Pattern Recognition 150 (2024) 110309

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Explainable AI for time series via Virtual Inspection Layers

Johanna Vielhaben [a], Sebastian Lapuschkin [a], Grégoire Montavon [c,b,d], Wojciech Samek [a,b,d,*]

[a] Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
[b] Department of Electrical Engineering and Computer Science, Technische Universität Berlin, 10587 Berlin, Germany
[c] Department of Mathematics and Computer Science, Freie Universität Berlin, 14195 Berlin, Germany
[d] BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany

ARTICLE INFO

Keywords:
Interpretability
Explainable Artificial Intelligence
Time series
Discrete Fourier Transform
Invertible transformations
Audio classification

ABSTRACT

The field of eXplainable Artificial Intelligence (XAI) has witnessed significant advancements in recent years. However, the majority of progress has been concentrated in the domains of computer vision and natural language processing. For time series data, where the input itself is often not interpretable, dedicated XAI research is scarce. In this work, we put forward a *virtual inspection layer* for transforming the time series to an interpretable representation and allows to propagate relevance attributions to this representation via local XAI methods. In this way, we extend the applicability of XAI methods to domains (e.g. speech) where the input is only interpretable after a transformation. In this work, we focus on the Fourier Transform which, is prominently applied in the preprocessing of time series, with Layer-wise Relevance Propagation (LRP) and refer to our method as *DFT-LRP*. We demonstrate the usefulness of *DFT-LRP* in various time series classification settings like audio and medical data. We showcase how DFT-LRP reveals differences in the classification strategies of models trained in different domains (e.g., time vs. frequency domain) or helps to discover how models act on spurious correlations in the data.

1. Introduction

The field of XAI has produced numerous methods that illuminate on the reasoning processes of black box machine learning models, in particular deep neural networks. *Local* XAI methods quantify the contribution of each input feature toward the model output on a per-sample basis. Prominent examples such as Layer-wise Relevance Propagation (LRP) [1], Integrated Gradients [2], LIME [3] or SHAP [4] provide valuable insights into the intricate decision function of a neural network. The feature-wise attribution scores they produce are usually presented as a heatmap overlaying the sample [5], such that they guide the eye to the important parts of the sample. In this manner, it is the human user who assumes the responsibility of conducting the actual interpretation, e.g. “The model focuses on the dog’s ears.”. These explanations work well for images or text, where XAI methods can rely on the visual interpretability of feature relevance scores. Here, we can observe the rationale behind the predominant development and testing of XAI methods within the domains of computer vision or natural language processing domains. The implicit requirement of feature interpretability is particularly challenged in the context of time series data, where single or collective time points are often not meaningful for humans [6]. To exemplify, consider the simple case of a model that classifies the frequency of a single sinusoid. Here, it is not important which minima or maxima the XAI method highlights, but how far the highlighted features are apart. In the more realistic case of a superposition of multiple sinusoids, it is impossible for the human user to derive the classification strategy from the heatmap. We see this as a reason, why only limited XAI research is available for time series [7].

In this study, we propose the concept of enhancing the interpretability of explanations for time series, by propagating them to an interpretable representation via a *virtual inspection layer*. A natural choice for an interpretable representation of time series is in the frequency or time–frequency domain. These domains are connected to the time domain via linear invertible transformations, namely the Discrete Fourier Transform (DFT) and the Short Time Fourier Transform (STDFT). We leverage this to propagate relevance scores for models trained in the time domain into the frequency or time–frequency domain without model re-training and without causing any change to the decision function. See Fig. 1 for an illustration of an audio signal and model relevances in all three domains. This idea generalizes to any other invertible linear transformation of the data or a representation in latent feature space that renders it interpretable.

[*] Corresponding author at: Department of Artificial Intelligence, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany.
E-mail addresses: johanna.vielhaben@hhi.fraunhofer.de (J. Vielhaben), sebastian.lapuschkin@hhi.fraunhofer.de (S. Lapuschkin), gregoire.montavon@fu-berlin.de (G. Montavon), wojciech.samek@hhi.fraunhofer.de (W. Samek).

https://doi.org/10.1016/j.patcog.2024.110309
Received 3 March 2023; Received in revised form 8 January 2024; Accepted 31 January 2024
Available online 2 February 2024

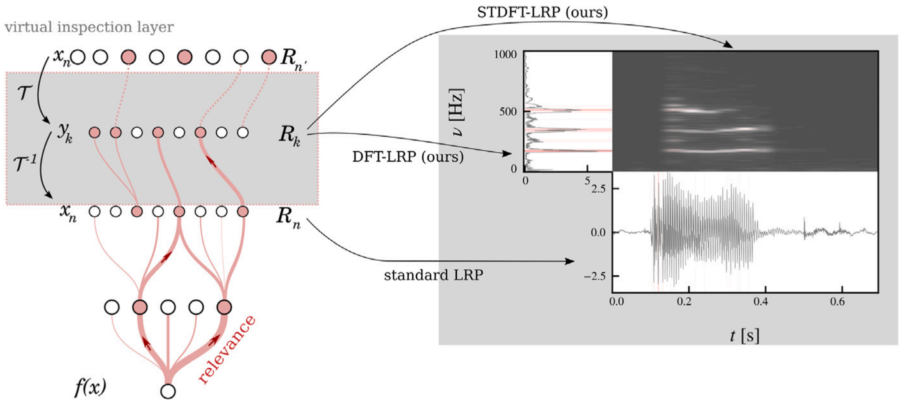

Fig. 1. Schematic overview of virtual inspection layers and (ST)DFT-LRP. **Left**: A virtual inspection layer is inserted before the original input layer, performing a transformation $\mathcal{T}$ of the original input data $x = \{x_n\}$ to an interpretable representation $y = \{y_k\}$ and back. Relevance is propagated from the output $f(x)$ to the original input x via LRP to arrive at relevance scores R_n. These are then propagated further through $\mathcal{T}^{-1}$ for relevance scores R_k on the interpretable representation y. **Right**: Explanations for sex classifier operating on raw waveforms of voice recordings. With standard LRP, only relevance in the time domain (lower panel) is accessible, which is distributed rather uniformly over the part of the signal with large amplitude, making it impossible to derive a classification strategy. For time series, DFT or STDFT as choices for $\mathcal{T}$ lead to DFT-LRP and STDFT-LRP and relevances in frequency (left panel) or time–frequency (center panel) domain. Only here it becomes apparent that the model is focusing on a fundamental frequency between 180 and 200 Hz and subsequent harmonics which is typical for female voices.

Technically, we attach two linear layers to the input: one that transforms the data from the original representation to the interpretable representation, and one that transforms it back (unmodified) to the original data format. Then, any local XAI method can be used to quantify relevance within the *new* and interpretable domain. Here, (modified) backpropagation-based methods like gradient-based methods or LRP have the advantage, that one needs to propagate the relevance scores only one layer further, i.e., just the original relevance scores are required instead of the entire model. Because of its successful application in a wide range of domains [6], in this work we will focus on LRP and refer to our method as *DFT-LRP*. However, we would like to stress that our idea of the virtual inspection layer can be combined with any other local XAI technique.

We see applications of our method in particular in domains involving acoustic or sensory data where interpretability of raw time series features, such as individual time points, poses significant challenges. First, our approach can be employed to render the explanations for an existing model trained in the time domain more interpretable, without the need to retrain a model in another domain. In other words, our approach allows interpreting a given model in (the original) time domain as well as in (the virtually constructed) frequency or time–frequency domain, practically without any additional overhead. Second, we can compare the strategies of models trained in different domains on the same representation of the data. In particular, in audio classification, finding the best input data representation, i.e. raw waveforms vs. spectrograms with different filters, is an important research question [8]. Here, our approach provides a well-informed basis for the selection of the final model, based on the model strategies (and their alignment with prior knowledge) beyond the measure of predictive accuracy.

Our contributions are the following:

- We propose a new form of explanation for models trained on time series data, that highlights relevant time steps as well as frequencies.
- We present a closed-form expression for relevance propagation through DFT and STDFT.
- We expand the scope of pixel-flipping-based evaluations to enable a comprehensive comparison of explanations presented in different formats, including time, frequency, or time–frequency representations.
- We demonstrate how DFT-LRP provides valuable insights on ML model strategies employed by audio and ECG classifiers in frequency domain. Additionally, we highlight how DFT-LRP can unveil potential "Clever Hans" strategies employed by these models [9].

