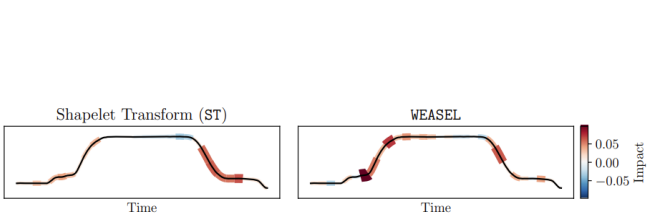
4 Line Plots with Attribution Extensions

Possible alternative abstract representations for attributions are often grounded in line plots and their extensions for time series. Temporal data has a long history of visualization techniques such as line plots, stacked plots, or horizon graphs [1]. However, especially adding more information to the temporal component visualization is not trivial. For instance, multi-variate data is often visualized in small multiples and line plots to incorporate more information [1]. As we want to add more data into the same plot, such a visualization is even harder to compare and interpret for most user groups.

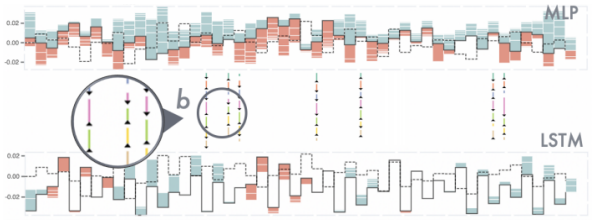
Siddiqui et al. [30] propose with TSinsight an attribution technique but also use multiple line plots to show the attribution as a line plot. Thus, they move the visualization towards small multiples in the form of multiple line plots. Their visualizations show the relevances of time series in another plot which enables a closer look at the initial data and the attributions to, e.g., compare different time points and their relevance more easily. However, the relation between time series and attribution data is more complex because locating the correct time point in the time series data is not trivial by focusing on another attribution line plot. Their focus moves from the combination of the time series and the attributions towards the attributions, which can help, for example, experts with domain knowledge of certain situations, e.g., time points on anomaly events.

Mujkanovic et al. [19] present another technique to generate attributions and an approach towards combining line plots with heatmaps. However, these heatmaps are not just a color gradient on or behind the line of a line plot but mapped on the color and size of a region around the line. They use a pipe around the line to present the relevance of the attributions (Fig. 3a). Such an approach has two advantages; on the one hand, the focus of the user switches to the larger pipes as the size and the color highlight the part of the time series, which can potentially have the highest relevance for the prediction. On the other hand, as little relevance can still be interesting, the color and a small pipe enable users to still find and see the less important regions and parts of the time series.

Line plots are generally the first visualization for temporal data [1]. However, in many cases, the temporal data is the only visualized component without further additional information towards, e.g., explanations. Combining both is challenging and often done with color gradients as heatmaps on the line or behind it in the line plot. However, such heatmaps line plots approaches are either difficult to understand for non-experts or reveal the issues of attribution techniques showing high next to low relevances. Some of these



(a) Mujkanovic et al. [19] use pipes and a color gradient around a line of a line plot to visualize the relevance of an extended SHAP algorithm for time series.



(b) Xu et al. [41] show the additive attributions of SHAP [14] as bar charts on the forecast target and further arrows to compare multiple models attributions.

Figure 3: Mujkanovic et al. [19] and Xu et al. [41] visualizations for time series attributions focusing on either a pipe enhancement for line plots to visualize relevance of SHAP or multiple models using the additive attribution SHAP technique for an improved comparison between various models. Mujkanovic et al. [19] extend the SHAP technique for time series and visualize the corresponding relevance as pipes around the line plot with a color scale. Xu et al. [41] incorporate the additive properties of SHAP to compare the performances of time series forecasting models by enabling explanations of single models and a direct comparison with arrows showing the differences in the attributions.

challenges can get mitigated by easy-to-understand visualizations on line plots with aggregations such as Mujkanovic et al. [19].

