# Conceptual Explanations of Neural Network Prediction for Time Series

Ferdinand Küsters
*IAV GmbH*
Gifhorn, Germany
ferdinand.kuesters@iav.de

Peter Schichtel
*IAV GmbH*
Gifhorn, Germany
peter.schichtel@iav.de

Sheraz Ahmed
*DFKI GmbH*
Kaiserslautern, Germany
sheraz.ahmed@dfki.de

Andreas Dengel
*DFKI GmbH*
Kaiserslautern, Germany
andreas.dengel@dfki.de

***Abstract*—Deep neural networks are black boxes by construction. Explanation and interpretation methods therefore are pivotal for a trustworthy application. Existing methods are mostly based on heatmapping and focus on locally determining the relevant input parts triggering the network prediction. However, these methods struggle to uncover global causes. While this is a rare case in the image or NLP modality, it is of high relevance in the time series domain.**

**This paper presents a novel framework, i.e. Conceptual Explanation, designed to evaluate the effect of abstract (local or global) input features on the model behavior. The method is model-agnostic and allows utilizing expert knowledge. On three time series datasets Conceptual Explanation demonstrates its ability to pinpoint the causes inherent to the data to trigger the correct model prediction.**

***Index Terms*—Machine Learning, Deep Learning, Interpretability, Explainability, Time Series**

## I. Introduction

Deep Neural Networks (DNNs) have been applied successfully in various domains on tasks like regression, classification, or anomaly detection. Due to their ability to extract important features of the input data automatically, they can be easily adapted to new problems [1].

By construction, DNNs are black boxes. Therefore, understanding the reason for a specific network decision or even the overall model behavior is difficult. This lack of transparency significantly hampers the applicability of DNNs in many sectors, e.g. health care, finance, and Industry 4.0. It has already been pointed out in the literature that network explanations are required to fully exploit the potential of DNNs [2].

Explainability of DNNs is an active field of research and a variety of interpretation methods have been proposed [3]. The methods differ strongly in resulting explanations, referring to input parts [4], relevant training samples [5] or to concepts relevant for the network decisions [6].

Most interpretation methods try to assign relevance to individual input parts. There are various variants of such heatmapping methods, for example Integrated Gradients [7], Layerwise Relevance Propagation [8], SmoothGrad [9] or Guided Backpropagation [10]. Other methods, like LIME [11] or Meaningful Perturbation [12] also point out the relevant input parts. These heatmapping methods are especially popular for natural language processing (NLP) and the image domain, as pinpointing towards a special shape or object in the input image or towards certain words makes the network decision more intelligible.

However, the use of heatmapping methods suffers greatly if the important input aspect cannot be localized, but is spread over the whole signal. While this is rarely the case for images, and certainly not meaningful for language processing, it is often an inherent property of time series. Trend, seasonality or frequency ranges are obviously non local, to name a few.

Conceptual Explanation has been developed specially for describing global input properties and is one of the few works directly addressed toward neural network interpretation for the time series domain. A concept is an abstract (local or global) input property that can be manipulated by a suitable filter. Conceptual Explanation evaluates the effect preprocessing the network input by different filters has on the network performance. This makes the method model-agnostic and ensures easily intelligible results.

The main contribution of this work is the introduction and characterization of Conceptual Explanation (Sec. III) as well as its evaluation on different datasets (Sec. IV).

## II. Related Work

Conceptual Explanations is a mask-based interpretation approach. In contrast to [4], [11], [12] it does not mask input regions, but input properties. While region-based masking usually adds unwanted side effects to the input, e.g. jumps and seasonality breaks, this problem does not occur for global filter-based masking.

Heatmapping methods [7], [8], [9], [10] are, as described above, suitable for finding relevant local, but not global input properties. A drawback of these methods is that they are sample-based. The relevant information is not the position of the important pixels (which has no dataset-wide meaning), but the object parts these pixels refer to. Therefore, manual inspection of the highlighted areas and aggregation for many samples is necessary. An automatic extraction together with a statistical evaluation is not possible.