In summary, we put forward a virtual inspection layer that allows for explanations of ML models of which the inputs are not directly interpretable. Our method does not require any model retraining or approximation. When used in combination with backpropagation-based methods such as LRP our method simply requires propagating one layer further. Our method is however also applicable alongside a broad family of local XAI methods, thereby widening the general applicability of XAI to time series models.

This paper is organized as follows: In Section 2, we give an overview of previous approaches of XAI for time series. We introduce the concept of virtual inspection layers and derive a closed-form expression for DFT-LRP in Section 3. In Section 4, we present experiments that qualitatively and quantitatively confirm the effectiveness of our method before we conclude in Section 5.

2. Related work

Prominent applications of deep learning for time series modeling include the domain of audio processing [8,10], the analysis of electronic health records like ECG or EEG [11] or forecasting in fields like finance [12] and addressing challenges related to public health.

Following its primary focus on computer vision and natural language processing, the field of XAI has experienced a notable surge of research efforts dedicated to time series analysis, see [7,13] for a systematic review. Often, XAI methods that originated from other domains such as computer vision can be readily applied to time series classifiers, as they are based on the same architectures such as CNNs or RNNs [14]. Prominently, LRP has been applied to explain time series classifiers in the domain of human gait analysis [15], audio classification [16], as well as ECG [17] and EEG analysis [18]. Other examples of established XAI methods applied to time series are Gradient × Input for ECG data [19], Integrated Gradient for hydrology [20], or [21] adopted Grad-CAM for generic time series from the UCR dataset. All of these methods produce attribution scores for single time points. However, the explanations cannot rely on the visual interpretability of the single features, i.e. time points, which limits their usability for time series [13]. This observation is in line with [5], who compare XAI methods across input domains (including computer vision, natural language processing, and time series) and find that users prefer nearest matching training samples as explanations over input overlaid by relevance scores for time series. In [22], the authors promote training time-series classifiers in frequency and time–frequency domain in order to make post-hoc explanations by Shapley values or Sobol indices more interpretable. By incorporating a virtual inspection layer that enables the assessment of relevance scores in an interpretable domain, e.g. the frequency or time–frequency domain, our approach facilitates explanations produced by all aforementioned feature-wise post-hoc XAI methods for classifiers operating in the time domain.

Recently, a novel category of XAI methods has emerged, generating concept-based explanations [23–25], that have been applied to hidden feature layers of time series classifiers [24,26]. While concept-based explanations increase interpretability by contextualizing the explanation with the help of concept prototypes, they still suffer from the limited interpretability of the input features they are based on.

Further, there is a multitude of XAI methods designed and specialized on time series data: In [27], CNNs are visualized by clustering filters and measuring their influence based on the gradient, [28] measures the impact of user-defined filters applied in the input space via classification accuracy, [29] constructs surrogate models using shapelets, and [30] explains predictions using counterfactual samples from the training set. These approaches try to improve interpretability by introducing novel forms of explanations that differ from traditional heatmaps generated by methods such as LRP or IG. One drawback of these strategies is the absence of theoretical guarantees for the explanation, in particular the lack of relevance conservation, implying that relevance scores sum up to the prediction. Another prominent research area in XAI is the evaluation of explanation techniques. While most evaluation techniques have been developed for general input domains such as pixel flipping and localization [31], few works address the question of specifically evaluating on time series data. Here, [32] proposes a method to evaluate model fidelity of single time point attribution XAI methods. However, we note that so far, there are no existing evaluation techniques available to compare explanations based on features from different representation domains as units of interpretability (e.g. time or frequencies), an aspect which we address in Section 4.3 of this paper.

3. Using LRP to propagate relevance to interpretable representations

3.1. Virtual inspection layer

Let us view a neural network as a composition of functions,

$$f(x) = f_L \circ \cdots \circ f_1(x).$$

where each function can be e.g. a layer or a block. We can quantify the relevance $R_f(x_i)$ of each feature i in x towards $y = f(x)$ by a local XAI method. While the representation of the datapoint x is not interpretable for humans, we assume there is an invertible transformation $\mathcal{T}(x) = \tilde{x}$, that renders x interpretable. Without the need to retrain the model on the representation $\tilde{x}$, we can now quantify the relevance of $\tilde{x}_i$,

$$f(x) = f_L \circ \cdots \circ f_1 \circ \mathcal{T}^{-1} \circ \underbrace{\mathcal{T}(x)}_{\tilde{x}}, \tag{1}$$

and compute the relevance scores $R'_f(\tilde{x}_i)$ for the interpretable representation of the data. In general, an interpretable-representation-inducing bottleneck can be inserted at any layer of the network, e.g.

$$f(x) = f_L \circ \cdots \circ \mathcal{T} \circ \mathcal{T}^{-1} \circ \cdots \circ f_1(x).$$

In the following, we will specialize to DFT regarding $\mathcal{T}$, as our focus in on time series classification, and LRP regarding the local XAI method.

3.2. Brief review of LRP

LRP is a backpropagation-based local XAI method, which decomposes the output of a deep neural network in terms of the input features in a layer-by-layer fashion to arrive at the *relevance* of the input features towards the final prediction. Its central property is the conservation of relevance at each layer. LRP propagates relevance R_j from layer with neurons j to the layer below with neurons i, by summing over all relevances passed from neurons j to neuron i,

$$R_i = \sum_j R_{i \leftarrow j}\,. \tag{2}$$

Generically,

$$R_{i \leftarrow j} = \frac{z_{i,j}}{\sum_i z_{i,j}} R_j \tag{3}$$

where $z_{i,j}$ quantifies how much neuron i contributed towards the activation of neuron j, and is usually dependent on the activation a_i and the weight w_{ij} between the neurons. The sum in the denominator ensures the conservation property $\sum_i R_i = \sum_j R_j$. There are numerous choices for $z_{i,j}$ corresponding to propagation rules. Which rules to choose depends on the model under consideration (see e.g. [33]). To summarize, LRP propagates relevance scores R_j at layer j onto neurons of the lower layer i by the rule,

$$R_i = \sum_j \frac{z_{i,j}}{\sum_i z_{i,j}} R_j \tag{4}$$

until the input layer is reached.

3.3. Relevance propagation for the discrete fourier transform

For a neural network trained in time domain, we can employ Eq. (4) to quantify the relevance of each time step towards the prediction. Here, we lay out how to propagate relevance one step further into the frequency domain. A signal in time domain x_n, $n = 0, \dots, N-1$ is connected to its representation in frequency domain $y_k \in \mathbb{C}$, $k = 0, \dots, N-1$, via the DFT. The DFT and its inverse are simply linear transformations with complex weights,

$$y_k = \mathrm{DFT}(\{x_n\}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_n \left[\cos(\frac{2\pi kn}{N}) - i \sin(\frac{2\pi kn}{N})\right] \tag{5}$$

$$x_n = \mathrm{DFT}^{-1}(\{y_k\}) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} y_k \left[\cos(\frac{2\pi kn}{N}) + i \sin(\frac{2\pi kn}{N})\right]. \tag{6}$$

We require relevances of y_k to be real. Thus, we proceed by writing the signal in frequency domain as a concatenation of real and imaginary parts, $[\mathrm{Re}\, y_0, \mathrm{Re}\, y_1, \dots, \mathrm{Re}\, y_{N-1}, \dots, \mathrm{Im}\, y_1, \dots, \mathrm{Im}\, y_{N-1}]$. As visualized in Fig. 1, we attach a layer that performs the inverse DFT in Eq. (6) to the model, before the first layer f_1 that operates on the signal in time domain. For real valued signals $x_n \in \mathbb{R}$ we can express the inverse DFT as,

$$x_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \mathrm{Re}(y_k) \cos\left(\frac{2\pi kn}{N}\right) - \mathrm{Im}(y_k) \sin\left(\frac{2\pi kn}{N}\right). \tag{7}$$

We assume that relevance values R_n for x_n are available and that they are of form $R_n = x_n c_n$ (a property ensured by most LRP rules, in particular, LRP-$0/\epsilon/\gamma$ [6]). Now, the question is how to transform relevance in time domain R_n to relevance in frequency domain. For LRP, the first question is how much each frequency component $\mathrm{Re}(y_k), \mathrm{Im}(y_k)$ contributes to each time point x_n, i.e. finding an expression for $z_{i,j}$ in Eq. (4). The inverse DFT in Eq. (7) is only a homogeneous linear model, i.e. of type $f(x) = w^\top x$. Thus, we can quantify the contribution of neuron $\mathrm{Re}(y_k)$, $\mathrm{Im}(y_k)$ to x_n by the value of the neuron itself times the weight,

$$z_{k,\mathrm{Re},n} = \mathrm{Re}(y_k) \cos(\frac{2\pi kn}{N}), \tag{8}$$

$$z_{k,\mathrm{Im},n} = -\mathrm{Im}(y_k) \sin(\frac{2\pi kn}{N}). \tag{9}$$

In fact, it can easily be shown, that LRP-0, Deep Taylor Decomposition, Integrated Gradients, PredDiff, and Shapley values all default to neuron times weight for homogeneous linear transformations if one sets the respective reference value to zero [6].

