POST-HOC SALIENCY METHODS FAIL TO CAPTURE LATENT FEATURE IMPORTANCE IN TIME SERIES DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Saliency methods provide visual explainability for deep image processing models by highlighting informative regions in the input images based on feature-wise (pixels) importance scores. These methods have been adopted to the time series domain, aiming to highlight important temporal regions in a sequence. This paper identifies, for the first time, the systematic failure of such methods in the time series domain when underlying patterns (e.g., dominant frequency or trend) are based on latent information rather than temporal regions. The latent feature importance postulation is highly relevant for the medical domain as many medical signals, such as EEG signals or sensor data for gate analysis, are commonly assumed to be related to the frequency domain. To the best of our knowledge, no existing post-hoc explainability method can highlight influential latent information for a classification problem. Hence, in this paper, we frame and analyze the problem of latent feature saliency detection. We first assess the explainability quality of multiple state-of-the-art saliency methods (Integrated Gradients, DeepLift, Kernel SHAP, Lime) on top of various classification methods (LSTM, CNN, LSTM and CNN trained via saliency guided training) using simulated time series data with underlying temporal or latent space patterns. In conclusion, we identify that Integrated Gradients and DeepLift, if redesigned, could be potential candidates for latent saliency scores.

1 INTRODUCTION

Saliency methods aim to explain the predictions of deep learning models by highlighting important input features. These methods often assign scores to individual inputs (Guidotti et al., 2018; Ismail et al., 2020), collectively resulting in the detection of class-distinctive patterns. For image data this means assigning scores to positional information, namely pixels. Such a strategy suits image data, as the label is often associated with specific input regions. Recently, image saliency methods have been adopted for time series data Loeffler et al. (2022); Schlegel et al. (2020). They similarly assign importance scores to the pixel counterparts, namely "time points". These methods suit the time series problem when a temporal pattern is indicative of the class. In some time series problems, however, the label may depend on the latent features such as dominant frequency, state-space model parameters, or the overall trend of a non-stationary time series. In these cases, even though the classifier might successfully capture the latent space, the positional scores extracted from the classifier will not directly explain the importance of the underlying latent features. Hence, the generated saliency maps will not be directly interpretable and thus fail to fulfill their purpose.

The goal of this paper is to introduce, formulate and analyze the problem of latent feature saliency in deep time series classification problems, focusing on the fundamental Fourier series latent model. By extension, our study is replicable for other latent models. We summarize our main contributions below:

1. We draw attention to the problem of latent feature saliency detection in time series data. We formulate the shapelet- vs. latent-based pattern in time series classification and propose a definition for an ideal latent feature saliency method (Section 2).
2. We provide a comprehensive study of popular time series saliency methods including Integrated Gradients, DeepLift, Kernel SHAP and Lime (Section3,Section 4) on top of multiple

classification methods (LSTM, CNN, LSTM and CNN trained via saliency guided training).

3. We identify effective methods that can be extended to potentially tackle the problem of latent space saliency (Section 5).

2 Problem Formulation

Let $D = (X, Y)$ with a univariate time series $X \in \mathcal{X}$ and the binary label $Y \in \{0, 1\}$ formulate a time series classification data set. Furthermore, let the mapping $f_{XY} : \mathcal{X} \mapsto \{0, 1\}$ represent a deep learning-based classifier. In *latent-representation learning*, we assume a latent space $\mathcal{Z}$, a mapping from feature to latent space $f_{XZ} : \mathcal{X} \mapsto \mathcal{Z}$ and a latent space to label mapping $f_{ZY} : \mathcal{Z} \mapsto \{0, 1\}$, such that the classifier f_{XY} can be learned via the feature-to-latent and latent-to-label mappings.This view has been adopted by several time series classifiers such as hidden Markov models (HMM) and recurrent neural networks (RNN). The learned latent representation, exhibits properties shown to be significant in terms of explainability Mikolov et al. (2013); Charte et al. (2020). Instead of estimating f_{XZ} as a black-box model, a parametric latent model (such as Fourier series models, state space models, linear and switching dynamical systems, or additive and multiplicative models) can be estimated via a neural network. These models are motivated by prior knowledge about the underlying data generation mechanism; thus, their parameters often are interpretable. A saliency method applied to this solution assigns scores to latent features in the $\mathcal{Z}$ space. In contrast, methods used for the black-box models usually lack explainability for the latent features.

The latent space assumption is relevant in many time series problems. Sound signals are often differentiated by amplitude and frequency; thus, the decision process behind audio classification is likely to be better explained by the Fourier latent space than by spatial importance scores. Vibration signal classification, as in earthquake or production line failure prediction, is likely to depend on frequency or amplitude as well. Financial time series classification often revolves around modeling trend and seasonality of the time series. Many signals in the medical domain, such as EEG signals, sensor data from wearable technologies for gait analysis for neurological disease progression, pain recognition, or medication level adjustment, are further strongly related to amplitude and frequency. These examples show that achieving time series explainability is heavily related to latent space assumptions.