TSXplain [13] combines heatmapping methods for finding the relevant input segments with the computation of statistical time-series properties to provide the user with a more insightful interpretation of the relevant input. As it is still based

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

on heatmapping methods, it also struggles with global input properties.

Kim et al. [6] propose testing concepts via corresponding activation vectors (TCAV). While our method is completely model-agnostic, TCAV utilizes the hidden layers of a network. Another difference worth pointing out is the different meaning of a concept: [6] defines a concept by a set of samples having the property of interest whereas we define a concept by its modification on given data.

Goyal et al. [14] estimate the causal concept effect. As the data generation process is usually unknown, the authors propose to approximate this process by a generative model, e.g. a conditional variational autoencoder. In the time series domain, the data generation process is usually simpler than for images and concepts can rather easily be added to or removed from a signal by using a suitable filter.

Palacio et al. [15] use a fine-tuned autoencoder to extract the input aspects relevant for a network and to suppress unimportant parts. This approach is also suitable if global input properties are relevant for the network. However, manual inspection is again necessary. Furthermore, the modified autoencoder adds some opaqueness to the explanation.

## III. Method

Conceptual Explanation quantifies the effect of different properties of the time series inputs on the network behavior. For this, the data is preprocessed by a filter removing or adding the property before feeding it to the network. This allows spotting both relevant as well as irrelevant signal properties. Each tested filter can be seen as a hypothesis regarding the relevant input aspects, making the result of each evaluation easily intelligible.

Conceptual Explanation makes use of the fact that many relevant concepts for time series can indeed be formulated mathematically. In fact, directly looking at (parts of) the input is much less informative for time series than it is for images. These functional concepts have the advantage of allowing statistical evaluation of their relevance for the model performance on the full data set.

### A. Details

The systematics of Conceptual Explanation are described in Fig. 1. Given a data set and a model trained on said data, one starts with a pre-defined set of basic filters, see Tab. I. Parametric filters allow for a fine-grained look on the effect of different model-properties.

| Filter | Parameter |
|---|---|
| Offset Removal | - |
| Trend Removal | - |
| Moving-Average | Window Size |
| Lowpass | Cutoff Frequency |
| Highpass | Cutoff Frequency |
| Additive Noise | Noise Scale |

TABLE I
Basic Filters used for Conceptual Explanation.

For each filter, model output for the filtered and the unfiltered data are compared. This informs if the checked property was used in the network decision process. Conceptual Explanation is an iterative process where filters can be added by construction of new ones or combination of existing ones. This way an ever deeper understanding of the network prediction can be achieved. In addition, prior knowledge may be utilized by experts to construct sophisticated filtering methods.

The result for each filter is a quantitative, data set wide score of the properties relevance. A manual inspection of individual samples is therefore not required. Note that this method does not add any additional complexity (like, for example, explanatory networks). Therefore, the quality and reliability of the interpretation are easily accessible.

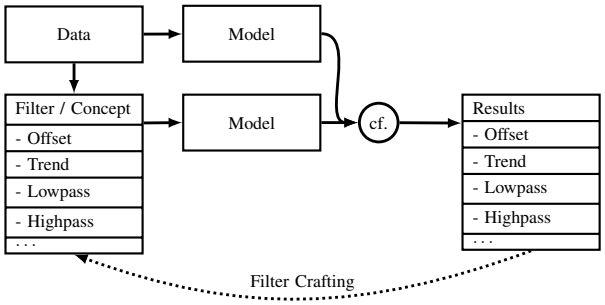


Fig. 1. Structure of Conceptual Explanations: A stack of filters (hypotheses) is tested for the given model and dataset by comparing the network behavior for filtered and unfiltered data. The filter stack is iteratively increased to allow for more detailed insights.

### B. Filter Crafting / Concept Definition

The design of suitable filters is a vital step for Conceptual Explanations, as the quality of the insights provided here directly depend on the filtered signal properties. The iterative filter design by construction allows for very case-specific filters, i.e. case-specific hypotheses on the given model and dataset. One can make use of previously tested filters by (visually) inspecting the effect of relevant filters. A common question in this regard is whether a filter can be localized, i.e. whether it has a similar effect if it is only applied to a subset of the signal defined by another function. An example for this is given in Sec. IV-C3.

