Learnable Masks for Time Series Explainability using Time-Frequency Representations

Theresa Dahl Frehr*[1], Thea Brüsch[1], and Tommy Sonne Alstrøm[1]

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark
{tdafr, theb, tsal}@dtu.dk

Abstract

The demand for explainable AI models continues to grow with the rise of AI-based solutions. Relatively few explainability methods address time series due to the complex nature of the data. We propose learning saliency maps over both time and frequency via the discrete wavelet transform and the short-time Fourier transform. Faithfulness scores show that our method is on par with current state-of-the-art methods.

1 Introduction

Several critical domains, such as climate [1], finance [2], and healthcare [3], heavily rely on time series data. With the increase of automated processes based on deep learning, the need for explainability of such models becomes increasingly more important. While a wide range of methods have been developed to explain predictions in image-based models, the same cannot be said for time series data.

Time series are difficult to interpret [4]. Despite many models being trained on the raw time signals, the important patterns may exist in a latent feature domain, such as the frequency domain [5].

Learnable masks are a popular choice for creating attribution maps in the time domain, making them a suitable option for explainable time series [6, 7]. The masks are learned through gradient descent using an objective function that masks out much of the input while not changing the model prediction. As these approaches assume that localized salient information is in the time domain, they fall short when relevance is found in the frequency domain.

To address this limitation, FreqRISE [8] was proposed as a model-agnostic framework that jointly learns salient features in both the time and frequency domains. The method estimates relevance using Monte Carlo sampling across multiple masks to identify key features. Although FreqRISE offers competitive performance to established baselines, such as Integrated Gradients [9] and Layer-wise Relevance Propagation [10], it suffers from inefficient sampling and can potentially introduce artefacts by zeroing out frequency components. To overcome these issues, FLEXtime [11] was introduced to explain time series purely in the frequency domain by decomposing the signal into frequency bands using a filterbank.

In this work, we combine the strengths of FLEXtime and FreqRISE by introducing a mask-based approach to identify cross-domain saliency. By representing the signal jointly in the time and frequency domains, our method aims to learn a mask that captures salient features across time and frequency simultaneously.

2 Methods

We consider two signal representations: the Discrete Wavelet Transform (DWT) using the `Symlet2` mother wavelet, and the Short-Time Fourier Transform (STFT). The objective is to learn a mask, M, that highlights the most relevant components of the input signal X for a given prediction task. The masked input is defined as the element-wise product between the time-frequency representation ($\tilde{X}$) of the input (X) and the mask: $X^M = \tilde{X} \circ M$. The prediction based on the masked input is then given by $\hat{y}^M = \boldsymbol{f}(X^M)$, where $\boldsymbol{f}(\cdot)$ represents the trained model. An optimal mask should preserve the model's predictive behaviour, i.e. $\hat{y}^M \approx \hat{y}$, where $\hat{y} = \boldsymbol{f}(X)$ is the original prediction. To measure the deviation caused by masking, we use the cross-entropy loss:

$$\mathcal{L}_D(\hat{y}, \hat{y}^M) = -\sum_{c=1}^{C} \hat{y}_c \log(\hat{y}_c^M). \tag{1}$$

As $\hat{y}^M = \hat{y}$ if M is all ones, we need to impose a sparsity constraint on the mask, which is done using the ℓ_1-norm:

$$\mathcal{L}_R(M) = \max\left(\frac{\|M\|_1}{L} - r, 0\right), \tag{2}$$

where L is the number of elements in the mask and r is a ratio parameter that controls the sparsity threshold. To ensure temporal consistency, we include a smoothness regularization term that penalizes abrupt changes in the mask values over time. For the STFT, where each coefficient corresponds to a uniform time interval, we use the following

[*] Corresponding Author.

Table 1. Hyperparameters.