2.1 Latent features vs. shapelets

Ye & Keogh (2009) define *shapelets* as variable-length subsequences of time series which are maximally representative of a class. We define a feature-to-shapelet mapping $f_{XS} : \mathcal{X} \mapsto [0, 1]^k$. Samples in $\mathcal{S}$ are normalized score vectors, determining which shapelet appears in a sample. Subsequently, shapelet-based classifiers predict the label based on an existing pattern in the time domain. These models are coordinated with saliency methods, which in this case are visually explainable since time points are directly expressive of both saliency scores and shapelets. The presence of informative shapelets does not contradict the assumption of a latent model. On the contrary, shapelets may appear as a proxy for latent information (see Figure 2). Nevertheless, from the explainability point of view, there is a notable difference between latent features and shapelets. As an example, a label correlated with the damping ratio of a vibration signal can be potentially predicted by shapelet-based classifiers; however, a conventional saliency method, applied to this problem, will only highlight a proxy of the informative latent feature, namely the existing fluctuations and oscillations of the time series. In conclusion, time series classification problems may be characterized by class differences in features which belong to the time domain as shapelets or to a latent domain. Current saliency methods can provide explainability for shapelets, but not directly for latent models.

2.2 Defining a desirable saliency method for time series

Figure 1 illustrates the setup of a time series classification problem with multiple possible intermediate latent spaces, enumerated with i, and denoted as $\mathcal{Z}^{(i)}$. A time series $X \in \mathcal{X}$ can be mapped to $\mathcal{Z}^{(i)}$ by the i-th chosen latent model $f_{XZ}^{(i)}$. Without loss of generality, we assume that there is only one latent feature Z^* which provides the best explanation for the classification task. The latent space that contains Z^* is denoted as $\mathcal{Z}^{(*)}$.

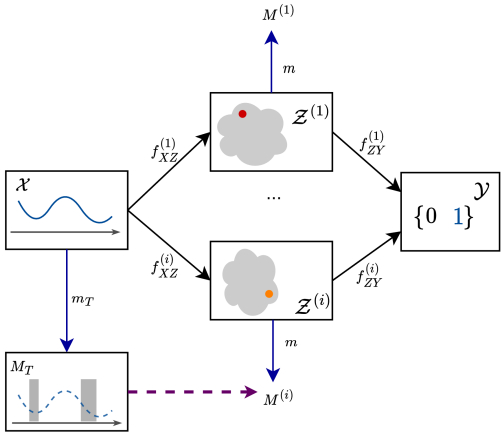

Figure 1: Time series classification schematic over the space $\mathcal{X} \times \mathcal{Y}$ with latent space representations $Z^{(i)}$, associated with saliency function $m(Z^{(i)})$ and resulting saliency map $M^{(i)}$. Current methods m_T measure saliency of the feature space, yielding the map M_T.

We define a saliency method as "reliable" if it assigns the highest score to Z^* above all other features throughout all latent spaces. To formulate the reliability definition, we consider a *latent-aware* saliency method $m : \mathcal{Z}^{(i)} \mapsto \mathbb{R}_+^{|Z^{(i)}|}$ which produces a saliency map $M^{(i)}$ for $\mathcal{Z}^{(i)}$. The reliability condition is then formulated as

$$\forall i \neq *, \quad \max M^{(*)} > \max M^{(i)}.$$

Note that during implementation, we have to define the possible set of latent models manually.

The fundamental problem of existing saliency methods is that they only estimate the saliency map for the time domain and therefore lack appropriate output for features in other domains. Hence none of the existing saliency methods meet the criteria for reliability. However, we argue that there might exist some *promising failing methods*, which requires only minor adjustments to serve as desired saliency methods for time series. We define a saliency method $m_T : \mathcal{X} \mapsto \mathbb{R}_+^{|X|}$ as promising if the produced map $M_T \in \mathbb{R}_+^T$ bears enough information to infer $M^{(i)}, \forall i$ (possibly via a simple mapping function, depicted as a purple arrow in Figure 1). In other words, m_T can capture information about latent saliency, even though it cannot directly explain it. In this case, an extension of the promising method, representing the mapping from M_T to $M^{(*)}$, establishes a desired latent saliency method.

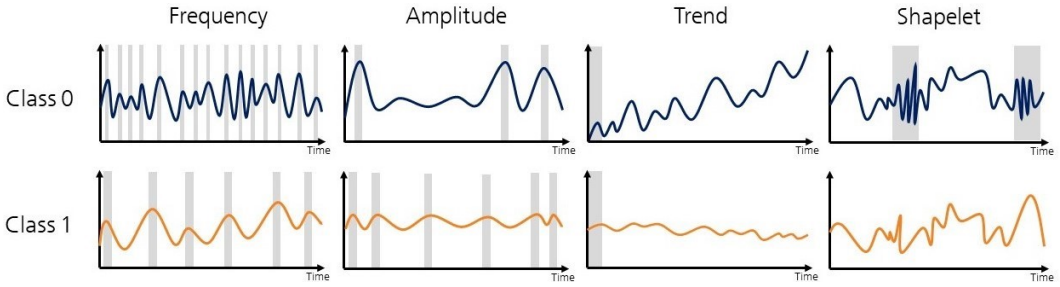

Figure 2: Toy examples of multiple label-making scenarios. Influential time steps (regions with high saliency scores) are shaded in grey for frequency (peaks), amplitude (highest peaks), trend (a window enough for inferring about the trend), and shapelet (presence of the informative pattern).

Figure 2 schematically depicts the output of a good failing method when the label is associated with either the frequency or amplitude of a Fourier model, the trend of an additive model, or shapelets. In particular, highlighted regions are sufficient to infer the latent parameter (or equally shapelet). Putting the experiment into practice, Figure 3 presents heat maps of importance scores resulting from two exemplary failing methods.

