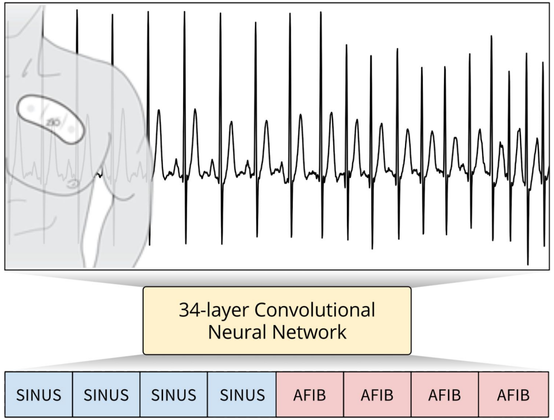
PERT: Learning Saliency Maps to Explain Deep Time Series Classifiers

Prathyush Parvatharaju, Ramesh Doddaiah, Tom Hartvigsen, Elke Rundensteiner

Worcester Polytechnic Institute

Deep Networks are powerful but complex



CNN detects arrhythmia[1]

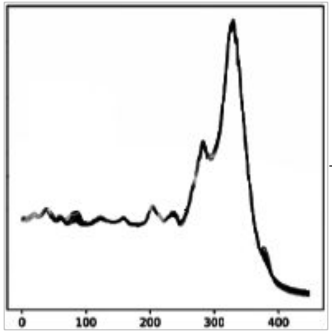
Healthcare - Slow adoption of deep learning models

- **Mistakes** can have catastrophic effects
- **Hard to trust** deep learning model predictions
- FDA mandates **doctor verification**

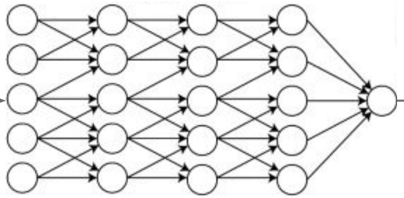
[1] Rajpurkar, Pranav et al. "Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks." ArXiv abs/1707.01836 (2017): n. pag.