5 Future Opportunities

Another approach by Xu et al. [41] incorporates the additive properties of SHAP [14] into their visualization to exchange the heatmap with bar charts on top and beneath a line representing the time series data seen in Fig. 3b. Such an approach enables overcoming attribution heatmaps with more data to allow experts to investigate the decisions of models in more detail. Significantly, such an approach helps present different features' influences in a multi-variate time series on the forecasting target. The bar charts above and beneath the line plot show the additive influences of the SHAP [14] attributions enabling a direct comparison of the features and also even models. Fig. 3b shows their approach on a Multi-Layer Perceptron (MLP) and a LSTM [9] model with a more widespread attribution for the MLP. They even introduce a comparison visualization to compare the attributions shown in (b) with arrows and color lines to encode the differences in direction and value. Their visualization lacks an application towards uni-variate time series as the additive property is not given, but such an approach works nicely for multi-variate time series and multiple models.

In general, such visualizations are hard to understand for non-experts and often also for experts [12]. Jeyakumar et al. [12] compare heatmap visualizations of attributions against explanations by example. Their user study clearly demonstrates that explanations by example are better suited for time series than heatmaps of Grad-CAM++ [7], Saliency [32], and SHAP [14] on time series data. Thus, especially for non-experts, other explanations than attribution heatmap visualizations are preferable and advisable. We argue for incorporating counterfactuals as explanations for time series to non-experts but also for experts as these and contrastive explanations are more robustly grounded in human decision making and their explanations [6]. Counterfactuals are instances of an input sample that flip the predictions of the input sample by changing it only marginally. With the help of the Shneiderman Mantra [28], we propose the following approach for time series explanations. As a first step, counterfactuals demonstrate the first glance as an explanation and give an overview of the model's internal decisions on specific samples. Thus, the original sample is visualized with corresponding counterfactuals visualizations nearby to the sample to enable a direct comparison either in the data domain or some projection of it. In the next step, an in-depth analysis can be supported by attribution visualizations by moving to the data domain level with a combination of time series and attribution data. By providing further interaction techniques, e.g., a what-if analysis, a probing of a user towards the selected sample, and individual time points, enables further investigation into the details of the model. For instance, with counterfactuals and assisted probing of time points, an in-depth analysis of decision boundaries of critical time points is possible to mitigate attacks and analyze the robustness of the model. Thus, the approach focuses on overview first (counterfactuals), zoom, and filter (individual attributions), with details on demand (attributions and time series data interaction). Such an approach includes the advantages of attributions and counterfactuals while mitigating many disadvantages of, e.g., attributions with additional information and limited heatmap visualizations as explanations.

6 Conclusion

We introduced attribution technique concepts for time series and their most common visualization technique (heatmaps). We further collected and presented related work for attribution visualizations for time series models based on these concepts. By analyzing these works, we argue for other visualization options such as enhanced line plots and argue for more abstract visualizations moving away from heatmaps and line plots to present explanations better for attributions. We argue for counterfactuals instead of attributions for time series for non-experts and attributions only for domain experts. At last, we present a small-scale workflow to combine counterfactuals and attributions into one pipeline approach to support non-experts and experts in their analysis of black-box time series models.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826494.

References

[1] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

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

[3] R. Assaf and A. Schumann. Explainable deep neural networks for multivariate time series predictions. In *IJCAI*, pp. 6488–6490, 2019.

[4] 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*, 2015. doi: 10.1371/journal.pone.0130140

[5] D. Balduzzi, M. Frean, L. Leary, J. Lewis, K. W.-D. Ma, and B. McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? *PMLR volume 70 (2017)*, Feb. 2017.

[6] R. M. J. Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-*

19, pp. 6276–6282. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/876

[7] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar 2018. doi: 10.1109/wacv.2018.00097

[8] 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), 2018. doi: 10.1145/3236009

[9] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8), 1997. doi: 10.1162/neco.1997.9.8.1735

[10] F. M. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–20, 2018. doi: 10.1109/TVCG.2018.2843369

[11] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*. 2019.

[12] J. V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, and M. Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33, 2020.