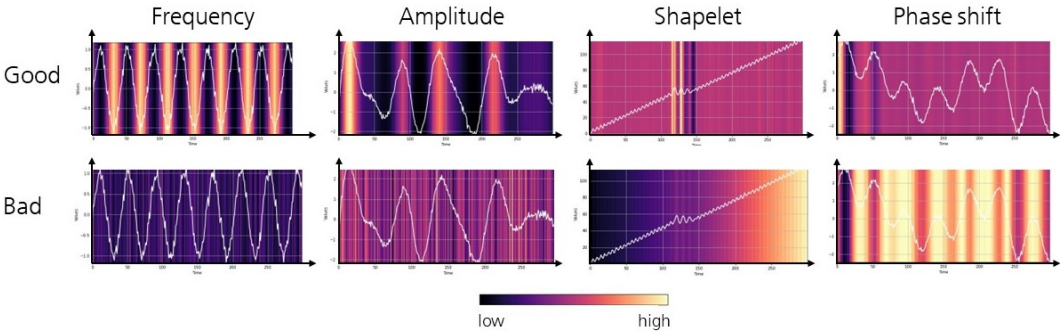

Figure 3: Examples of well-performing explainability methods (top row) providing to some extend interpretable explanations and completely uninterpretable saliency results (bottom row).

3 EXPERIMENTAL FRAMEWORK

As a preliminary step for presenting the results of the empirical study, this section introduces the examined time series saliency methods and data sets and implementation details.

Our study focuses on *post-hoc* saliency methods designed to explain single classification instances of trained models. Here, we investigate the following state of the art saliency methods and group them into four families.

(1) *Gradient-based feature attribution (FA)* methods infer input feature importance based on the magnitude of the gradient of the output with respect to the input features. The attribution method Saliency Simonyan et al. (2014) directly employs gradients to generate saliency maps. Extensions of this basic method are Gradient $\times$ Input Shrikumar et al. (2016), DeCovNet Zeiler & Fergus (2014), Guided Backpropagation Springenberg et al. (2015) and SmoothGrad Smilkov et al. (2017). DeepLift Shrikumar et al. (2017) utilizes a neuron attribution-based difference-from-reverence approach to assigning scores. Integrated Gradients (IG) Sundararajan et al. (2017) calculates the path integral from a non-informative baseline input to the respective input feature, tackling the problem of gradient saturation Bastings & Filippova (2020). Relevance-based methods, e.g., Layer-wise Relevance Propagation (LRP) Bach et al. (2015), Deep Taylor Decomposition Montavon et al. (2017), calculate attribution scores by propagating relevance scores from the output back through the network via designed propagation rules.

(2) *Model-agnostic* FA methods can be applied to any black-box classifier without access to the models' parameters Carrillo et al. (2021); Petsiuk et al. (2018). Methods such as Occlusion Zeiler & Fergus (2014), Meaningful Perturbations Fong & Vedaldi (2017) and RISE Petsiuk et al. (2018) assign saliency scores relative to the change in output when the respective feature is perturbed. LIME Ribeiro et al. (2016) fits local interpretable surrogate models to the classifier in the neighborhood of the target sample and calculates the saliency based on these models' parameters. Other methods are inspired by theorems from the field of game theory Datta et al. (2016); Lipovetsky & Conklin (2001); Štrumbelj & Kononenko (2014). In particular, the application of the Shapley Value Shapley (1953) has achieved great popularity. Lundberg & Lee (2017) introduce the SHAP values method to measure feature importance by the Shapley value of a conditional expectation function of the to-be-explained model. (3) A different class of post-hoc methods generates *counterfactual explanations* (CF) as LASTS Guidotti et al. (2020), time series tweaking Karlsson et al. (2020), LatentCF++ Wang et al. (2021), CoMTE Ates et al. (2021) and Native Guide Delaney et al. (2021). These methods identify counter-samples to provide explainability by estimating the required variation in individual input features to change the classification outcome. Since our experiments focus on saliency maps, we exclude CF methods from our investigations in this paper.

For our study we selected four candidate methods from different classes of post-hoc methods: Integrated Gradients (IG), Deep-Lift (DL), LIME and Kernel SHAP (SHAP). As for the classifiers, we utilize long-short term memory networks (LSTM) Hochreiter & Schmidhuber (1997) and convolutional neural networks (CNNs) Le Cun et al. (1989). Since the experiments focus on saliency detection, we also train the LSTM and CNN networks via a saliency-guided training procedure

(SGT) Ismail et al. (2021). This procedure allows networks to produce more consistent saliency scores, as the saliency feedback is used for training the network.

3.1 DATA SET GENERATION

To demonstrate our findings, we designed a simulation study in which time-series data is generated based on the Fourier series model. The Fourier series is a well-known latent model for many natural scenarios Geweke & Singleton (1981); Bracewell (2000) and it is proven that any given univariate time series can be reconstructed from its Fourier latent space using a Fourier transformation function. The Fourier latent space can be defined as a matrix with three rows representing frequencies, amplitudes and phase shifts. In our experiments, the Fourier latent space is a matrix of 3x10 parameters.

