Learning Saliency Maps to Explain Deep Time Series Classifiers

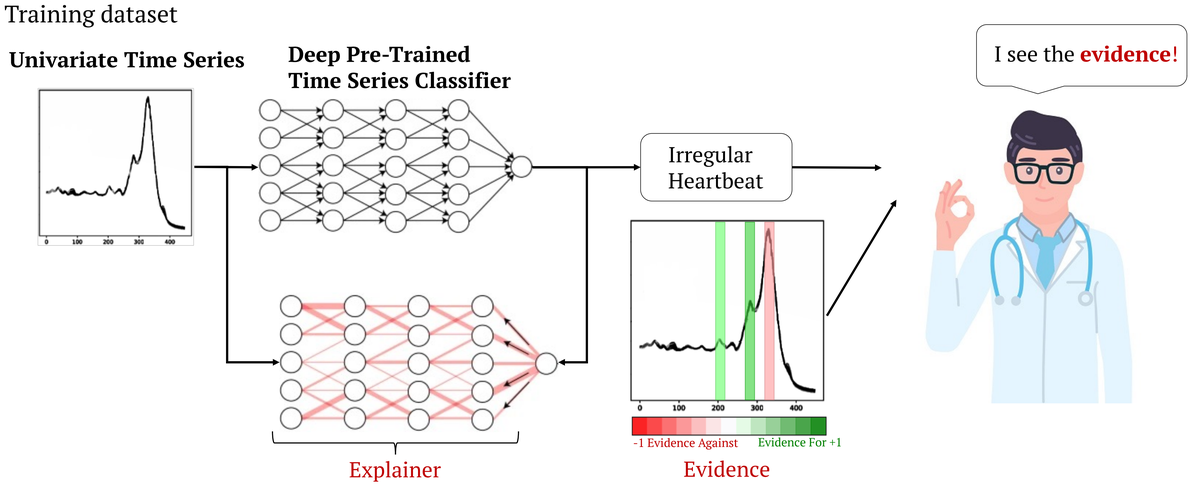
Prathyush Parvatharaju, Ramesh Doddaiah, Tom Hartvigsen, Elke Rundensteiner

CIKM 2021 1-5 NOVEMBER

Problem Definition

Given:
- A univariate time series
- Deep pre-trained classifier
- Training dataset

Goal: Assign one value $\theta_t \in [-1, 1]$ per timestep indicating evidence **for** & **against** the classifier's prediction (saliency map)

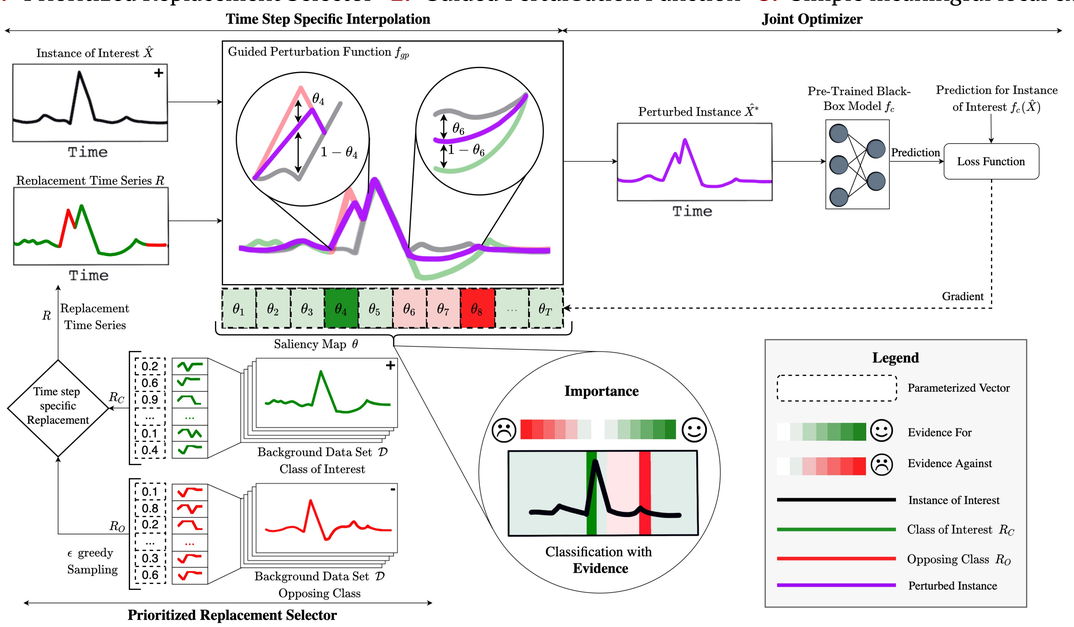


Proposed Method: **PE**rturbation by Prioritized **R**eplacemen**T**

Main idea: Learn classifier's sensitivity to change in input time series

Key Innovations:

1. Prioritized Replacement Selector 2. Guided Perturbation Function 3. Simple meaningful local explanation



Simple Meaningful and Local Explanations

$$L(P(\hat{X}); \theta) = (L_{\text{preservation}} + L_{\text{budget}} + L_{\text{TV}})$$

Component Coefficient; Prediction for Original Instance; Prediction for Perturbed Instance

$$L_{\text{preservation}} = \lambda_1 \left(\frac{1}{\|\hat{X}\|} \sum_{t=0}^{T} (f_c(\hat{X}) - f_c(f_p(\hat{X}; \theta)))^2 \right)$$

Mean Square Error

Component Coefficient; Time step specific Importance value

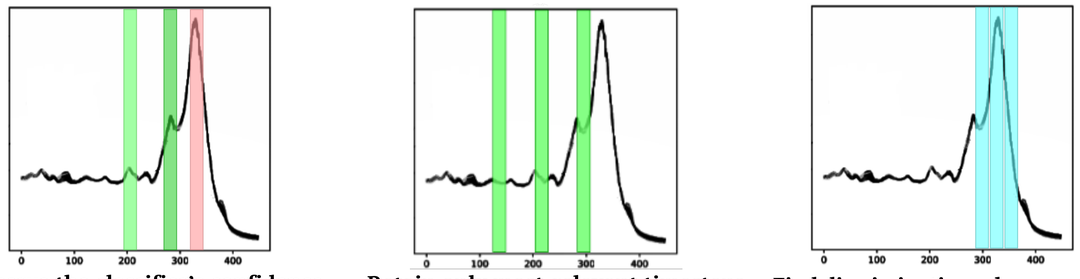
$$L_{\text{budget}} = \lambda_2 \left(\frac{1}{\|\theta\|} \sum_{t=0}^{T} |\theta_t| \right)$$

Minimize the sum of Saliency Map (θ)

Component Coefficient

$$L_{\text{TV}} = \lambda_3 \left(\frac{1}{\|\theta\|} \sum_{t=0}^{T-1} (\theta_t - \theta_{t+1})^2 \right)$$

Total Variance Normalization



Preserve the classifier's confidence | **Retain only most-relevant timesteps** | **Find discriminative subsequences**

Metrics and Evaluation

- 3 Metrics • 9 Real World Datasets • 2 Black Box Models (FCN and RNN) • 1 Baseline, 5 SOTA Explainers

Methods	Datasets								
	WAFER	GUNPOINT	COMPUTERS	EARTHQUAKES	FORDA	FORDB	CRICKETX	PTB	ECG
Random	0.01 (.01)	0.03 (.01)	0.01 (.01)	0.04 (.01)	0.01 (.01)	0.01(.01)	-0.01 (.01)	0.07 (.04)	0.01 (.06)
RISE	0.13 (.01)	0.10 (.01)	-0.01 (.02)	0.23 (.05)	0.15 (.01)	0.11 (.02)	0.42 (.01)	0.10 (.05)	0.19 (.07)
LEFTIST	0.16 (.01)	0.15 (.03)	-0.16 (.01)	0.53 (.03)	0.15 (.02)	0.15 (.01)	-0.10 (.01)	0.42 (.01)	0.51 (.01)
LIME	0.07 (.01)	0.02 (.01)	0.05 (.03)	-0.02 (.01)	0.01 (.01)	0.01 (.01)	0.03 (.01)	0.12 (.07)	0.09 (.06)
SHAP	-0.15 (.01)	-0.01 (.01)	0.10 (.01)	**0.80 (.03)**	0.23 (.01)	-0.17 (.01)	0.30 (.01)	-0.14 (.01)	0.08 (.09)
MP	0.55 (.01)	0.02 (.01)	0.16 (.01)	0.30 (.01)	0.47 (.01)	0.39 (.01)	0.23 (.01)	0.30 (.01)	-0.15 (.01)
PERT	**0.78 (.01)**	**0.48 (.01)**	**0.92 (.01)**	**0.82 (.01)**	**0.70 (.01)**	**0.70 (.01)**	**0.68 (.01)**	**0.52 (.01)**	**0.57 (.01)**

Table : Average performance of the AUC-difference metric with the RNN black-box model.

PERT outperforms state-of-the-art methods by an average of **26%**

Code is publicly-available at **https://github.com/kingspp/timeseries-explain/**

OAK RIDGE INSTITUTE FOR SCIENCE AND EDUCATION Shaping the Future of Science (CA W911NF-16-2-0008, W911NF20-2-0232)
(P200A180088)
National Science Foundation (NSF 1910880, CSSI: FAIN: 2103832)
Weights & Biases
WPI

Thanks to DAISY Group and Academic Research Computing at WPI