Contrary to heatmapping methods, which require the inspection of single samples, Conceptual Explanation, like TCAV [6], requires the design of hypotheses. The hypotheses then need to be translated into a functional form.

While the evaluation of a hypothesis is straightforward, the design of suitable filters depends on the user. To help novice users, a stack of basic filters can be provided. Experts might utilize their dataset insights to define more fine-grained filters and to test their ideas on the network reasoning.

Other works use machine learning or other optimization techniques to find suitable input modifications, e.g. minimal modifications flipping the prediction or an autoencoder trained to reconstruct the essential input components. Transferred to

our setting, these approaches would allow for an easier filter crafting at the cost of (filter) interpretability. In this work, we do not try to automate the concept definition as we would like to ensure that the results are always intelligible to the user.

### C. Local and global concepts

The filters presented so far are related to global properties, as they modify the whole time series. Local properties can be described by filters modifying only few related timesteps, e.g. by masking peaks of the considered signal, see for example Sec. IV-C3.

Once a suitable global filter is found, one can evaluate if the relevant signal property is indeed global or can be localized. For this, the effect of applying the (global) filter only to parts of the signal is investigated. This either ensures that the relevant property is truly globally represented in the signal, or it helps to further specify the relevant aspects by localizing them.

Finally, working the other way round can also be useful: If heatmapping methods point to certain time intervals in the data, but the important aspect of those intervals is not directly intelligible, one can apply Conceptual Explanation on the marked areas to get a better understanding of the network's decision process.

### D. Multivariate Time Series

The procedure can directly be extended to a multivariate setting. One can apply filters to single signals, to all signals, or - based on previous experiments or dataset insights - on suitable subsets of data.

### E. Further Benefits

The Conceptual Explanation of a network can be used for creating new and possibly smaller networks by utilizing it for feature extraction.

Another benefit is that it hints at suboptimal performing networks. If the network performance improves due to a certain filter, one can conclude that the network has not enough capacity or training.

## IV. Evaluation

### A. Data

*1) Trend Anomaly:* The first data set is synthetic, consisting of 100,000 univariate time series with 50 time steps each[1]. The time series are noisy (std 0.6), have random offsets (in $[-10, 10]$) and - in 30% of the samples - a seasonality. A fifth of the data is anomalous, i.e. has a small positive trend (0.02). Samples of both normal signals as well as anomalies are given in Fig. 2.

By design, the relevant property of anomalous signals is global and all time steps of the input signal are equally important for classification. Hence, heatmapping methods fail to provide a meaningful explanation.

[1]The data is available at https://tinyurl.com/TrendAnomaly.

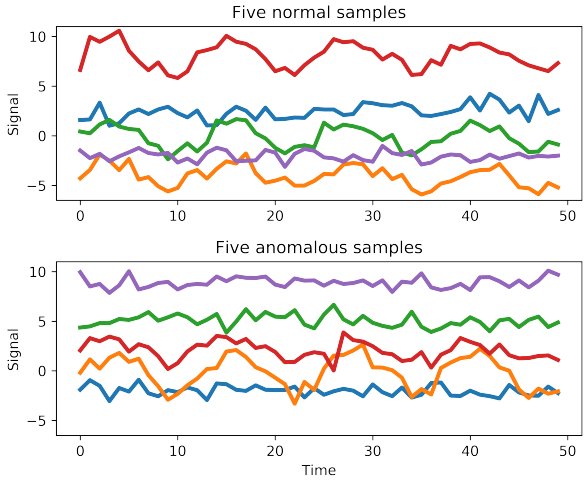


Fig. 2. Normal and anomalous samples of the Trend Anomaly Data Set. The positive trend of anomalous samples is barely visible.

*2) Ford-A:* The FordA-Dataset [16] describes an anomaly detection problem for an automotive subsystem. Each measurement is a univariate time-series of the engine noise. There are 3601 training and 1320 test samples with 500 time steps each. Examples of a normal and an anomalous measurement are given in Fig. 3. The data set was used in a competition in WCCI 2008, details on the anomaly properties are not available.