We generated a total of ten experiments to understand the response of different saliency methods to different patterns. Our ten experiments include four experiments with temporal shapelet patterns, two with latent amplitude patterns, two with latent frequency patterns, and two with latent phase shift patterns. In each experiment, we build a data set containing 2560 time series of equal length divided into two equally sized classes. For the shapelet experiments, each sample in the data set is generated by first randomly sampling from the latent space and then applying a Fourier transformation to reconstruct its temporal signal from the latent space matrix. Afterwards the time series samples in class 1 were superimposed with a dominant shapelet pattern positioned either at a random location (experiment 1), the end (experiment 2), middle (experiment 3) or start (experiment 4) in the temporal time series. For the latent feature experiments, the latent space matrices for class 0 were sampled from a latent space different from the latent space for class 1. The difference was defined in terms of sampling intervals for frequency, amplitude or phase shift. A detailed description of the sampling distributions per experiment is presented in Table 3 in Appendix A.2. For each experiment, the training, validating and testing sets were generated by random sampling without replacement with a ratio of 80%, 10% and 10%, respectively.

For assigning the labels to the data samples, we induced a simple linear relation between the latent or temporal patterns and the class labels. In the latent scenarios, two classes are distinguishable using a single decision boundary defined as $Z^* = const.$, meaning that only one latent feature is class-distinctive. Likewise, in shapelet-related scenarios, the presence or absence of a specific shapelet decides the label of the data. This allows us to study the latent features individually and in a controlled manner. In such settings, potential poor results can be confidently attributed to the intrinsic weakness of the saliency methods, rather than inappropriate classifiers. The data generation mechanism and the resulting data sets are presented and described in detail in Appendix A.1 and A.2, respectively.

3.2 IMPLEMENTATION DETAILS

In this paper, we investigate the performance of both the classifiers and the saliency methods with special focus on the interpretability of saliency methods. To ensure there is a uniform power between all classifiers, they were designed as simple one-layer networks with no dropouts or other forms of additional regularization. The performance of saliency methods is strongly correlated with the classification performance, which is typically increased through more sophisticated and deeper networks. Therefore, by keeping the architecture simple, we intended to objectively evaluate and compare the explainability methods without the influence of optional variations, preventing overfitting or performance boosting.

All algorithms were implemented in the Python programming language. The classifiers were implemented using the deep learning library *PyTorch* Paszke et al. (2019) with the help of the wrapper *PyTorch Lightning* Falcon (2019). Hyper-parameter optimization was performed through the library *Optuna* Akiba et al. (2019). For the feature attribution techniques, the implementations from the PyTorch based model interpretability library *Captum* Kokhlikyan et al. (2020) were employed.

4 RESULTS

Table 1 reports the average accuracy and F1 score of the chosen classifiers across our ten data sets grouped by the type of the experiments. The results show that overall the CNN trained via saliency-guided training achieves the highest classification performance.

Table 1: Average classification performance on test data across all synthetic data sets.

	Shapelet		**Frequency**		**Phase shift**		**Amplitude**	
Classifier	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
LSTM	0.8535	0.8466	0.9749	0.9470	0.5157	0.4914	0.9981	0.9981
LSTM + SGT	0.8242	0.8417	0.9082	0.9117	0.5352	0.4145	0.9160	0.9230
CNN	0.6221	0.7439	**0.9610**	**0.9633**	0.9629	0.9625	0.9981	0.9981
CNN + SGT	**0.8721**	**0.9138**	**0.9610**	**0.9633**	**0.9649**	**0.9634**	**1.0000**	**1.0000**

It appears that the *LSTM* classifier is seriously challenged during phase-shift experiments. This could be due to the *vanishing gradient* problem of LSTMs, which hinders proper classification if informative patterns are placed in the early time points. Surprisingly, LSTM with saliency-guided training procedure performs slightly worse than LSTM. Unlike the LSTM, the *CNN* largely benefits from the saliency-guided training procedure, especially in the shapelet experiment.