Deep Time Series Classifiers are not Explainable

ECG Time Series



Deep Time Series Classifier



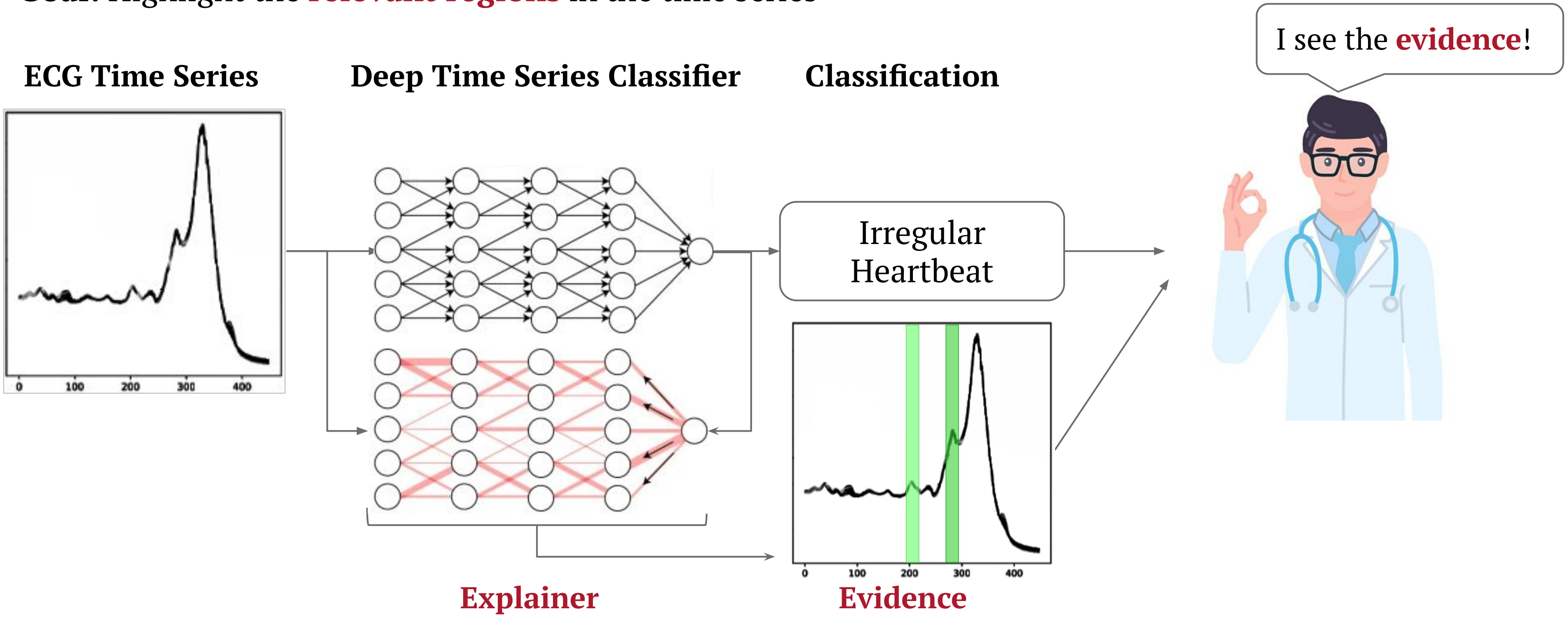
Classification

Irregular Heartbeat

Which **evidence** is used for the decision ?

Making Deep Time Series Classifiers Explainable

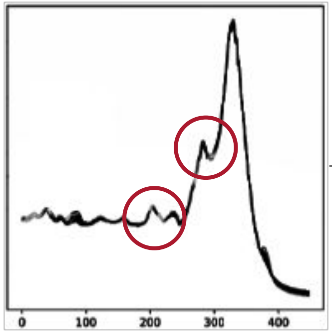
Goal: Highlight the **relevant regions** in the time series

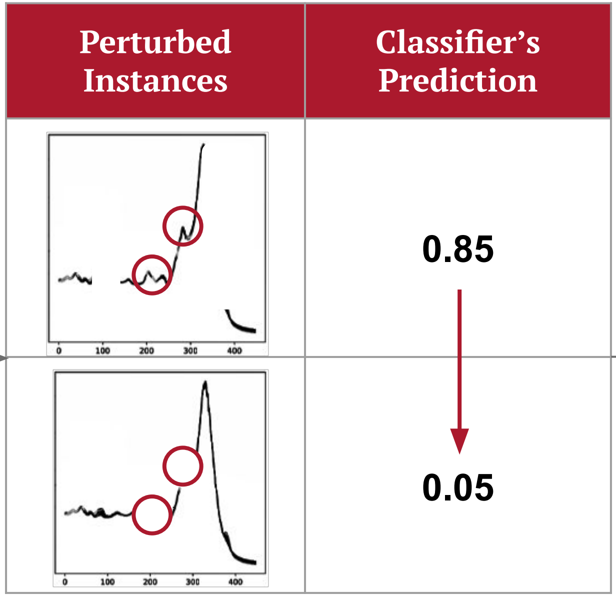


Perturbation produces state-of-the-art explanations

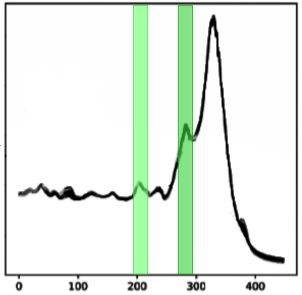
Approach: Perturb regions[1] of time series and observe effects on classifier's predictions

ECG Time Series





Explanation



[1] Ribeiro et al. Why Should I Trust You? Explaining the Predictions of Any Classifier. KDD, 2016.