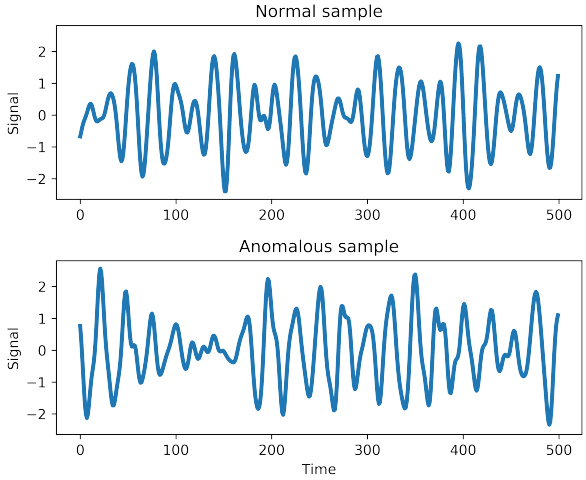


Fig. 3. Normal and anomalous sample of the Ford-A dataset.

*3) Machine Anomaly Detection:* The Machine Anomaly Detection Data Set is a synthetic multivariate time series data set curated by [17]. It consists of 60,000 samples, having three channels with 50 time steps each. The samples labeled anomalous contain a peak in one of the three channels. A third of the data are anomalous. The relevant data property is inherently local. See Fig. 4 for examples.

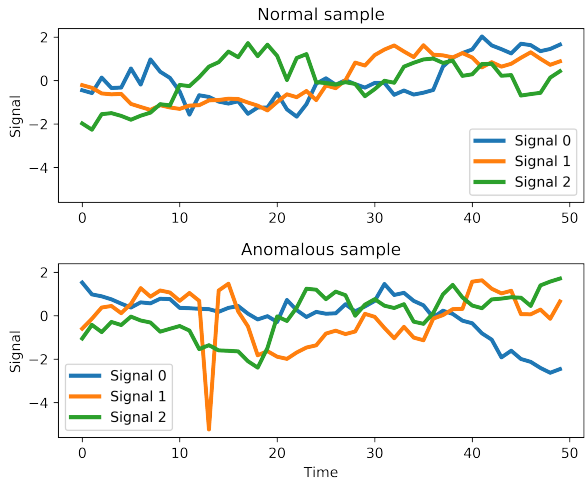


Fig. 4. Normal and anomalous sample of the Machine Anomaly Detection Data Set. The peak in the anomalous sample is clearly visible.

### B. Models

Conceptual Explanation is completely model agnostic. Hence we do not put any emphasis on the network modeling. The explanatory abilities of Conceptual Explanations are illustrated using simple networks. For Trend Anomaly, we use a single layer dense network with a test set accuracy of 96.2%. In the case of Ford-A the network chosen is a five-layer CNN with a test set accuracy of 91.7%. Last but not least the Machine Anomaly Data Set is evaluated on a three-layer CNN with a test set accuracy of 95.8%.

### C. Results

We applied the filters given in Tab. I with different parameters. The presentation here is restricted to the most insightful ones. In order to access the effect of different filters, considering the model accuracy as well as the class recalls turned out to be most helpful. The model accuracy gives a good impression of the overall effect a filter has on the network behavior. The class recall, i.e. the fraction of class samples predicted correctly, allows to access the filter effect on individual classes.

*1) Trend Anomaly:* As one can see in Tab. II, the data offset has no effect on the model prediction. Filtering the data trend drastically reduces the model performance below that of the most simple model. (80% of the data is normal.) While filtering the trend has little effect on the prediction of normal samples, it flips the prediction of most anomalies. This shows that Conceptual Explanation is able to spot the data trend as the signal property crucial for the the network prediction.

As noted before, each data point is equally important for the network prediction. Hence, localization methods like heatmapping cannot give meaningful explanations for this data set. An example is given in Fig. 5.