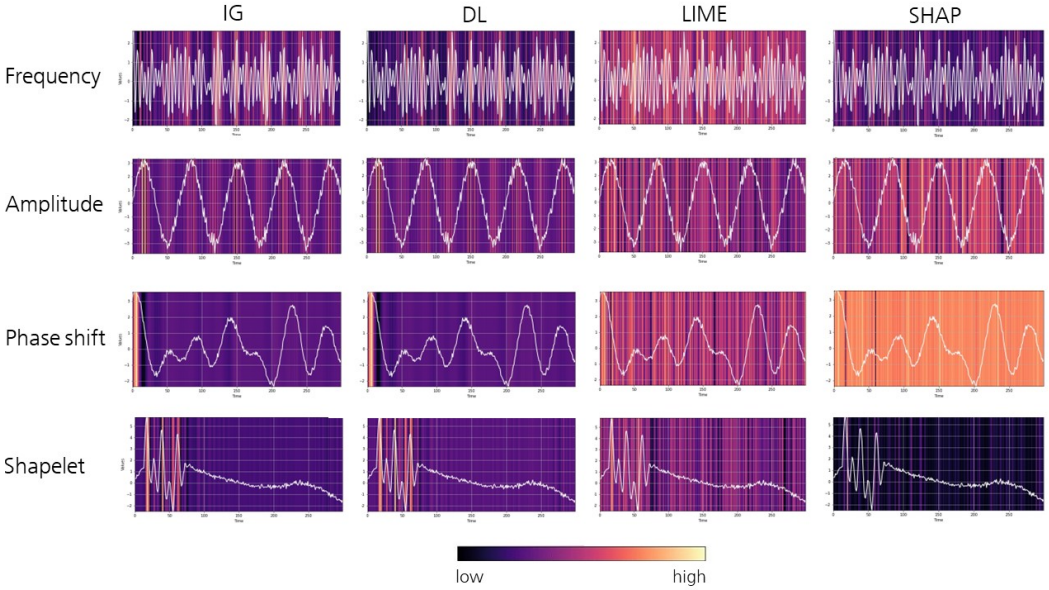

Figure 4: Saliency maps by IG, DL, Lime and SHAP for the CNN+SGT on a frequency, amplitude, phase shift and shapelet experiment, respectively. Explanations by IG and DL clearly focus on aspects related to the latent feature (peaks and valleys for amplitude and frequency, beginning of time sequence for phase shift) and the shapelet. Maps of Lime and SHAP are visually uninterpretable.

Next, to investigate the explainability of the saliency methods, we visualize their output via color-coded heat maps and overlay them onto the original time series (Figure 4). This allows us to assess the relevance of the saliency scores and the positional information directly. In the shapelet experiments, we expect the maps to highlight the shapelet itself. In the amplitude and frequency experiments, we expect an oscillating heat map with a focus on the peaks (or valleys) and extreme values of the time series, respectively. Finally, in the phase shift experiments, we expect an emphasis on the beginning of the time series. Figure 4 compares the saliency maps of the post-hoc saliency methods (IG, DL, LIME, and SHAP) plotted for one sample per experiment group (shapelet- and latent- experiments). Visual explanations provided by IG and DL align with our expectations for all experiments and are comparatively easy to interpret. For example, in the amplitude experiments, IG and DL highlight the peaks whose values are the direct proxies for the latent feature. On the other hand, the heat maps of SHAP and LIME do not yield the expected visual patterns.

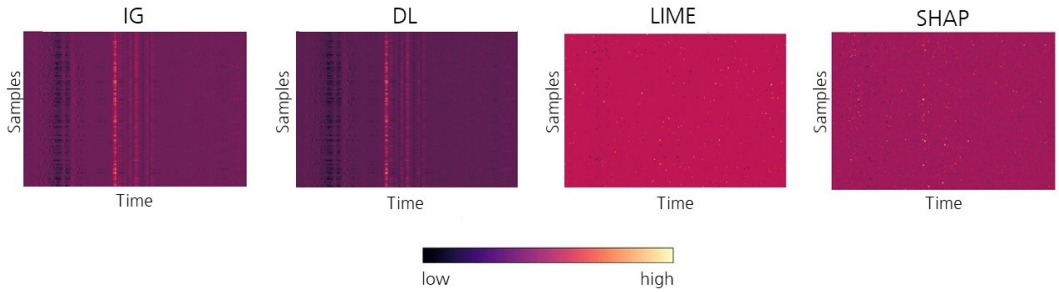

Figure 5: Saliency heat map of IG, DL, LIME and SHAP across all samples of the positive class (occurrence of a shapelet) in the test data set. The IG and DL heat maps show a clear saliency pattern in the middle of the time series, in which the shapelet occurred. The SHAP and LIME heat maps, however, resemble a random saliency assignment.

We expected that the four saliency methods perform reasonably at least for the shapelet experiments. To investigate this further on the entire data set, we generated Figure 5. These heat maps depict aggregated scores of the saliency methods for the middle-positioned shapelet experiments. In these maps, each row represents a test sample and each column a time point. Figure 5 shows that both SHAP and LIME fail to discover the shapelet pattern across the entire data set. The other two methods IG and DL, however, performed successfully: their aggregated heat map clearly highlights the position of the middle shapelet.

5 Discussion

Promised effectiveness of saliency methods for shapelet-related classification The goal of this paper was to demonstrate the fundamental problem of adopted saliency methods for time series data in latent-related classification problems. The methods were expected to be effective in case of the presence of positional information, i.e., shapelets. However, experiments show that some of the methods performed poorly even in simple shapelet scenarios. In particular, explanations provided by different methods mostly did not align. This finding is in accordance with Neely et al. (2021). Our observation raises caution regarding the use of saliency methods for time series data, previously pointed out by Loeffler et al. (2022); Parvatharaju et al. (2021); Schlegel & Keim (2021). In our findings, IG and DL showed reliable performances throughout the experiments when paired with effective classifiers. Nevertheless, we encourage using various explainability methods as multiple explanations can coexist Wiegreffe & Pinter (2019).

Need for latent feature saliency methods for time series classification We emphasize the need for developing latent feature saliency methods for time series classification. Adopted image saliency methods are unable to parse explainable and meaningful saliency scores for time series data with class-distinctive latent patterns. As discussed in Section 2.2, we proposed a definition for "promising failing" methods as ones which produce positional scores associated with informative latent parameters. In the case of Fourier series models, this corresponds to highlighting peaks or valleys, highest peaks, or early time points in case of frequency-, amplitude- and phase-shift-related classification problems, respectively. Not all SoA methods could exhibit such behavior. We hypothesize that this was caused by the independence assumption between neighboring data points, which is made by the tested approaches. Under this assumption, the model neglects the relative temporal ordering of input features, leading to the inability to detect temporal dependencies. This finding is also reported by Lim et al. (2021).