Now, we apply Eq. (4) to aggregate the contributions of each neuron $R_{k,\mathrm{Re}}$, $R_{k,\mathrm{Im}}$ towards the model output and find,

$$R_{k,\mathrm{Re}} = \mathrm{Re}(y_k) \sum_n \cos(\frac{2\pi kn}{N}) \frac{R_n}{x_n} \tag{10}$$

$$R_{k,\mathrm{Im}} = -\mathrm{Im}(y_k) \sum_n \sin(\frac{2\pi kn}{N}) \frac{R_n}{x_n} \,. \tag{11}$$

Here, we assume $R_k = 0$ if $x_k = 0$ and define $0/0 = 0$. In practice, we add a small-valued constant ϵ to the denominator for numerical stability. Now, leveraging additivity of LRP attributions, we define $R_k = R_{k,\mathrm{Re}} + R_{k,\mathrm{Im}}$. To abbreviate the form of the sum, we separate y_k into amplitude r_k and phase φ_k, i.e. $\mathrm{Re}(y_k) = r_k \cdot \cos(\varphi_k)$ and $\mathrm{Im}(y_k) = r_k \cdot \sin(\varphi_k)$, and find,

$$\boxed{R_k = r_k \sum_n \cos(\frac{2\pi kn}{N} + \varphi_k) \frac{R_n}{x_n}} \,. \tag{12}$$

3.4. Relevance propagation for the short-time discrete fourier transform

For slowly varying, quasi-stationary time series like audio signals, one is interested in how the frequency content varies over time. Here, one applies the short-time DFT (STDFT) which connects the signal in time to the time–frequency domain. For the STDFT, one computes the DFT of potentially overlapping windowed parts of the signal [34],

$$v_{m,k} = \mathrm{DFT}(\underbrace{x_n \cdot w_m(n)}_{sm,n}) \,, \tag{13}$$

where $w_m(n)$ is a window function with window width H selecting the segment of the signal to be analyzed while convolving in steps of $m \cdot D$ time points over the input sequence. To sequentially cover the whole signal, we require $0 < D \leq H$ for the shift length. To recover the original signal $\{x_n\}$ given $\{S_{m,n}\}$, we first compute the inverse DFT in Eq. (6) of $\{v_{m,k}\}$ to obtain $\{s_{m,n}\}$. Second, we rescale $\{s_{m,n}\}$ by the sum over the windows $w_m(n)$ over shifts m to obtain $\tilde{x}_n$:

$$\tilde{x}_n = \frac{\sum_m \mathrm{DFT}^{-1}(\{v_{m,k}\})}{\sum_m w_m(n)} \,. \tag{14}$$

This so-called weighted overlap-add technique imposes only a mild condition on the windows $w_m(n)$ for perfect reconstruction $\tilde{x}_n = x_n$, which is,

$$\sum_m w_m(n) \neq 0 \ \ \forall n.$$

In the following, we write $W_n = \sum_m w_m(n)$. In the supplementary material, we show an alternative formulation of the inverse STDFT, which imposes stricter conditions on the windows. Analogous to Eq. (10), we propagate the relevance $R(x_n)$ to the real $\mathrm{Re}(z_{mk}) = r_{m,k} \cos(\varphi_{m,k})$, and imaginary part $\mathrm{Im}(z_{mk}) = r_{m,k} \sin(\varphi_{m,k})$ of z_{mk},

$$R_{m,k,\mathrm{Re}} = r_{m,k} \cos(\varphi_{m,k}) \sum_n \cos(\frac{2\pi kn}{N}) \cdot W_n^{-1} \frac{R_n}{x_n}$$

$$R_{m,k,\mathrm{Im}} = -r_{m,k} \sin(\varphi_{m,k}) \sum_n \sin(\frac{2\pi kn}{N}) \cdot W_n^{-1} \frac{R_n}{x_n} \,.$$

Aggregating the relevance of real and imaginary part yields,

$$\boxed{R_{m,k} = r_{m,k} \sum_n \cos(\frac{2\pi kn}{N} + \varphi_{m,k}) \cdot W_n^{-1} \frac{R_n}{x_n}} \,. \tag{15}$$

We now specialize to an appropriate choice for the window function $w_m(n)$. The DFT of the product between the signal in time domain and the window function is the convolution between the DFT of the original signal and the DFT of the windowing function. Thus, the latter introduces new frequency components, known as spectral leakage.[1] Depending on the shape of the windowing function, spectral leakage can cause two opposing issues. On the one hand, it can restrict the ability to resolve frequencies that are very close but have a similar amplitude (*low resolution*). On the other hand, it can limit the ability to resolve frequencies that are far apart from each other but have dissimilar frequencies (*low dynamic range*). Windows with a rectangular shape have a high resolution but a low dynamic range. On the other end of the spectrum, windows with much more moderate changes on the edges like the half-sine window have a high dynamic range but a low resolution. At this point, the window function and shift can be chosen according to the requirements of the time series at hand.

3.5. Properties of DFT-LRP

Here, we present the conservation and symmetry properties exhibited by (ST)DFT-LRP, which are inherited from LRP and DFT.

(1) Total relevance conservation. The total relevance in frequency domain equals the total relevance in time domain, i.e. $\sum_k R_k = \sum_n R_n$. This is easily validated by examining

$$\sum_k R_k = \sum_n \underbrace{\sum_k r_k \cos(\frac{2\pi kn}{N} - \varphi_k)}_{x_n} \frac{R_n}{x_n} \,,$$

for DFT-LRP and

$$\sum_{k,m} R_{k,m} = \sum_m \sum_n \underbrace{\sum_k r_{m,k} \cos(\frac{2\pi kn}{N} - \varphi_{m,k}) \cdot W_n^{-1}}_{x_n \cdot I_{n \in m}} \frac{R_n}{x_n}$$
$$= \sum_m \sum_{n \in m} R_n = \sum_n R_n \,,$$

for STDFT-LRP. In particular, due to the rescaling with W^{-1}, this is given for any window choice and overlap.

(2) Relevance conservation in time bins. In time–frequency domain, we might require more fine-grained relevance conservation over time bins in some settings. Here, we want to obtain the total relevance over time interval $n \in m$ in time domain when we sum over frequency bins in time bin m in the time–frequency domain, i.e. $\sum_k R_{k,m} = \sum_{n \in m} R_n$. In the case of overlapping windows with shift $D < H$, the signal is stretched in time domain and there is no clear assignment between relevance in time bins in time and time–frequency domain. Thus, we can assign this property to STDFT-LRP only when $D = H$. This singles out the rectangular window, because windows with smoothed edges and no overlap suffer from information loss at the edges where the signal receives weights close to zero when $D = H$. To summarize, when we require fine-grained relevance conservation over time bins, such as in Section 4.2, we need to restrict to the rectangular window with shift $D = H$.

(3) Symmetry. We only consider real signals $x_n \in \mathbb{R}$, for which the spectrum is even symmetric $y_k = y_{-k \bmod N}$. It is apparent from Eq. (12) that this symmetry also holds true for R_k. This property can be leveraged for reduced computational cost, as one needs to evaluate R_k only for $k \in [0, N/2 + 1]$.

4. Results

First, we empirically evaluate our method on a synthetic dataset with ground-truth annotations in Section 4.2 and on a real-world dataset via feature flipping in Section 4.3. Next, we demonstrate the utility of our approach in two use-cases: We compare the strategies of two audio classifiers trained on different input domains in Section 4.4 and show how DFT-LRP reveals Clever Hans strategies of audio and ECG-classifiers in Section 4.5.

[1] In fact, this is inevitable for the DFT of any signal, not just for STDFT, because a discrete and finite signal is always subject to sampling and windowing.

4.1. Datasets and models

Here we present all datasets and models which are used in the following sections.