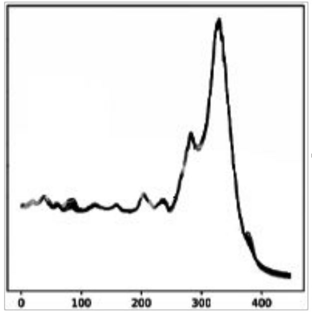
Problem Definition: Perturbation learning for time series models

Given:

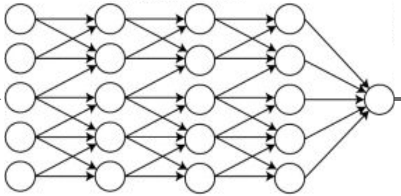
- A univariate time series
- Deep pre-trained classifier
- Training dataset

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

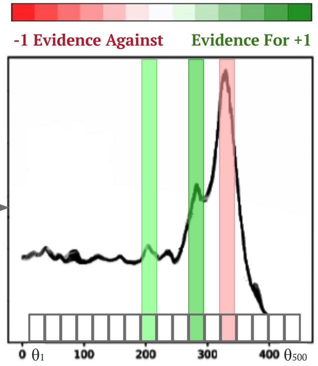
Univariate time series instance



Deep Pre-Trained Time Series Classifier



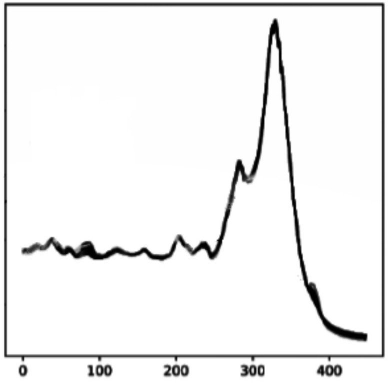
P(Irregular Heartbeat) = 0.85



Saliency Map θ

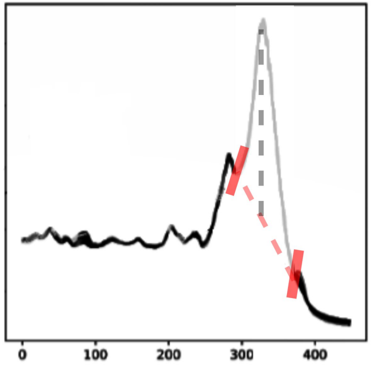
Challenges

1. **Generating Realistic Perturbations**
2. Heterogenous series
3. Perturbing long time series



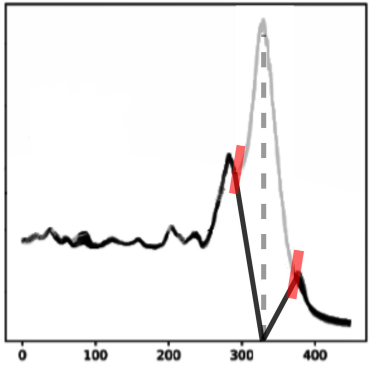
Perturbations

Deletion



X

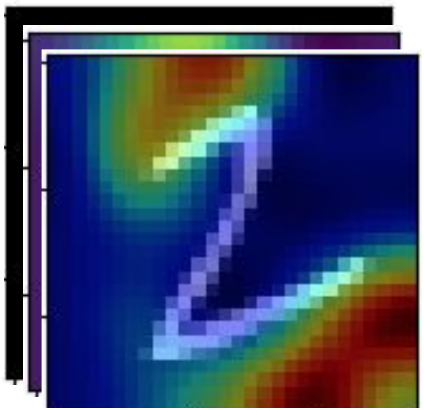
Zero Replacement



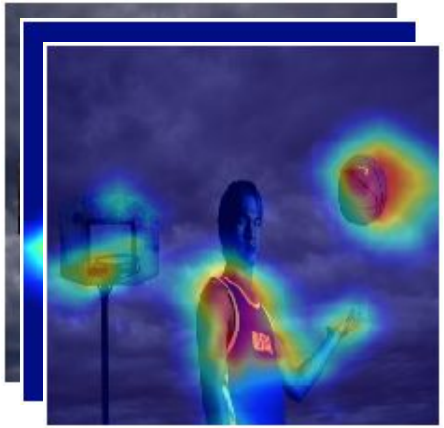
X

Challenges

1. Generating Realistic Perturbations
2. **Heterogenous series**
3. Perturbing long time series



MNIST (2)