We observed that the IG and DL methods consistently performed well for shapelet-related problems and produced useful saliency maps for latent-related problems. Note that despite calling these methods "promising", the need for directly scoring the latent parameters remains. We expect this problem to exacerbate for latent-related settings whose features contain less legible associations with the positional information, e.g., rates of changes in state-space models.

Future work To extend the empirical investigations, we suggest considering other time series latent models. We further encourage the development of methods that can incorporate multiple feature

spaces into the saliency analysis. With this regard, there is a potential for extracting latent saliency scores directly from positional saliency maps, given that the target latent model is known. Our findings show that the output of IG and DL are associated with the Fourier latent model. This approach (i.e., mapping positional scores to latent scores) serves well as a baseline method.

Throughout our study, the evaluation of saliency maps was performed by visual inspection only, since the primary purpose of this paper to formulate the latent feature saliency problem and motivate further investigation of this topic through a simple experimental framework. For future work, we encourage using quantitative evaluation metrics to assess the performance of different saliency methods objectively. Furthermore, we motivate the extension of our experiments to more complex real-world data sets.

Our analyses were done on the sample level, i.e., we studied individual saliency maps to infer the underlying classification mechanism. Intra-class studies of variability and variance of saliency maps might uncover further information regarding the classification.

6 Conclusion

Explainability of time series models is an uprising field of research. To build trust in AI, interpretation and explanation of black-box classifiers are crucial. Various image saliency methods have been introduced to time series problems. They focus on positional information of the input features, providing spatial explanations. In time series data, however, the class label may depend on a latent model instead of positional information. To the best of our knowledge, performance and behavior of saliency methods in such settings have not been explored, neither has a saliency model accounting for latent features been developed. We demonstrated this problem by empirically showing that if the class label is associated with latent features of the time series instead of the presence of a specific shape, common saliency methods do not provide accurate or interpretable explanations. Finally, we presented an outline for future research to develop extensions for existing saliency methods providing latent saliency results based on time-step-wise importance scores. Our work highlights the need for research on latent saliency detection for deep time series classification.

Acknowledgments and Funding

To ensure the integrity of blind review process, the acknowledgement and funding statements will be added after the review process.

Conflicts of interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Reproducibility Statement

The synthetic data generation algorithm is described in Appendix A.1. The specific data sets employed are stated in terms of the sampling intervals of the latent features in Appendix A.2. Implementation details such as employed libraries were provided in Section 3.2. A GitHub repository containing the complete code base will be published upon final paper publication.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19)*, Anchorage, USA, 2019.

Emre Ates, Burak Aksar, Vitus J. Leung, and Ayse K. Coskun. Counterfactual explanations for multivariate time series. In *Proceedings of the 2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, pp. 1–8, Halden, Norway, 2021.

S. Bach, A. Binder, G. Montavon, F. Klauschen, KR. Müller, and Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, 10(7), 2015. doi: 10.1371/journal.pone.0130140.

Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the 2020 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, online, 2020.

Ronald Newbold Bracewell. *The Fourier transform and its applications*. McGraw-Hill: New York, USA, 3 edition, 2000.

Alfredo Carrillo, Luis F. Cantú, and Alejandro Noriega. Individual explanations in machine learning models: A survey for practitioners. *ArXiv*, 2021. doi: 10.48550/arXiv.2104.04144.

David Charte, Francisco Charte, Maria J del Jesus, and Francisco Herrera. An analysis on the use of autoencoders for representation learning: Fundamentals, learning task case studies, explainability and challenges. *Neurocomputing*, 404:93–107, 2020. doi: 10.1016/j.neucom.2020.04.057.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP)*, pp. 598–617, San Jose, USA, 2016.

Eoin Delaney, Derek Greene, and Mark T. Keane. Instance-based counterfactual explanations for time series classification. In *Proceedings of the Case-Based Reasoning Research and Development: 29th International Conference (ICCBR 2021)*, pp. 32–47, Salamanca, Spain, 2021.

William Falcon. Pytorch lightning, 2019.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy*, pp. 3449–3457, 2017.

John F. Geweke and Kenneth J. Singleton. Latent variable models for time series : A frequency domain approach with an application to the permanent income hypothesis. *Journal of Econometrics*, 17:287–304, 1981. doi: 10.1016/0304-4076(81)90003-8.

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.*, 51(5), 2018. doi: 10.1145/3236009.

Riccardo Guidotti, Anna Monreale, Francesco Spinnato, Dino Pedreschi, and Fosca Giannotti. Explaining any time series classifier. In *Proceedings of the 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*, pp. 167–176, Atlanta, USA, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 9 1997.

Aya Abdelsalam Ismail, Mohamed K. Gunady, Héctor Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6441–6452, online, 2020.

Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. Improving deep learning interpretability by saliency guided training. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS2021)*, pp. 26726–26739, online, 2021.

Isak Karlsson, Jonathan Rebane, Panagiotis Papapetrou, and Aristides Gionis. Locally and globally explainable time series tweaking. *Knowledge Information Systems*, 62:1671–1700, 2020. doi: 10.1007/s10115-019-01389-4.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch. *arXiv*, 2020. doi: 10.48550/arXiv.2009.07896.