Synthetic Data. The signal is a simple superposition M sinusoids,

$$x_n = \sum_j^M a_j \cdot \sin(\frac{2\pi n}{N k_j} + \varphi_j) + \sigma y$$

with amplitude a_j, frequency $2\pi/Nk_j$, random phase φ_j, and additive Gaussian noise $y \sim \mathcal{N}(0,1)$ with strength σ. We choose the signal length as $N = 2560$ and restrict to $0 < k_j < 60$. The task is to detect a combination of one to four frequencies from the set $k^* = \{k_1, k_2, k_3, k_4\}$ in the time representation of the signal. Here, each combination of $\{k_i\}$ from the powerset of k^*, i.e. $\{\}, \{k_1\}, \dots, \{k_1, k_2\}, \dots, \{k_1, k_2, k_3\}, \dots, \{k_1, k_2, k_3, k_4\}$, corresponds to a label. We choose $k^* = \{5, 16, 32, 53\}$ for the set. We train a simple Multi-Layer-Perceptron model with two hidden layers and ReLU activation on 10^4 samples on a *baseline* task with noise strength $\sigma = 0.01$ and a *noisy* task with $\sigma = 0.8$. The model reaches an accuracy of 99.9% and 99.7% on the test set with 1000 samples, respectively.

AudioMNIST. This dataset by [16] consists of 3000 recordings of spoken digits (0–9) in English with 50 repetitions of each digit by each of 60 speakers. Besides the actual spoken digit, the dataset contains meta-information such as biological sex and accent of all speakers. Following [16], we down-sample recordings from 16 kHz to 8 kHz and zero-pad them, such that each recording is represented by a vector of length 8000. We train the same 1d CNN classifier as in [16] on the raw waveforms and achieve an accuracy of 92% and 96% on the sex and digit classification task, respectively.

MIT-BIH. The ECG arrhythmia database by [35] consists of ECG recordings from 47 subjects, with a sampling rate of 360 Hz. The preprocessing follows [36], who isolated the ECG lead II data, split and padded the data into single beats with a fixed length of 1500 ms at a sampling rate of 125 Hz. At least two cardiologists have annotated each beat and grouped the annotations into five beat categories: (1) normal beats etc., (2) supraventricular premature beats, etc, (3) premature ventricular contraction and ventricular escape, (4) fusion of ventricular and normal, and (5) paced/ unclassifiable, etc. in accordance with the AAMI EC57 standard. The model under consideration is a 1d CNN with three convolutional layers and a classification head consisting of three dense layers, all with ReLU activations, that classifies an ECG signal in time domain into five beat categories with an accuracy of 95.3%. For further insights into the classification performance, we show confusion matrices for the AudioMNIST and MIT-BIH models in the supplementary material.

4.2. Evaluation on synthetic data with ground truth

We evaluate (ST)DFT-LRP in a setting where ground truth relevance attributions in frequency and time–frequency domain are available for a simple task on synthetic data. First, we quantitatively evaluate how well (ST)DFT-LRP explanations and explanations of attribution methods equipped with a virtual DFT layer align with the ground truth. Second, we qualitatively evaluate the interpretability of explanations in time versus frequency domain.

4.2.1. Quantitative evaluation

We base this evaluation on explanations of the frequency detection models trained on the *baseline* and *noisy* task for the respective test split of the synthetic dataset described in Section 4.1. We compute LRP relevances using the ϵ-rule in time domain and apply (ST)DFT-LRP according to Eq. (12) and Eq. (15) to transform them to frequency and time–frequency domain. We compare to other local attribution methods, namely Sensitivity [37], Gradient times Input (G × I) (e.g. [38]), and Integrated Gradient (IG) [2] which we equip with a virtual inspection layer. To this end, we attach an inverse (ST)DFT layer according to Eq. (7) and Eq. (14) to the input layer like in Fig. 1. Then, we perform the attribution method for the new model that takes the signal in frequency (time–frequency) domain (split into real and imaginary part) as input. For IG, we use $x_n = y_k = z_{mk} = 0$ as a baseline. For all attribution methods we make use of implementations readily available via the zennit package [39]. Given the simplicity of the task and the high test set accuracy of close to 100%, we can assume that ground-truth explanations correspond to attributing positive relevance only to the subset of k^* related to the respective label. To quantitatively evaluate how well the explanations align with this ground truth, we define a *relevance localization score* λ,

$$\lambda = \sum_{k \in k^*} R_k / \sum_k R_k I_{R_k > 0}\,, \tag{16}$$

which measures the ratio of the positive relevance that is attributed to the informative features $\{k_i\}$. A high λ corresponds to accurate relevances in frequency or time–frequency domain.

Table 1
Positive relevance localization λ of explanations in the frequency and time–frequency domain for synthetic frequency detection tasks with low (baseline) and high (noisy) additive noise. Relevances from LRP, IG, and G × I all show equal localization scores. The error is below 0.01 in all cases.

Task	Baseline		Noisy	
method	{LRP, IG, G × I}	Sens.	{LRP, IG, G × I}	Sens.
λ_{DFT}	0.94	0.51	0.80	0.46
$\lambda_{STDFT-N/10}$	0.36	0.51	0.29	0.46
$\lambda_{STDFT-N/4}$	0.67	0.51	0.55	0.46
$\lambda_{STDFT-N/2}$	0.80	0.51	0.67	0.46

In Table 1, we show the mean λ_{DFT} and λ_{STDFT} for heatmaps in frequency and time–frequency domain across 1000 test set samples. In time–frequency domain, we evaluate λ for STDFTs with window widths $D = N/10, N/4, N/2$.

Since the simple MLP model with only two hidden layers and ReLU activation is only slightly non-linear, and LRP, G × I, and IG reduce to the same attribution for a linear model [6], the attributions among these methods are very similar, resulting in equal relevance localization scores λ.

For the *baseline* task, we observe almost perfect relevance localization λ_{DFT} in the frequency domain for DFT-LRP and equivalent methods. For the *noisy* task, λ_{DFT} reduces to 0.80. When cutting the sum in Eq. (16) at the maximum frequency of the signal ($k = 60$), this gap disappears, revealing that DFT-LRP and equivalent methods mix a small part of the total relevance with noise.

Further, we observe that λ_{STDFT} is generally lower than λ_{DFT} and increases with higher window width D for STDFT-LRP and equivalent. This is due to the time–frequency resolution trade-off inherent to STDFT. The higher D, the higher the frequency resolution, i.e. the ability to resolve similar frequencies, and the lower the time resolution. Since the signal is stationary, λ_{STDFT} is affected only by the increase in frequency resolution, not the decrease in time resolution.

Lastly, Sensitivity mostly shows much lower localization scores λ than LRP and other methods. This is can be explained by Sensitivity highlighting local effects instead of overall feature contributions. Only for $\lambda_{STDFT-N/10}$, Sensitivity has the highest score. This is because Sensitivity solely relies on the gradient to determine attribution scores, and does not take the signal itself into account. Furthermore, the weights in the inverse STDFT layer are the same for each window shift as for the DFT layer. Thus, λ_{DFT} and λ_{STDFT} are equal and STDFT-Sensitivity does not suffer from the limited frequency resolution.

In summary, DFT-LRP is superior to Sensitivity and equivalent to IG and G × I for this simple task. DFT-LRP and STDFT-LRP can reliably recover ground-truth explanations, up to the slight mixing of relevance with noise and limitations inherent to STDFT, i.e. the time–frequency resolution trade-off.

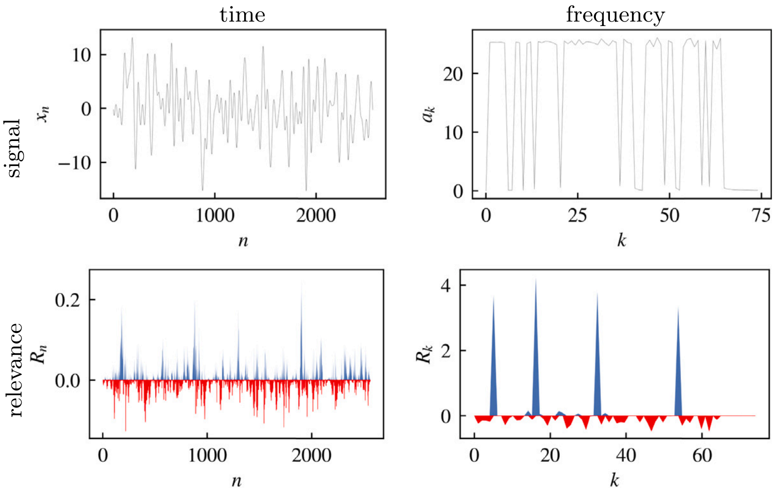

Fig. 2. Time and frequency signal (first row) and relevances (second row) for the frequency detection task on the synthetic data. Relevances are based on the LRP-ϵ rule and (ST)DFT-LRP. In the time domain relevance tends to be distributed quite uniformly over the signal, but is clearly localized on the frequencies to detect, i.e. $\{k_1, k_2, k_3, k, 4\}$, in frequency domain, revealing the classifier strategy.