Imagenet(Basketball)

255

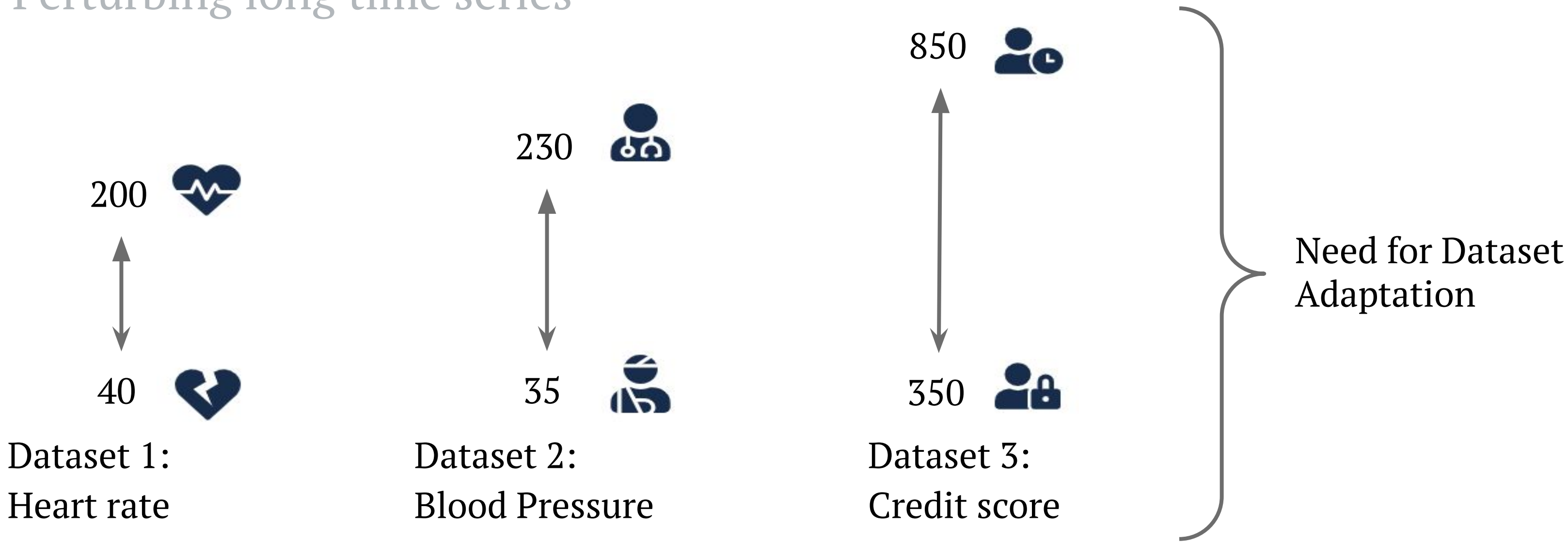
0

Image Datasets:
MNIST, CIFAR
ImageNet, COCO
(Pixel range)

Scope for Generalization

Challenges

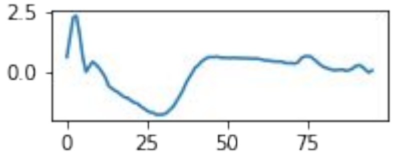
1. Generating Realistic Perturbations
2. **Heterogenous series**
3. Perturbing long time series



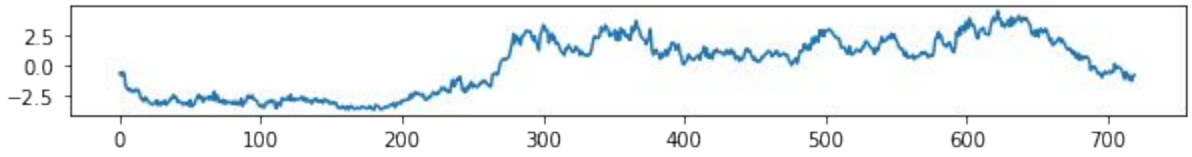
Challenges

1. Generating Realistic Perturbations
2. Heterogenous series
3. **Perturbing long time series**

Short Time Series: (Relatively easy to perturb)