Y. Le Cun, L.D. Jackel, B. Boser, J.S. Denker, H.P. Graf, I. Guyon, D. Henderson, R.E. Howard, and W. Hubbard. Handwritten digit recognition: applications of neural network chips and automatic learning. *IEEE Communications Magazine*, 27(11):41–46, 1989. doi: 10.1109/35.41400.

Bryan Lim, Sercan Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37:1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17:319–330, 2001. doi: 10.1002/asmb.446.

Christoffer Loeffler, Wei-Cheng Lai, Bjoern Eskofier, Dario Zanca, Lukas Schmidt, and Christopher Mutschler. Don't get me wrong: How to apply deep visual interpretations to time series. *arXiv*, 2022. doi: 10.48550/ARXIV.2203.07861.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp. 4768–4777, Long Beach, USA, 2017.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Conference on Neural Information Processing Systems (NeurIPS 2013)*, Lake Tahoe, USA, 2013.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017. doi: 10.1016/j.patcog.2016.11.008.

Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic. Order in the court: Explainable ai methods prone to disagreement. In *Proceedings of the ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*, online, 2021.

Prathyush S. Parvatharaju, Ramesh Doddaiah, Thomas Hartvigsen, and Elke A. Rundensteiner. Learning saliency maps to explain deep time series classifiers. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Virtual Event*, pp. 1406–1415, Queensland, Australia, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2019)*, pp. 8024–8035, Vancouver, Canada, 2019.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the 29th British Machine Vision Conference (BMVC)*, Newcastle, UK, 2018.

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 (KDD 16)*, pp. 1135–1144, New York, USA, 2016.

Udo Schlegel and Daniel A. Keim. Time series model attribution visualizations as explanations. In *Proceedings of the 2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pp. 27–31, New Orleans, USA, 2021.

Udo Schlegel, Daniela Oelke, Daniel A. Keim, and Mennatallah El-Assady. An empirical study of explainable ai techniques on deep learning models for time series tasks. In *Proceedings of the Pre-registration workshop NeurIPS (2020)*, Vancouver, Canada, 2020.

Lloyd S. Shapley. *A value for n-person games*, pp. 307–317. Princeton University Press, Princeton, 1953.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. In *Proceedings of the 33rd International Conference on Machine Learning (ICML'16)*, volume 48, New York, USA, 2016.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, volume 70, Sydney, Australia, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv*, abs/1312.6034, 2014. doi: 10.48550/arXiv.1312.6034.

Daniel Smilkov, Nikhil Thorat, Been Kim, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv*, 2017. doi: 10.48550/arXiv.1706.03825.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv*, 2015. doi: 10.48550/arXiv.1412.6806.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, pp. 3319–3328, Sydney, Australia, 2017.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

Zhendong Wang, Isak Samsten, Rami Mochaourab, and Panagiotis Papapetrou. Learning time series counterfactuals via latent space representations. In *Proceedings of the 24th International Conference on Discovery Science (DS 2021)*, pp. 369–384, Halifax, Canada, 2021.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, 2019.

Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD09)*, pp. 947–956, Paris, France, 2009.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the Computer Vision (ECCV 2014)*, pp. 818–833, Zurich, Switzerland, 2014.

A Appendix

A.1 Synthetic data generation

Based on the Fourier series latent model, a time series $x_t, t = 1, ..., T$ is modeled as

$$\begin{aligned} x_t &= a_0 + \sum_{n=1}^{\infty} a_n \cos(\omega_n t) + \sum_{n=1}^{\infty} b_n \sin(\omega_n t) \\ &= a_0 + \sum_{n=1}^{\infty} A_n \cos(\omega_n t + \phi_n) \\ &= a_0 + \sum_{n=1}^{\infty} A_n \sin(\omega_n t + \phi_n + \frac{\pi}{2}). \end{aligned}$$

To simulated data, let $\tilde{n}$ represent the number of amplitudes present in the series, i.e. $\forall i > \tilde{n}, A_i = 0$. For simplicity, we consider centered stationary periodic time series in the data generation process,

i.e. $a_0 = 0$. In this case, the value at every time step t is calculated as

$$x_t = \sum_{i=1}^{\tilde{n}} A_i \sin(\omega_i t + \phi_i + \frac{\pi}{2}). \tag{1}$$

We refer to the notions amplitude A, frequency ω, phase shift ϕ as *concepts*. The separate Fourier coefficients A_i, ω_i, ϕ_i for $i = 1, ..., \tilde{T}$ are referred to as latent features. The latent features frequency ω_i and phase shift ϕ_i are each sampled from a uniform distribution. The sampling intervals are chosen with respect to the specific intention in the experiment design. To simulate the amplitude parameters A_i, a *dominant amplitude* A_1 is sampled. The next amplitudes are calculated considering an exponential decay with a fixed rate dec:

$$A_i = A_1 \exp(-i \cdot dec), \quad i = 1, ..., \tilde{n}.$$

This makes the first frequency i.e. ω_1 to be the dominant frequency of the Fourier series. Throughout the experiments, all time series were generated with an equal length of 300 time steps. i.e. $T = 300$.

For assigning class labels to the time series samples, we consider the following two scenarios.