4.2.2. Qualitative evaluation

We briefly demonstrate the advantage of relevance propagation to frequency domain. We show (ST)DFT-LRP relevances in time and frequency domain in Fig. 2 for the baseline task and a signal corresponding to the label $\{5, 16, 32, 53\}$. In the time domain, relevance tends to be distributed quite uniformly over the entire signal. In contrast, in frequency domain, relevance is clearly localized on the ground-truth informative frequencies k^* frequency domain. We argue, that the classifier strategy is only comprehensible after relevance propagation to frequency domain.

4.3. Evaluation on real-world data

For real-world audio data, we (1) test which feature domain — time, frequency or time-frequency — is the most *informative* to the model across different XAI methods, and (2) compare the *faithfulness* of different XAI methods in each feature domain. Specifically, we measure the complexity of heatmaps [40] to assess informativeness and perform feature flipping experiments to quantify faithfulness. The latter are analogous to Pixel-flipping, which is often deployed to benchmark XAI methods in computer vision [31].

We base our evaluation on the digit classification model trained on the AudioMNIST dataset. Again, we consider LRP, IG, G × I, and Sensitivity. We compute LRP relevances in time domain by applying the z^+-rule [41] to convolutional and the ϵ-rule [1] to dense layers. Then, we apply ST(DFT)-LRP via Eq. (7) and Eq. (14) to propagate relevances R_n from time domain x_n to frequency y_k and time–frequency $v_{m,k}$ domain. Relevance scores for Sensitivity, G × I, and IG in all domains are computed like in the previous section, i.e. by attaching a virtual inspection layer to the original input layer that performs an inverse Fourier Transform. We choose a rectangular window of size $H = N/10$ and hop length $D = H$ for the STDFT.[2]

First, we compute the Shannon entropy of the heatmaps to measure their complexity. In the most *informative* domain, relevance will be concentrated on only a few features that are sufficient for the prediction, which results in heatmaps with low complexity. Second, we perform feature flipping in time, frequency, and time–frequency domain. Here, we either flip features to a zero baseline in order of their relevance scores (smallest destroying feature, SDF) or start with an empty signal and add the most relevant features first (smallest constructing feature, SCF). After each feature modification, i.e. addition or deletion, we measure the model's output probability for the true class. To flip a feature in frequency or time–frequency domain, we set the amplitude of $y_k, z_{k,m}$ to zero for $k = 0, \dots N/2$, accounting for the symmetry of the signal in these domains. In time domain, we set the time point x_n to zero. For comparability of the feature flipping curve across domains, we scale them to the ratio of modified features, where 100% correspond to N features modified in time domain, $N/2$ features in frequency domain and $N/H \cdot N/2$ features in time–frequency domain. To reduce the results to a scalar score, we compute the area under the curve (AUC) of the feature flipping curves. A relevance attribution method that is *faithful* to the model reflects in a steep descent or ascent in true class probabilities after flipping or adding the truthfully as most important annotated features, respectively.

In Fig. 3 we show the true class probability against the ratio of deleted and added features, i.e. for SDF and SCF respectively, for all attribution methods and input domains. We list the corresponding AUC scores of the feature flipping curves and the mean complexity over all heatmaps in Table 2. For a qualitative comparison of explanations across feature domains, we show LRP heatmaps for each domain in Fig. 4 for a randomly selected sample correctly classified as a seven. We show relevances for additional samples of correctly and incorrectly digits in the supplementary material.

Now, we turn to the question of which feature domain is the most *informative* to the model. To this end, we compare complexity scores across domains for each XAI method. For each method except Sensitivity, the frequency domain shows the lowest complexity, i.e. is most informative with respect to the model, followed by time and time–frequency domain. However, the visual impression of the heatmaps in Fig. 3 contradicts this ranking, as relevance shows distinct peaks at certain frequencies in frequency *and* time–frequency domain, but is distributed rather uniformly in time domain. Thus, we suspect that the higher complexity of time–frequency heatmaps compared to time domain might result from the fringes in the spectrum, produced by the sharp edges of the rectangular window. Again, the complexity of Sensitivity heatmaps is the same for frequency and time–frequency features because the method only takes into account the gradient, i.e. the weights of the Fourier Transform, as already described in Section 4.2.1. At this point we emphasize that we cannot accurately compare informativeness via feature flipping. This is because probability decrease/increase might not only result from deducting/adding information, but also from off-manifold evaluation of the model on samples with unknown artifacts that result from setting features to the

[2] We choose a rectangular window, so we can choose the hop size to equal the window width, in order to not introduce artifacts by flipping a time–frequency feature that overlaps with another feature in time.

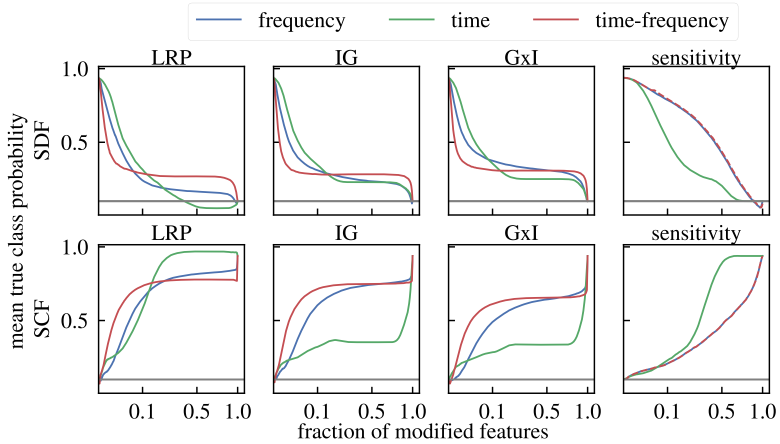

Fig. 3. Evaluation of LRP, IG, G × I and Sensitivity attributions for a model trained on the AudioMNIST digit classification task via feature flipping: Mean true class probability after feature deletion (SDF) and feature addition (SCF) in time (x_n), frequency (y_k) and time–frequency ($v_{m,k}$) domain. The horizontal axis is square root scaled. The gray horizontal line corresponds to the chance level, i.e. a probability of 0.1.

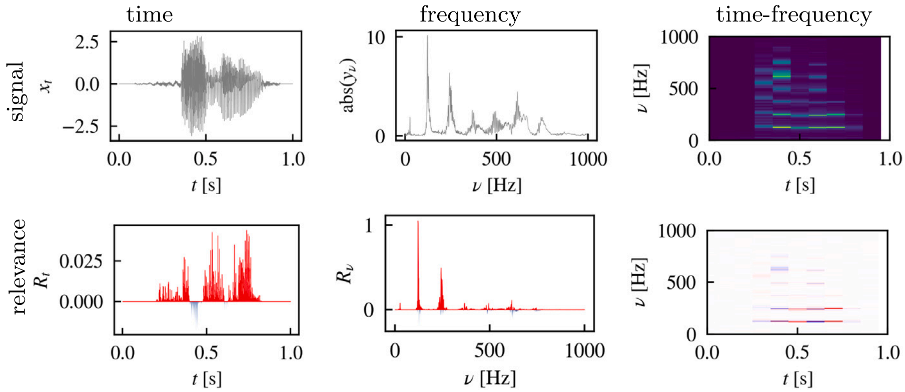

Fig. 4. Time, time–frequency, and frequency signal (first row) and relevances (second row) for the digit detection task on the AudioMNIST data. The signal corresponds to a spoken seven. Relevances are based on the LRP-z^+-rule for convolutional and LRP-ϵ rule for dense layers and (ST)DFT-LRP. In the time domain relevance tends to be distributed uniformly over the signal but is more localized in frequency and time–frequency domain.

Table 2
Evaluation of LRP, IG, G × I and sensitivity relevances for a model trained on the AudioMNIST digit classification task: AUC of feature flipping curves for adding (SCF) and deleting (SDF) features in order of their relevance, and complexity scores. The method with the globally highest faithfulness per domain, i.e. highest (↑) AUC for SCF and lowest (↓) AUC for SDF, is marked in bold. Further, the domain with the lowest complexity is marked in bold for each attribution method. The AUC scores correspond to the feature flipping curves in Fig. 3, where the horizontal axis is square root scaled.