Long Time Series: (Difficult to perturb)



Need for gradient based search space optimization

Proposed Method: PERT*

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

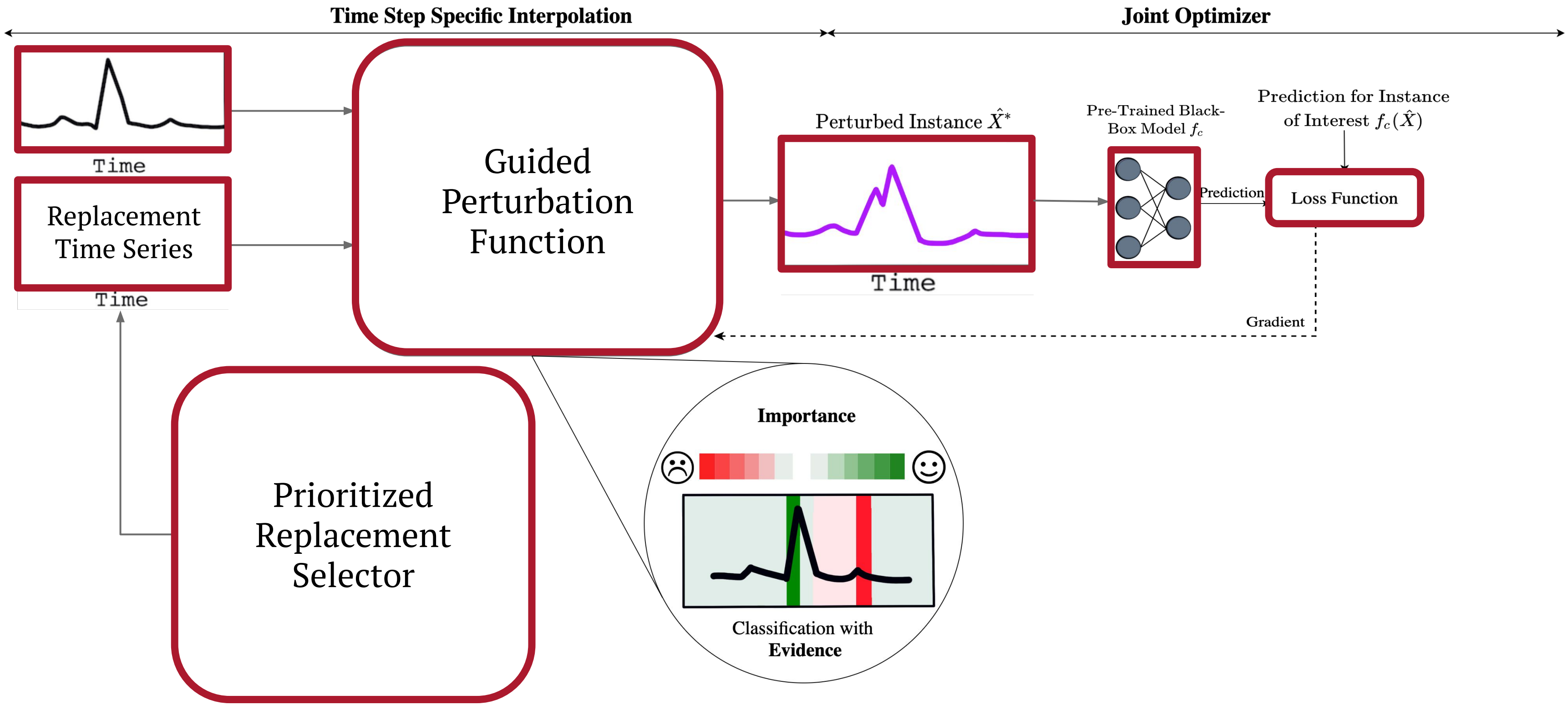
Approach: Use of gradient descent to generate guided perturbations