*2) Ford-A:* The data set has neither trends nor offset. Looking at the effect of smoothing the data by a moving average filter of window size $k$, denoted by $\mathrm{MA}_k$, we obtain

| Filter | Accuracy | Recall Normal | Recall Anomaly |
|---|---|---|---|
| Unfiltered | 96.2% | 98.3% | 88.0% |
| Offset | 96.3% | 98.3% | 88.1% |
| Trend | **70.0%** | 83.0% | **17.0%** |

TABLE II
FILTER EFFECTS FOR THE TREND ANOMALY DATA SET. SMALL NUMBERS INDICATE STRONG EFFECT.

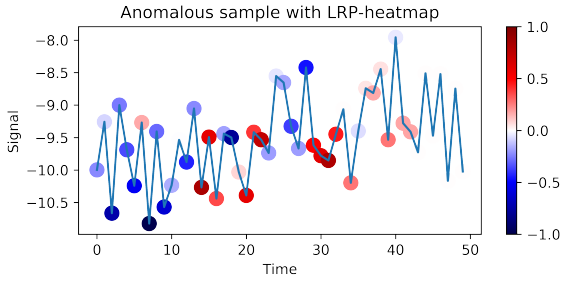


Fig. 5. This anomalous sample of the Trend Anomaly Dataset is correctly classified by the network. The heatmap provided by LRP ([18]) is illustrated by the point markers ranging from blue (very negative effect on the anomaly prediction) to red (very positive effect). Apparently, the heatmap is not helpful for explaining the network prediction of this example. Indeed, a heatmap cannot explain the network prediction for anomalies of this dataset.

the results presented in Tab. III. Here, a closer look at the

| Filter | Accuracy | Recall Normal | Recall Anomaly |
|---|---|---|---|
| Unfiltered | 91.7% | 89.5% | 93.8% |
| $\mathrm{MA}_5$ | 90.5% | 92.6% | 87.8% |
| $\mathrm{MA}_9$ | **65.7%** | 92.6% | **40.4%** |
| Highpass | **52.4%** | **1.6%** | 100% |
| Lowpass | **62.1%** | 94.8% | **31.4%** |

TABLE III
FILTER EFFECTS FOR THE FORD-A DATA SET. SMALL NUMBERS INDICATE STRONG EFFECT.

modified data shows that for the larger smoothing window, many local extrema vanish. Thus washing out the signal too much.

The most insightful results for this model are obtained by looking at different frequency parts of the data. For the high- and lowpass-filter, we set the cutoff frequency to 0.15 times the Nyquist-frequency. This cutoff frequency was chosen by evaluating the effect of different filter parameters.

Removing high-frequency parts by using a lowpass filter

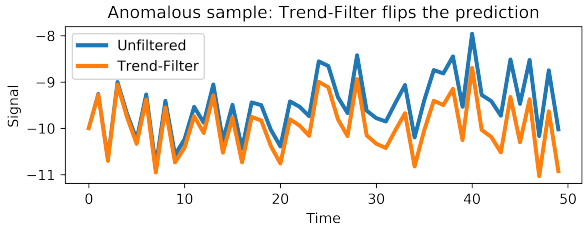


Fig. 6. The Trend filter explains the Trend Anomaly Dataset. As for most anomalous samples, the prediction for the given sample flips after applying the trend filter.

strongly affects the model performance on the anomalous samples, see Tab. III. Contrary the low-frequency part (evaluated by using the highpass filter) is important for the prediction of normal samples. This shows that high frequencies are crucial for the prediction of an anomaly. The moving-average-filter gives a consistent result as it 'smoothens out' high frequencies.

In Fig. 7 both filters are applied to the same sample. Interestingly, it is the highpass-filter which modifies the data visibly. This shows that the high frequencies amount little to the overall signal, while having an enormous effect on the network prediction.