Method	Domain	Faithfulness across methods		Informativeness across domains
		SCF (↑)	SDF (↓)	Complexity (↓)
LRP	Frequency	**0.66**	**0.28**	**6.00**
	Time	**0.73**	**0.28**	6.69
	Time-freq.	**0.69**	**0.31**	7.26
IG	Frequency	0.60	0.32	**6.97**
	Time	0.34	0.35	7.14
	Time-freq.	0.67	0.31	8.52
G × I	Frequency	0.51	0.38	**7.03**
	Time	0.32	0.37	7.19
	Time-freq.	0.58	0.33	8.58
Sensitivity	Frequency	0.36	0.59	7.66
	Time	0.51	0.41	**6.13**
	Time-freq.	0.36	0.59	9.96

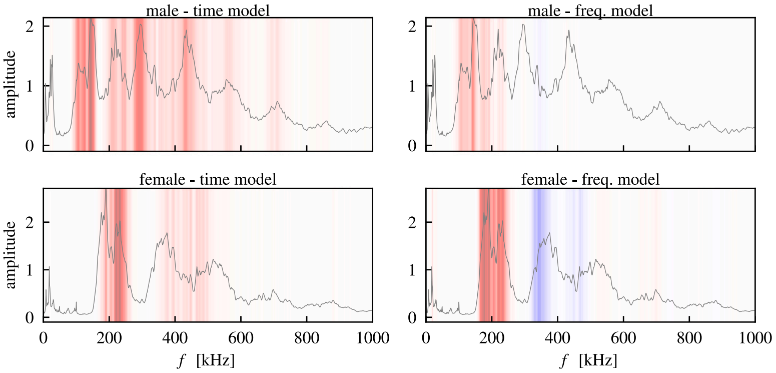

Fig. 5. We evaluate LRP relevances for models trained on the AudioMNIST sex classification task. We show the mean spectrum and mean DFT-LRP relevances across the test dataset for male/female samples for the time and frequency model. The time model uses fundamental frequency and subsequent harmonics as features while the frequency model focuses only on the fundamental frequency.

baseline. Importantly, this effect might differ between feature domains. Still, if we focus on the first part of the SDF and SCF feature flipping curves in Fig. 3, where the least artifacts exist, the initial steep decrease/increase in time–frequency and frequency domain supports our findings. We would like to stress, that the ranking of informativeness of input domains is dependent on the quality of the time series to classify, e.g. for time series with time-localized characteristics, time domain might be more informative to the model than frequency domain, which in turn will manifest in the expressions of the attribution maps. Our method novelly enables the comparison between the domains and allows analyzing the model strategy in the most interpretable domain.

Lastly, we compare the faithfulness between XAI methods in each domain in terms of the feature flipping results. For all domains, ((ST)DFT)-LRP delivers the most *faithful* relevance heatmaps, followed by IG, G × I, and Sensitivity, according to both, SCF and SDF AUC scores.

4.4. Use case I: Data representations for audio classifiers

The best choice of data representation — e.g. raw waveforms, spectrograms or spectral features — is an important aspect of deep learning-based audio analysis [8]. Previous work benchmarks different representations by measuring classification accuracy [42]. Novelly, DFT-LRP allows for a comparison of the underlying strategies of two audio classifiers trained in the frequency and time domain, which we leverage in this case study to demonstrate the utility of our approach.

Here, we compare the 1d CNN sex audio classifier operating on the raw waveforms of the AudioMNIST dataset (*time model*), to a model of the architecture, but trained on absolute values of the signal in frequency domain (*frequency model*). The frequency model achieves an accuracy of 98% on the sex classification task (the time model has an accuracy of 92%).

To compare the classification strategies of the two models, we show the mean relevance in frequency domain for female and male samples for both models in Fig. 5 across 3000 test set samples (1500 female). The correlation between the relevances of the frequency and time model in the frequency domain is only 0.43 on average, already revealing that the two models have picked up different classification strategies. Before we look into these in more detail, we list the characteristics of female and male voices from the literature: The fundamental frequency of the male voice is between 85–155 Hz for males and 165–255 Hz for females [43], the subsequent harmonics are integer-multiples of this value. To quantify the classification strategy of the two models, we list the frequency bands for which the mean relevance exceeds the 90% percentile. For male samples and the time model, the mean relevance exceeds the 90% percentile for frequency intervals 99–156 Hz, 276–307 Hz, and 425–438 Hz. For male samples and the frequency model, this is the case for 83–160 Hz, plus for a small number 24 frequencies in the intervals 293–300 Hz, 335–340 Hz, and 423–436 Hz. For the female samples, the analog threshold is exceeded between intervals 182–255 Hz and 392–487 Hz for the time model and between 18–20 Hz (noise) and 157–253 Hz for the frequency model. For comparison, the fundamental frequency of the male voice is between 85–155 Hz for males and 165–255 Hz for females [43], the subsequent harmonics are integer-multiples of this value. In summary, the time model focuses on the fundamental frequency and the first two (one) subsequent harmonics of the male (female) voice, whereas the frequency model considers mostly the fundamental frequency as a relevant feature for male and female samples. Interestingly, for the female samples also low frequencies corresponding to noise are relevant for the frequency model.

4.5. Use case II: DFT-LRP reveals Clever Hans strategies in frequency domain

4.5.1. Artificial noise in audio data

Noise is separated from the signal in frequency domain, but not in time domain. We mimic a scenario which is realistic in various real world audio classification problems. We add noise to one class of the AudioMNIST digit classification task. Likely, noise as a spurious correlation also exists in real-world data due to class-dependent recording techniques or environment. A model that learns to separate classes by spurious correlations is deemed a Clever Hans classifier [9]. Here, we demonstrate, that Clever Hans strategies leveraging noise can only be detected after propagating relevance from time to frequency domain.

To this end, we compare a model trained on the original AudioMNIST digit classification data and a model trained on the modified data, where pink noise was added only to the spoken zeros.

As in the previous section, both models are trained in the time domain and achieve an accuracy of about 94%. We confirm that the model is using the Clever Hans strategy, which we tried to induce by introducing the spurious correlation in the training data, by finding that it classifies 98% of samples with added noise as zero, regardless of the actual digit spoken. Now, we try to infer this behavior from the explanations in Fig. 6, showing LRP relevances of the same sample with label zero for each classifier the in three domains. In the time domain, the only visible difference between the Clever Hans and the regular classifier is that the beginning and end of the signal, where no digit is spoken, is relevant for the decision of the Clever Hans but not for the regular model. Otherwise, relevance is spread rather uniformly over the signal in both cases. The difference between the classifiers only becomes perfectly clear in the time or time–frequency domain. The regular classifier focuses on the fundamental frequencies and subsequent harmonics towards the beginning and end of the spoken digit, whereas the relevance of the Clever Hans classifier is concentrated on frequencies between zero and 50 Hz, which correspond to noise.

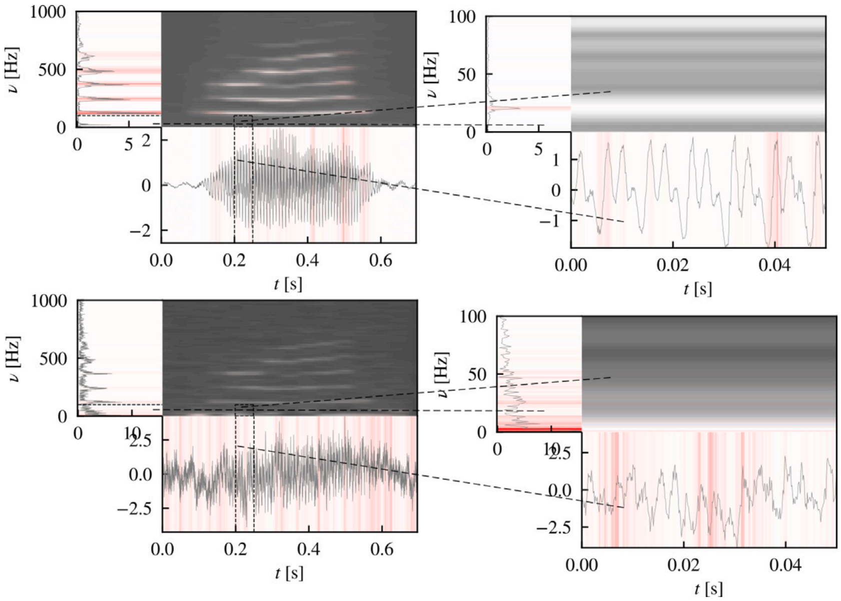

Fig. 6. We compare a model trained on data without additional noise (upper) to a model trained on data with Clever Hans noise (lower), for a sample without and with added pink noise with strength $\sigma = 0.8$. The left column shows the whole signal, while the right column shows a zoom on the part of the signal marked in the left column. The upper and lower row depict the same sample of a spoken zero, but noise was added to the signal in the lower row. During training, noise was added only to samples of a spoken zero to induce a Clever Hans strategy. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