Scenario 1: Label based on the presence of a shapelet
For assigning shape-based labels to the time series, a shapelet is inserted at a random or fixed position into all time series $X \in D$ belonging to one class. The shapelet is a second simulated Fourier series of length $l \leq T$, where $l =$ window-ratio $\cdot\, T$ for a chosen window ratio. We define the sampling intervals for the latent features of the shapelet to be non-intersecting with the sampling intervals of the latent features of the original time series X. The resulting shapelet replaces the original time series in the interval $[j, j + l]$, where

$$j \sim \mathcal{U}(1, T - l).$$

Scenario 2: Label based on differences in the latent features
Following the investigation of the effectiveness of explainability methods for latent features, we introduce a second simulation scenario where the labels depend on a difference in the sampling distribution of latent features of the time series. This scenario highlights the main focus of this project, and represents our novel view of explainability methods for time series. Similar to the first scenario, the time series are sampled as discretized Fourier series with latent variables ω, A and ϕ. The latent-dependency is induced as follows:

1. Two normal distributions with different means (based on Table 3) are selected for classes 0 and 1. For positive parameters, the distributions are log-normal.
2. Per each class, $N/2$ Fourier parameters are sampled from the given distributions.
3. The rest of the parameters are sampled from the same distribution for both classes.
4. Sampled parameters are given to the deterministic Fourier series in Equation 1 to generate the temporal samples. Rows are then labeled with the associated class, from the corresponding distribution of which the informative parameters are sampled.

A.2 Data set description

Based on the data generation method described above, we design ten different mechanisms for binary classification of univariate time series. Table 2 lists the parameters and algorithms for assigning labels to each sample. In table 3 the parameters used for sampling the Fourier series are presented. The complete simulation code base will be published in a GitHub repository upon final publication.

Table 2: Label-making features per experiment. The overlapping ranges refer to the sampling intervals for frequency and phase shift.

Experiment	Label feature	Description of shapelet
1	Shapelet	Random position, window length of $0.2 * \text{sequence length}$
2	Shapelet	Fixed position, last $0.2 * \text{sequence length}$ time steps
3	Shapelet	Fixed position, starting at time step $0.4 * \text{sequence length}$ with window length $0.2 * \text{sequence length}$
4	Shapelet	Fixed position, first $0.2 * \text{sequence length}$ time steps
5	Frequency	Overlapping frequency ranges
6	Frequency	Overlapping frequency ranges
7	Phase shift	Non-overlapping phase shift ranges
8	Phase shift	Non-overlapping phase shift ranges
9	Amplitude	Different dominant amplitude
10	Amplitude	Different dominant amplitude

Exp.	Number of sines	Freq. low	Freq. high	Phase low	Phase high	Dominant amplitude	Decay rate	Noise ratio
1	10	$\frac{\pi}{300}$	$\frac{\pi}{60}$	$\frac{-\pi}{4}$	$\frac{\pi}{4}$	1	0.3	0.1
2	10	$\frac{\pi}{300}$	$\frac{\pi}{20}$	$\frac{-\pi}{4}$	$\frac{\pi}{4}$	1	0.3	0.1
3	10	$\frac{\pi}{300}$	$\frac{\pi}{20}$	$\frac{-\pi}{4}$	$\frac{\pi}{4}$	1	0.3	0.1
4	10	$\frac{\pi}{300}$	$\frac{\pi}{20}$	$\frac{-\pi}{4}$	$\frac{\pi}{4}$	1	0.3	0.1
5	10/10	$\frac{\pi}{300}/\frac{\pi}{100}$	$\frac{\pi}{20}/\frac{\pi}{2}$	$\frac{-\pi}{4}/\frac{-\pi}{4}$	$\frac{\pi}{4}/\frac{\pi}{4}$	1 / 1	0.3 / 0.3	0.1 / 0.1
6	1/1	$\frac{\pi}{300}/\frac{\pi}{100}$	$\frac{\pi}{20}/\frac{\pi}{2}$	$\frac{-\pi}{4}/\frac{-\pi}{4}$	$\frac{\pi}{4}/\frac{\pi}{4}$	1 / 1	0.3 / 0.3	0.1 / 0.1
7	1/1	$\frac{\pi}{300}/\frac{\pi}{300}$	$\frac{\pi}{20}/\frac{\pi}{20}$	$0 / \frac{-\pi}{4}$	$\frac{\pi}{4}/\frac{\pi}{2}$	1 / 1	0.3 / 0.3	0.1 / 0.1
8	10/10	$\frac{\pi}{300}/\frac{\pi}{300}$	$\frac{\pi}{20}/\frac{\pi}{20}$	$0 / \frac{-\pi}{4}$	$\frac{\pi}{4}/\frac{\pi}{2}$	1 / 1	0.3 / 0.3	0.1 / 0.1
9	10/10	$\frac{\pi}{300}/\frac{\pi}{300}$	$\frac{\pi}{20}/\frac{\pi}{20}$	$0 / \frac{-\pi}{4}$	$\frac{\pi}{4}/\frac{\pi}{4}$	1 / 3	0.3 / 0.3	0.1 / 0.1
10	1/1	$\frac{\pi}{300}/\frac{\pi}{300}$	$\frac{\pi}{20}/\frac{\pi}{20}$	$\frac{-\pi}{4}/\frac{-\pi}{4}$	$\frac{\pi}{4}/\frac{\pi}{4}$	1 / 3	0.3 / 0.3	0.1 / 0.1

Table 3: Overview of simulation parameters of the Fourier series. If two entries are present in one cell, each the classes were sampled from different distributions. The first entry in each cell corresponds to the sampling parameter of class 0, the second entry to class 1.