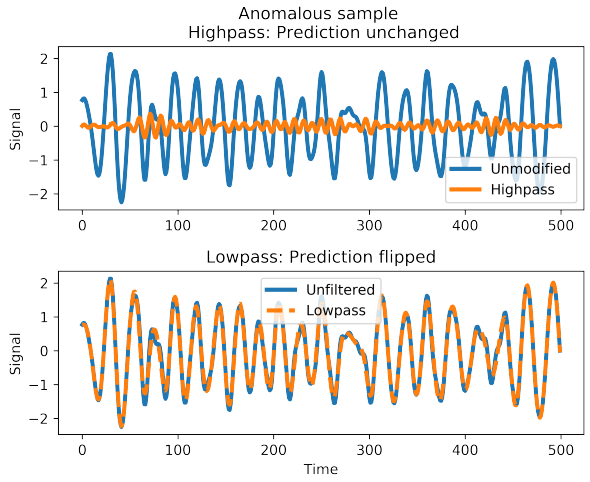


Fig. 7. An anomalous sample of the Ford-A data set together with filtered versions of the sample. The lowpass has visually little impact on the signal, but changes the model prediction. The highpass affects the signal strongly, but does not affect the model prediction.

Again the relevant input properties are indeed global: To see this, we compare the $\mathrm{MA}_9$ mask to a localized version of it. As shown in Tab. III, the $\mathrm{MA}_9$ mask has an immense effect on the prediction of anomalous signals. By $\mathrm{MA}_9^{\mathrm{firsthalf}}$ and $\mathrm{MA}_9^{\mathrm{secondhalf}}$ we define variants of $\mathrm{MA}_9$ restricted to the first and second half of the time interval, respectively. Compared to $\mathrm{MA}_9$, their effect on the anomaly recall is rather small, see Tab. IV. Even combined, the cannot explain the large drop in the anomaly recall of $\mathrm{MA}_9$. One gets comparable results using other subset strategies. Thus, the network apparently aggregates information over the full sample rather than focusing on short intervals. As one would expect by this, heatmapping methods do not yield helpful insights here.

| Filter | Accuracy | Recall Normal | Recall Anomaly |
|---|---|---|---|
| Unfiltered | 91.7% | 89.5% | 93.8% |
| $\mathrm{MA}_9$ | **65.7%** | 92.6% | **40.4%** |
| $\mathrm{MA}_9^{\mathrm{firsthalf}}$ | 88.3% | 93.0% | 84.0% |
| $\mathrm{MA}_9^{\mathrm{secondhalf}}$ | 90.5% | 90.1% | 90.8% |

TABLE IV
FILTER EFFECTS FOR THE FORD-A DATA SET. SMALL NUMBERS INDICATE STRONG EFFECT.

*3) Machine Anomaly Detection:* An additive noise (normally distributed with zero mean and standard deviation 0.5, denoted by $\mathrm{AddNoise}_{(0,0.5)}$) hampers the prediction of normal samples while not affecting the model performance for anomalies, see Tab. V. A moving-average filter of window size five ($\mathrm{MA}_5$), on the other hand, is able to mask the anomalies almost completely. $\mathrm{MA}_5$ is a global filter, affecting the whole signal. By restricting this mask to particular time intervals, we can show that the relevant signal property is actually local.

| Filter | Accuracy | Recall Normal | Recall Anomaly |
|---|---|---|---|
| Unfiltered | 95.8% | 97.8% | 91.7% |
| $\mathrm{AddNoise}_{(0,0.5)}$ | **65.8%** | **52.4%** | 92.6% |
| $\mathrm{MA}_5$ | **67.9%** | 100.0% | **3.3%** |
| MaskPeak | **67.6%** | 98.7% | **4.9%** |

TABLE V
FILTER EFFECTS FOR THE MACHINE ANOMALY DATA SET. SMALL NUMBERS INDICATE STRONG EFFECT.

The filter can be localized by applying it only to regions where the signal $x$ deviates significantly from the averaged signal $\mathrm{MA}_5(x)$, i.e. where it holds

$$|x - \mathrm{MA}_5(x)| > \underbrace{3 \cdot \mathrm{std}(x - \mathrm{MA}_5(x))}_{=:K}$$

where std is the standard deviation. This results in a peak filter:

$$\mathrm{MaskPeak}(x) = \begin{cases} \mathrm{MA}_5(x), & \text{if } |x - \mathrm{MA}_5(x)| > K, \\ x, & \text{else.} \end{cases}$$


Fig. 8. Applying the $\mathrm{MA}_5$-filter flips the prediction for the anomalous sample, but affects the full time series. The MaskPeak-filter has almost the same effect on the network prediction, but modifies only few data points. Here, it affects the peak of Signal 1. (Modified data areas are marked gray.)