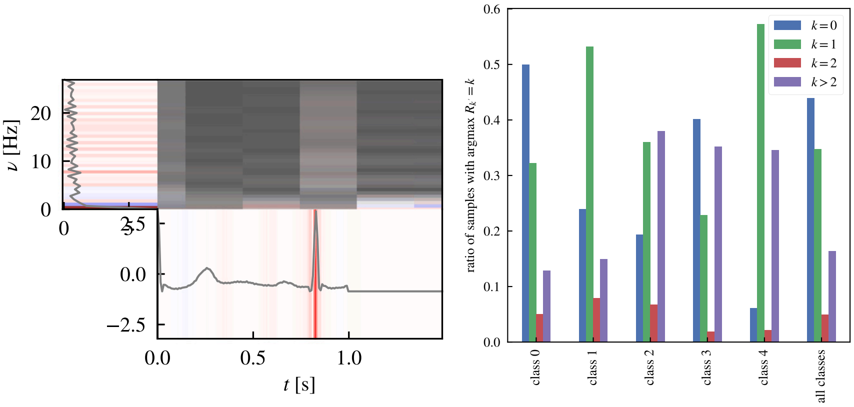

Fig. 7. Left: LRP relevances for an ECG classifier trained on the MIT-BIH dataset in time, frequency and time–frequency domain for a correctly predicted random normal beat. Right: Ratio of samples, for which the maximum relevance $max_{k'} R_{k'}$ lies on the respective frequency component k. Among all classes, the classifier focuses on the mean of the signal ($k = 0$) for 44.3% of the samples. This value is even higher for class 0 (normal beats).

4.5.2. ECG classifier

We now demonstrate how DFT-LRP helps to discover Clever Hans behavior of the ECG classifier described in Section 4.1.

The classification strategy of a trustworthy ECG model should align with the signal characteristics analyzed by cardiologists, such as the amplitude of the QRS complex and ST segment, the duration of segments or peak ratios. For instance, a normal heartbeat that originates at the atrium and traverses the normal conduction path is characterized by a sharp and narrow QRS complex with a broad peak at a frequency of 8 Hz [44]. To assess the classification strategy of the ECG classifier we show ((ST)DFT)-LRP relevances in time, frequency and time–frequency domain for a normal beat in Fig. 7 (left). Additional samples of correctly and incorrectly classified beats are displayed in the supplementary material. For visual clarity, we depict only frequency components up to $k = 20$, where the majority of relevance is located. For the sample in Fig. 7, relevance in time and time–frequency domain suggests that the model focuses on the QRS complex, in particular on frequencies around $k = 10$. However, the relevance in frequency domain reveals that a large part of the total relevance is attributed to $k = 0$ which corresponds to the mean of the signal. In Fig. 7 (right) we show the ratio of samples in the test set for which the maximum of R_k lies on frequencies $k = 0, 1, 2$ or $k > 2$ for each class. We observe that the model focuses on the mean of the signal for a majority of samples for all classes (44.3%). This tendency is even more pronounced for class 0 (normal beats). Based on these observations, we conclude that the classifier learned to focus on the mean of the signal instead of relying on signal characteristics considered by cardiologists. This suggests that the model relies on a Clever Hans strategy. Only by transforming the relevances from the time to the frequency domain using DFT-LRP, can we unveil this behavior in the model's explanations.

5. Conclusion

We have put forward virtual inspection layers that perform an identity loop via an interpretable representation to facilitate comprehensible explanations. We have specialized in DFT for the virtual inspection layer and in LRP for the XAI method. In this way, we have demonstrated how to extend LRP to provide interpretable explanations for time series classifiers in both the frequency and time–frequency domain. We have established the validity of our approach through testing on a ground-truth test bed and on real audio data. Further, we demonstrated the benefits of our methods bring in real-world scenarios, such as the analysis of input representations and detection of Clever Hans behavior. We envision applications of DFT-LRP in domains where interpreting the time domain representation of the signal is particularly challenging, such as audio, sensor data or electronic health records.

So far, we have focused on univariate time series. While we can apply our method to multi-variate time series straightforwardly by applying DFT-LRP to each channel separately, this approach is limited in the sense that it cannot reveal relevant interactions between channels. Understanding such interactions could be an important aspect, e.g. for ECG classifiers acting on multiple channels.

In future research, it would be interesting to explore the use of other invertible transformations such as PCA as virtual inspection layers at the input or at hidden feature layers of the model. Even non-linear but approximately invertible transformations, e.g. an autoencoder that learned an interpretable representation, could serve as virtual inspection layer.

CRediT authorship contribution statement

Johanna Vielhaben: Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Sebastian Lapuschkin:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Grégoire Montavon:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Wojciech Samek:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Federal Ministry of Education and Research (BMBF), Germany as grants [BIFOLD (01IS18025A, 01IS180 371I), SyReal (01IS21069B)]; the European Union's Horizon Europe research and innovation programme (EU Horizon Europe) as grant TEMA (101093003); the European Union's Horizon 2020 research and innovation programme (EU Horizon 2020) as grant iToBoS (965221); the state of Berlin within the innovation support programme ProFIT (IBB) as grant BerDiBa (10174498); and the German Research Foundation as grant DFG KI-FOR 5363.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patcog.2024.110309.

References

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015) e0130140, http://dx.doi.org/10.1371/journal.pone.0130140.

[2] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, in: ICML, 2017, pp. 3319–3328, http://dx.doi.org/10.1145/1993574.1993601.

[3] M.T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD '16, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450342322, 2016, pp. 1135–1144, http://dx.doi.org/10.1145/2939672.2939778.

[4] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774, http://dx.doi.org/10.5555/3295222.3295230.

[5] J.V. Jeyakumar, J. Noor, Y.-H. Cheng, L. Garcia, M. Srivastava, How can I explain this to you? An empirical study of deep neural network explanation methods, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 4211–4222, http://dx.doi.org/10.5555/3495724.3496078.

[6] W. Samek, G. Montavon, S. Lapuschkin, C.J. Anders, K.-R. Müller, Explaining deep neural networks and beyond: A review of methods and applications, Proc. IEEE 109 (3) (2021) 247–278, http://dx.doi.org/10.1109/JPROC.2021.3060483.

[7] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, N. Díaz-Rodríguez, Explainable artificial intelligence (XAI) on TimeSeries data: A survey, 2021, http://dx.doi.org/10.48550/ARXIV.2104.00950, arXiv preprint 2104.00950.

[8] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, T. Sainath, Deep learning for audio signal processing, IEEE J. Sel. Top. Sign. Proces. 13 (2) (2019) 206–219, http://dx.doi.org/10.1109/JSTSP.2019.2908700.

[9] C.J. Anders, L. Weber, D. Neumann, W. Samek, K.-R. Müller, S. Lapuschkin, Finding and removing clever hans: Using explanation methods to debug and improve deep models, Inf. Fusion (ISSN: 1566-2535) 77 (2022) 261–295, http://dx.doi.org/10.1016/j.inffus.2021.07.015.

[10] G. Deshpande, A. Batliner, B.W. Schuller, AI-based human audio processing for COVID-19: A comprehensive overview, Pattern Recognit. (ISSN: 0031-3203) 122 (2022) 108289, http://dx.doi.org/10.1016/j.patcog.2021.108289.

[11] B. García-Martínez, A. Fernández-Caballero, R. Alcaraz, A. Martínez-Rodrigo, Assessment of dispersion patterns for negative stress detection from electroencephalographic signals, Pattern Recognit. (ISSN: 0031-3203) 119 (2021) 108094, http://dx.doi.org/10.1016/j.patcog.2021.108094.

[12] D. Cheng, F. Yang, S. Xiang, J. Liu, Financial time series forecasting with multimodality graph neural network, Pattern Recognit. (ISSN: 0031-3203) 121 (2022) 108218, http://dx.doi.org/10.1016/j.patcog.2021.108218.

[13] A. Theissler, F. Spinnato, U. Schlegel, R. Guidotti, Explainable AI for time series classification: A review, taxonomy and research directions, IEEE Access 10 (2022) 100700–100724, http://dx.doi.org/10.1109/ACCESS.2022.3207765.