	λ_R	λ_S	r
DWT	0.5	10	0.1
STFT	2	5	0.05

smoothness loss:

$$\mathcal{L}_S = \frac{1}{L} \sum_{f \in \mathcal{F}} \sum_{t=1}^{T-1} (M_{f,t+1} - M_{f,t})^2, \qquad (3)$$

where $\mathcal{F}$ is the set of frequency bands. For the DWT, where time–frequency resolution varies across scales, a weighted version of this constraint is applied, assigning higher penalties to frequency bands with lower temporal resolution. The overall loss function thus becomes

$$\mathcal{L} = \mathcal{L}_D(\hat{y}, \hat{y}^M) + \lambda_R \mathcal{L}_R(M) + \lambda_S \mathcal{L}_S(M). \qquad (4)$$

Evaluation of an explanation is a challenging task. As no ground truth exists, the quality of an explanation is usually determined by measuring different properties that are deemed desirable. In this work, we focus on faithfulness, which measures how aligned an explanation is with the prediction of a model. The idea is that by removing the 10% most important features as given by the explanation, the mean true class probability drastically drops. Likewise, by adding the 10% most important features, we expect it to be high.

3 Experimental Setup

Experiments are conducted on the AudioMNIST dataset [12], which consists of 30,000 audio recordings of spoken digits (0–9), each repeated 50 times by 60 different speakers (12 female/48 male). All recordings are downsampled to 8kHz and zero-padded to a fixed length of 8,000 samples, following the procedure from [12]. The 1D convolutional neural network proposed by [12] is used as the model $\boldsymbol{f}$. The network was trained and released in [8], and the same weights are used in this study to ensure comparability with previous work. The selected hyperparameters are listed in Table 1. For the DWT representation, the mask is initialized with 10% non-zero values per frequency band, drawn uniformly from $[0, 0.1]$. For the STFT representation, 10% of elements are randomly distributed across time and frequency with values from $[0, 1]$.

Experiments are reported as an average of 300 samples run over 50 different seeds, which can be seen in Table 2. For comparison, FreqRISE gets an insertion faithfullness score of 0.86 and FLEXtime of 0.91 [11]. In Figure 1, an example explanation is shown for the time series corresponding to the spoken digit “4” (top row), using both the DWT (second row) and STFT (third row) representations. Both methods highlight higher-frequency components in the interval $[0.2, 0.4]$ s and lower-frequency components in the interval $[0.4, 0.65]$ s. The mask obtained with the DWT appears less smooth compared to the STFT.

Table 2. Mean faithfulness scores across 50 experiments. I = insertion, D = deletion, RI = random insertion, RD = random deletion.

	I(↑)	D(↓)	RI(↑)	RD(↓)
DWT	0.91	0.63	0.25	0.95
STFT	0.96	0.77	0.41	0.96

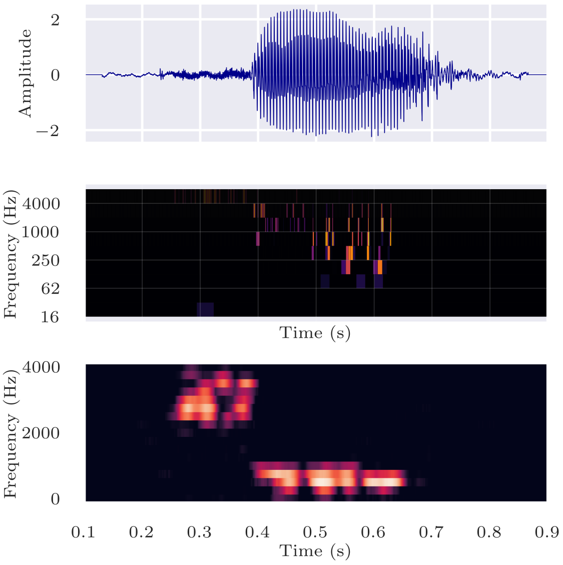

Figure 1. Top: Time series of the digit ”4”. **Middle**: Learned attribution map DWT. **Bottom**: Learned attribution map using STFT.

4 Discussion and Conclusion

We proposed a framework for learning saliency maps in the time–frequency domain using different signal representations. Both the STFT-based and the DWT-based approaches achieved high insertion faithfulness scores, with STFT performing slightly better and even surpassing established baselines such as FreqRISE and FLEXtime. The attribution maps generated from the DWT representation appeared less smooth and displayed more jitter. This may be due to the non-uniform resolution of the DWT. This suggests that the fixed time–frequency resolution of the STFT may be more suitable for audio data.