The MaskPeak-Filter has a comparable effect on the model as the $\mathrm{MA}_5$-Filter while only modifying 0.9% of the data.

An example is given in Fig. 8. Hence, the explanation given by this refined filter is much more precise than that of the moving-average filter. It shows that the relevant data aspects are local and also allows their localization. Tab. VI shows that the Conceptual Explanation given by the peak-filter is comparable to that of Layerwise-Relevance-Propagation (LRP) in terms of both completeness of the explanation and precision.

| Method | Masked Data Points | Accuracy After Masking |
|---|---|---|
| Peak Filter | **0.9%** | **67.6%** |
| LRP [8] | 2.2% | 70.4% |

TABLE VI
COMPARISON OF CONCEPTUAL EXPLANATION (PEAK-FILTER) AND LRP FOR THE MACHINE ANOMALY DATA SET.

## V. CONCLUSION

Conceptual Explanations is a novel neural network interpretation framework designed specifically for time series. It is conceptually simple and transparent, but allows for easily understandable, quantitative and data set based evaluations.

With the help of hand crafted data, namely the trend anomaly, Conceptual Explanations demonstrates its functionality in principle. Furthermore, using the publicly available Ford-A data Conceptual Explanations is able to pinpoint the trigger of the network decision to the existence of non-local high frequencies present in the signal. A result certainly not possible with local methods and a fact unknown at least to the authors. Last but not least Conceptual Explanations is able to handle the transition from global to local allowing to interact with other methods, too.

Little user-interaction is required for this method, whose insights differ significantly from other network interpretation methods. Therefore, more useful and most notably statistically meaningful interpretations can be expected in the future.

## REFERENCES


[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT press, 2016.

[2] Robert Andrews, Joachim Diederich, and Alan B Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-based systems*, vol. 8, no. 6, pp. 373–389, 1995.

[3] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 93:1–93:42, 2018.

[4] Scott M Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[5] Pang Wei Koh and Percy Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1885–1894.

[6] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," *arXiv preprint arXiv:1711.11279*, 2017.

[7] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3319–3328.

[8] Alexander Binder, Sebastian Bach, Gregoire Montavon, Klaus-Robert Müller, and Wojciech Samek, "Layer-wise relevance propagation for deep neural network architectures," in *Information Science and Applications (ICISA) 2016*, pp. 913–922. Springer, 2016.

[9] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[10] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.

[12] Ruth C Fong and Andrea Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3429–3437.

[13] Mohsin Munir, Shoaib Ahmed Siddiqui, Ferdinand Küsters, Dominique Mercier, Andreas Dengel, and Sheraz Ahmed, "TSXplain: Demystification of DNN Decisions for Time-Series Using Natural Language and Statistical Features," in *International Conference on Artificial Neural Networks (ICANN-2019)*. 2019, vol. 11731, pp. 426–439, Springer, Cham.

[14] Yash Goyal, Uri Shalit, and Been Kim, "Explaining classifiers with causal concept effect (cace)," *arXiv preprint arXiv:1907.07165*, 2019.

[15] Sebastian Palacio, Joachim Folz, Jörn Hees, Federico Raue, Damian Borth, and Andreas Dengel, "What do deep networks like to see?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3108–3117.

[16] "Ford-A Anomaly Detection Dataset," www.timeseriesclassification.com/description.php?Dataset=FordA.

[17] Shoaib Ahmed Siddiqui, Dominique Mercier, Mohsin Munir, Andreas Dengel, and Sheraz Ahmed, "Tsviz: Demystification of deep learning models for time-series analysis," *IEEE Access*, vol. 7, pp. 67027–67040, 2019.

[18] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans, "innvestigate neural networks!," *CoRR*, vol. abs/1808.04260, 2018.