[14] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller, Deep learning for time series classification: a review, Data Min. Knowl. Discov. (ISSN: 1573-756X) 33 (4) (2019) 917–963, http://dx.doi.org/10.1007/s10618-019-00619-1.

[15] D. Slijepcevic, F. Horst, B. Horsak, S. Lapuschkin, A.-M. Raberger, A. Kranzl, W. Samek, C. Breiteneder, W.I. Schöllhorn, M. Zeppelzauer, Explaining machine learning models for clinical gait analysis, ACM Trans. Comput. Healthc. 3 (2) (2022) 1–27, http://dx.doi.org/10.1145/3474121.

[16] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, W. Samek, AudioMNIST: Exploring explainable artificial intelligence for audio analysis on a simple benchmark, J. Franklin Inst. B (ISSN: 0016-0032) 361 (1) (2024) 418–428, http://dx.doi.org/10.1016/j.jfranklin.2023.11.038.

[17] N. Strodthoff, P. Wagner, T. Schaeffter, W. Samek, Deep learning for ECG analysis: Benchmarks and insights from PTB-XL, IEEE J. Biomed. Health Inf. 25 (5) (2021) 1519–1528, http://dx.doi.org/10.1109/JBHI.2020.3022989.

[18] I. Sturm, S. Lapuschkin, W. Samek, K.-R. Müller, Interpretable deep neural networks for single-trial EEG classification, J. Neurosci. Methods 274 (2016) 141–145, http://dx.doi.org/10.1016/j.jneumeth.2016.10.008.

[19] N. Strodthoff, C. Strodthoff, Detecting and interpreting myocardial infarction using fully convolutional neural networks, Physiol. Meas. 40 (1) (2019) 015001, http://dx.doi.org/10.1088/1361-6579/aaf34d.

[20] F. Kratzert, M. Herrnegger, D. Klotz, S. Hochreiter, G. Klambauer, NeuralHydrology – interpreting LSTMs in hydrology, in: W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing, Cham, ISBN: 978-3-030-28954-6, 2019, pp. 347–362, http://dx.doi.org/10.1007/978-3-030-28954-6_19.

[21] Z. Wang, W. Yan, T. Oates, Time series classification from scratch with deep neural networks: A strong baseline, in: International Joint Conference on Neural Networks, IJCNN, IEEE, 2017, pp. 1578–1585, http://dx.doi.org/10.1109/IJCNN.2017.7966039.

[22] R. Mochaourab, A. Venkitaraman, I. Samsten, P. Papapetrou, C.R. Rojas, Post hoc explainability for time series classification: Toward a signal processing perspective, IEEE Signal Process. Mag. 39 (4) (2022) 119–129, http://dx.doi.org/10.1109/MSP.2022.3155955.

[23] J. Vielhaben, S. Bluecher, N. Strodthoff, Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees, Trans. Mach. Learn. Res. (ISSN: 2835-8856) (2023).

[24] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin, From attribution maps to human-understandable explanations through concept relevance propagation, Nat. Mach. Intell. 5 (2023) 1006–1019, http://dx.doi.org/10.1038/s42256-023-00711-8.

[25] S. Gautam, M.M.-C. Höhne, S. Hansen, R. Jenssen, M. Kampffmeyer, This looks more like that: Enhancing self-explaining models by prototypical relevance propagation, Pattern Recognit. (ISSN: 0031-3203) 136 (2023) 109172, http://dx.doi.org/10.1016/j.patcog.2022.109172.

[26] D. Mincu, E. Loreaux, S. Hou, S. Baur, I. Protsyuk, M. Seneviratne, A. Mottram, N. Tomasev, A. Karthikesalingam, J. Schrouff, Concept-based model explanations for electronic health records, in: Proceedings of the Conference on Health, Inference, and Learning, ACM, 2021, http://dx.doi.org/10.1145/3450439.3451858.

[27] S.A. Siddiqui, D. Mercier, M. Munir, A.R. Dengel, S. Ahmed, TSViz: Demystification of deep learning models for time-series analysis, IEEE Access 7 (2019) 67027–67040, http://dx.doi.org/10.1109/ACCESS.2019.2912823.

[28] F. Küsters, P. Schichtel, S. Ahmed, A. Dengel, Conceptual explanations of neural network prediction for time series, in: 2020 International Joint Conference on Neural Networks, IJCNN, 2020, pp. 1–6, http://dx.doi.org/10.1109/IJCNN48605.2020.9207341.

[29] R. Guidotti, A. Monreale, F. Spinnato, D. Pedreschi, F. Giannotti, Explaining any time series classifier, in: 2020 IEEE Second International Conference on Cognitive Machine Intelligence, CogMI, IEEE Computer Society, Los Alamitos, CA, USA, 2020, pp. 167–176, http://dx.doi.org/10.1109/CogMI50398.2020.00029.

[30] E. Ates, B. Aksar, V.J. Leung, A.K. Coskun, Counterfactual explanations for multivariate time series, in: 2021 International Conference on Applied Artificial Intelligence, ICAPAI, 2021, pp. 1–8, http://dx.doi.org/10.1109/ICAPAI49758.2021.9462056.

[31] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, Evaluating the visualization of what a deep neural network has learned, IEEE Trans. Neural Netw. Learn. Syst. 28 (11) (2017) 2660–2673, http://dx.doi.org/10.1109/TNNLS.2016.259982.

[32] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D.A. Keim, Towards a rigorous evaluation of XAI methods on time series, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop, ICCVW, 2019, pp. 4197–4201, http://dx.doi.org/10.1109/ICCVW.2019.00516.

[33] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, S. Lapuschkin, Towards best practice in explaining neural network decisions with LRP, in: Proceedings of the IEEE International Joint Conference on Neural Networks, IJCNN, 2020, pp. 1–7, http://dx.doi.org/10.1109/IJCNN48605.2020.9206975.

[34] J.B. Allen, L.R. Rabiner, A unified approach to short-time Fourier analysis and synthesis, Proc. IEEE 65 (11) (1977) 1558–1564, http://dx.doi.org/10.1109/PROC.1977.10770.

[35] G. Moody, R. Mark, The impact of the MIT-bih arrhythmia database, IEEE Eng. Med. Biol. Mag. 20 (3) (2001) 45–50, http://dx.doi.org/10.1109/51.932724.

[36] M. Kachuee, S. Fazeli, M. Sarrafzadeh, ECG heartbeat classification: A deep transferable representation, in: 2018 IEEE International Conference on Healthcare Informatics, ICHI, 2018, pp. 443–444, http://dx.doi.org/10.1109/ICHI.2018.00092.

[37] N.J. Morch, U. Kjems, L.K. Hansen, C. Svarer, I. Law, B. Lautrup, S. Strother, K. Rehm, Visualization of neural networks using saliency maps, in: Proceedings of ICNN'95-International Conference on Neural Networks, Vol. 4, IEEE, 1995, pp. 2085–2090, http://dx.doi.org/10.1109/ICNN.1995.488997.

[38] M. Ancona, E. Ceolini, C. Öztireli, M.H. Gross, Towards better understanding of gradient-based attribution methods for deep neural networks, in: International Conference on Learning Representations, 2017.

[39] C.J. Anders, D. Neumann, W. Samek, K.-R. Müller, S. Lapuschkin, Software for dataset-wide XAI: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy, 2021, http://dx.doi.org/10.48550/arXiv.2106.13200, arXiv preprint 2106.13200. abs/2106.13200.

[40] A. Hedström, L. Weber, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M.M.C. Höhne, Quantus: An explainable AI toolkit for responsible evaluation of neural network explanation, J. Mach. Learn. Res. 24 (34) (2023) 1–11.

[41] G. Montavon, S. Bach, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognit. 65 (2017) 211–222, http://dx.doi.org/10.1016/j.patcog.2016.11.008.

[42] L. Hertel, H. Phan, A. Mertins, Comparing time and frequency domain for audio event recognition using deep learning, in: 2016 International Joint Conference on Neural Networks, IJCNN, 2016, pp. 3407–3411, http://dx.doi.org/10.1109/IJCNN.2016.7727635.

[43] J.L. Fitch, A. Holbrook, Modal vocal fundamental frequency of young adults, Arch. Otolaryngol. 92 (4) (1970) 379–382, http://dx.doi.org/10.1001/archotol.1970.04310040067012.

[44] K. Minami, H. Nakajima, T. Toyoshima, Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network, IEEE Trans. Biomed. Eng. 46 (2) (1999) 179–185, http://dx.doi.org/10.1109/10.740880.