Future work will focus on optimizing hyperparameters, extending the evaluation to additional datasets and metrics, and conducting a more systematic comparison with other explainability methods.

Acknowledgments

This work was supported by the William Demant Fonden under grant number 24-5326.

References

[1] J. González-Abad, J. Baño-Medina, and J. M. Gutiérrez. "Using Explainability to Inform Statistical Downscaling Based on Deep Learning Beyond Standard Validation Approaches". en. In: *Journal of Advances in Modeling Earth Systems* 15.11 (Nov. 2023), e2023MS003641. ISSN: 1942-2466, 1942-2466. DOI: 10.1029/2023MS003641. URL: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023MS003641 (visited on 10/13/2025).

[2] P. Giudici and E. Raffinetti. "SAFE Artificial Intelligence in finance". en. In: *Finance Research Letters* 56 (Sept. 2023), p. 104088. ISSN: 15446123. DOI: 10.1016/j.frl.2023.104088. URL: https://linkinghub.elsevier.com/retrieve/pii/S1544612323004609 (visited on 10/13/2025).

[3] H. Phan and K. Mikkelsen. "Automatic sleep staging of EEG signals: recent development, challenges, and future directions". en. In: *Physiological Measurement* 43.4 (Apr. 2022), 04TR01. ISSN: 0967-3334, 1361-6579. DOI: 10.1088/1361-6579/ac6049. URL: https://iopscience.iop.org/article/10.1088/1361-6579/ac6049 (visited on 10/13/2025).

[4] T. Rojat, R. Puget, D. Filliat, J. D. Ser, R. Gelin, and N. Díaz-Rodríguez. *Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey.* en. arXiv:2104.00950 [cs]. Apr. 2021. DOI: 10.48550/arXiv.2104.00950. URL: http://arxiv.org/abs/2104.00950 (visited on 10/13/2025).

[5] M. Schröder, A. Zamanian, and N. Ahmidi. "Post-hoc Saliency Methods Fail to Capture Latent Feature Importance in Time Series Data". en. In: *Trustworthy Machine Learning for Healthcare.* Ed. by H. Chen and L. Luo. Vol. 13932. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023, pp. 106–121. ISBN: 978-3-031-39538-3 978-3-031-39539-0. DOI: 10.1007/978-3-031-39539-0_10. URL: https://link.springer.com/10.1007/978-3-031-39539-0_10 (visited on 10/13/2025).

[6] J. Crabbé. "Explaining Time Series Predictions with Dynamic Masks". en. In: (2021).

[7] J. Enguehard. "Learning Perturbations to Explain Time Series Predictions". en. In: (2023).

[8] T. Brusch, K. K. Wickstrøm, M. N. Schmidt, T. S. Alstrøm, and R. Jenssen. "FreqRISE: Explaining time series using frequency masking". en. In: *Northern Lights Deep Learning Conference 2025* (2024).

[9] M. Sundararajan, A. Taly, and Q. Yan. "Axiomatic Attribution for Deep Networks". en. In: *Proceedings of the 34th International Conference on Machine Learning.* 70 (2017), pp. 3319–3328.

[10] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation". en. In: *PLOS ONE* 10.7 (July 2015). Ed. by O. D. Suarez, e0130140. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0130140. URL: https://dx.plos.org/10.1371/journal.pone.0130140 (visited on 10/13/2025).

[11] T. Brüsch, K. K. Wickstrøm, M. N. Schmidt, R. Jenssen, and T. S. Alstrøm. *FLEXtime: Filterbank learning to explain time series.* en. arXiv:2411.05841 [cs]. Apr. 2025. DOI: 10.48550/arXiv.2411.05841. URL: http://arxiv.org/abs/2411.05841 (visited on 10/10/2025).

[12] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, and W. Samek. "AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark". en. In: *Journal of the Franklin Institute* 361.1 (Jan. 2024), pp. 418–428. ISSN: 00160032. DOI: 10.1016/j.jfranklin.2023.11.038. URL: https://linkinghub.elsevier.com/retrieve/pii/S0016003223007536 (visited on 10/10/2025).