[13] 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. 1–12, 2017. doi: 10.1016/j.jns.2003.09.014

[14] S. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, 2017. doi: 10.3321/j.issn:0529-6579.2007.z1.029

[15] 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.

[16] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[17] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65, 2017.

[18] G. Montavon, W. Samek, and K. R. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, 73, 2018. doi: 10.1016/j.dsp.2017.10.011

[19] F. Mujkanovic, V. Doskoč, M. Schirneck, P. Schäfer, and T. Friedrich. timexplain–a framework for explaining the predictions of time series classifiers. *arXiv preprint arXiv:2007.07606*, 2020.

[20] S. Raghunath, A. E. U. Cerna, L. Jing, D. P. vanMaanen, J. Stough, D. N. Hartzel, J. B. Leader, H. L. Kirchner, M. C. Stumpe, A. Hafez, A. Nemani, T. Carbonati, K. W. Johnson, K. Young, C. W. Good, J. M. Pfeifer, A. A. Patel, B. P. Delisle, A. Alsaid, D. Beer, C. M. Haggerty, and B. K. Fornwalt. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine*, 26(6):886–891, may 2020. doi: 10.1038/s41591-020-0870-z

[21] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778

[22] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

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

[24] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim. Towards a Rigorous Evaluation of XAI Methods on Time Series. In *ICCV Workshop on Interpreting and Explaining Visual Artificial Intelligence Models*, 2019.

[25] U. Schlegel, D. Oelke, D. A. Keim, and M. El-Assady. An empirical study of explainable AI techniques on deep learning models for time series tasks. *Pre-registration workshop NeurIPS*, 2020.

[26] U. Schlegel, D. L. Vo, D. A. Keim, and D. Seebacher. Ts-mule: Local interpretable model-agnostic explanations for time series forecast models. In *Advances in Interpretable Machine Learning and Artificial Intelligence (AIMLAI) Workshop*. ECML-PKDD The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2021.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *International Conference on Computer Vision*, vol. 2017-Octob, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74

[28] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pp. 336–343, 1996. doi: 10.1109/VL.1996.545307

[29] A. Shrikumar, P. Greenside, and A. Kundaje. Learning Important Features Through Propagating Activation Differences. *International Conference on Machine Learning*, 2017. doi: 10.1109/IALP.2010.4

[30] M. S. A. Siddiqui, D. Mercier, A. Dengel, and S. Ahmed. Tsinsight: A local-global attribution framework for interpretability in time-series data. *ArXiv e-prints*, abs/2004.02958:1–16, 4 2020.

[31] S. A. Siddiqui, D. Mercier, M. Munir, A. Dengel, and S. Ahmed. Tsviz: Demystification of deep learning models for time-series analysis. *IEEE Access*, 7:67027–67040, 2019. doi: 10.1109/ACCESS.2019.2912823

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

[33] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. How can we fool LIME and SHAP? Adversarial Attacks on Post hoc Explanation Methods. *arXiv preprint arXiv:1911.02508*, 2019.

[34] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. SmoothGrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[35] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics*, 2019.

[36] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[37] A. Subramanya, V. Pillai, and H. Pirsiavash. Fooling Network Interpretation in Image Classification. *International Conference on Computer Vision*, Dec. 2019.

[38] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*, pp. 3319—3328. JMLR. org, Mar. 2017. doi: 10.1007/s10144-009-0162-4

[39] J. Van Der Westhuizen and J. Lasenby. Techniques for visualizing lstms applied to electrocardiograms. *arXiv preprint arXiv:1705.08153*, 2017.

[40] F. Viton, M. Elbattah, J.-L. Guérin, and G. Dequen. Heatmaps for visual explainability of cnn-based predictions for multivariate time series with application to healthcare. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1–8, 2020. doi: 10.1109/ICHI48887.2020.9374393

[41] K. Xu, J. Yuan, Y. Wang, C. Silva, and E. Bertini. *MTSeer: Interactive Visual Exploration of Models on Multivariate Time-Series Forecast*. Association for Computing Machinery, New York, NY, USA, 2021.

[42] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833. Springer, 2014.