Key Innovations:

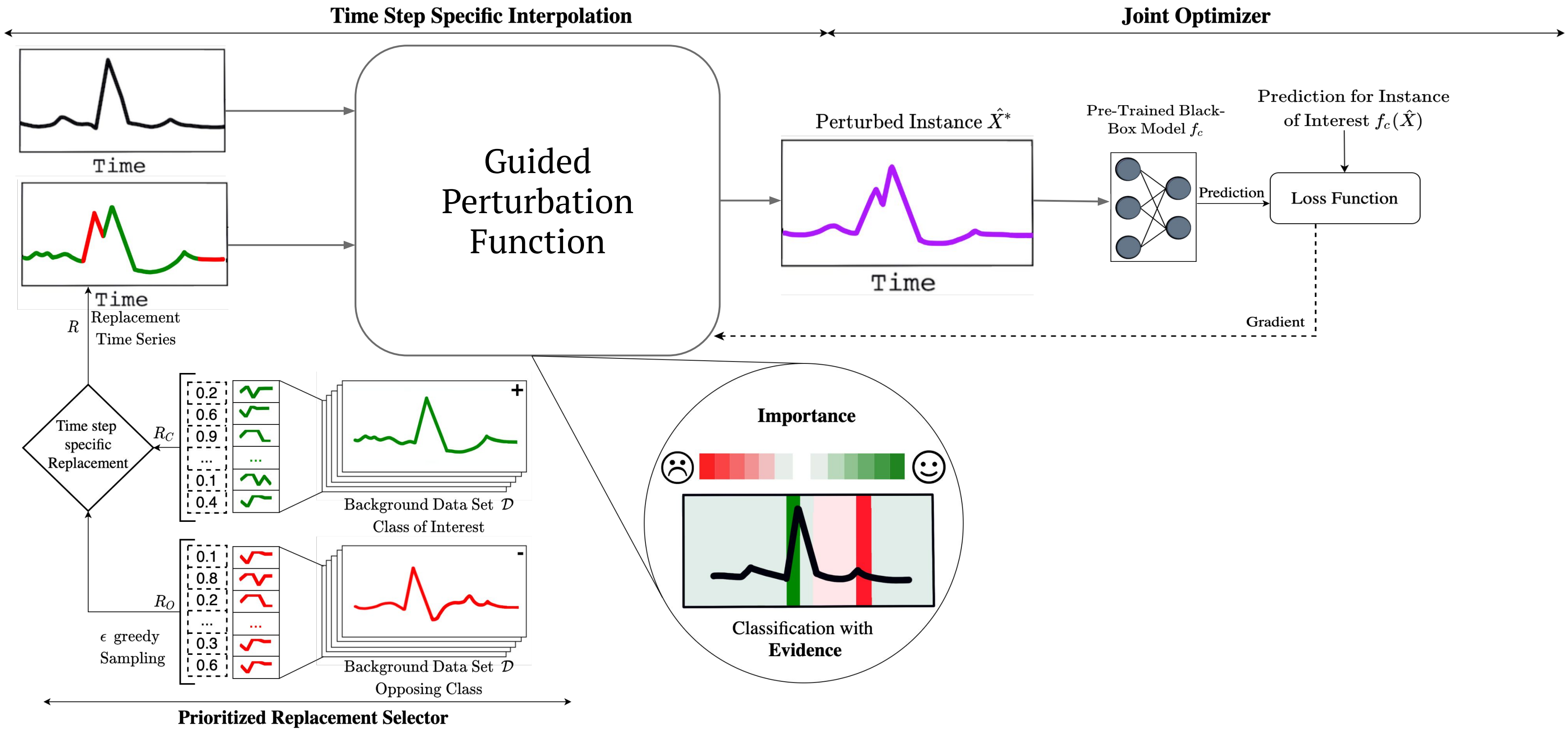
1. Prioritized-Time step specific replacement strategy
2. Guided Perturbation Function
3. Simple meaningful local explanation

*__PE__rturbation by Prioritized __R__eplacemen__T__

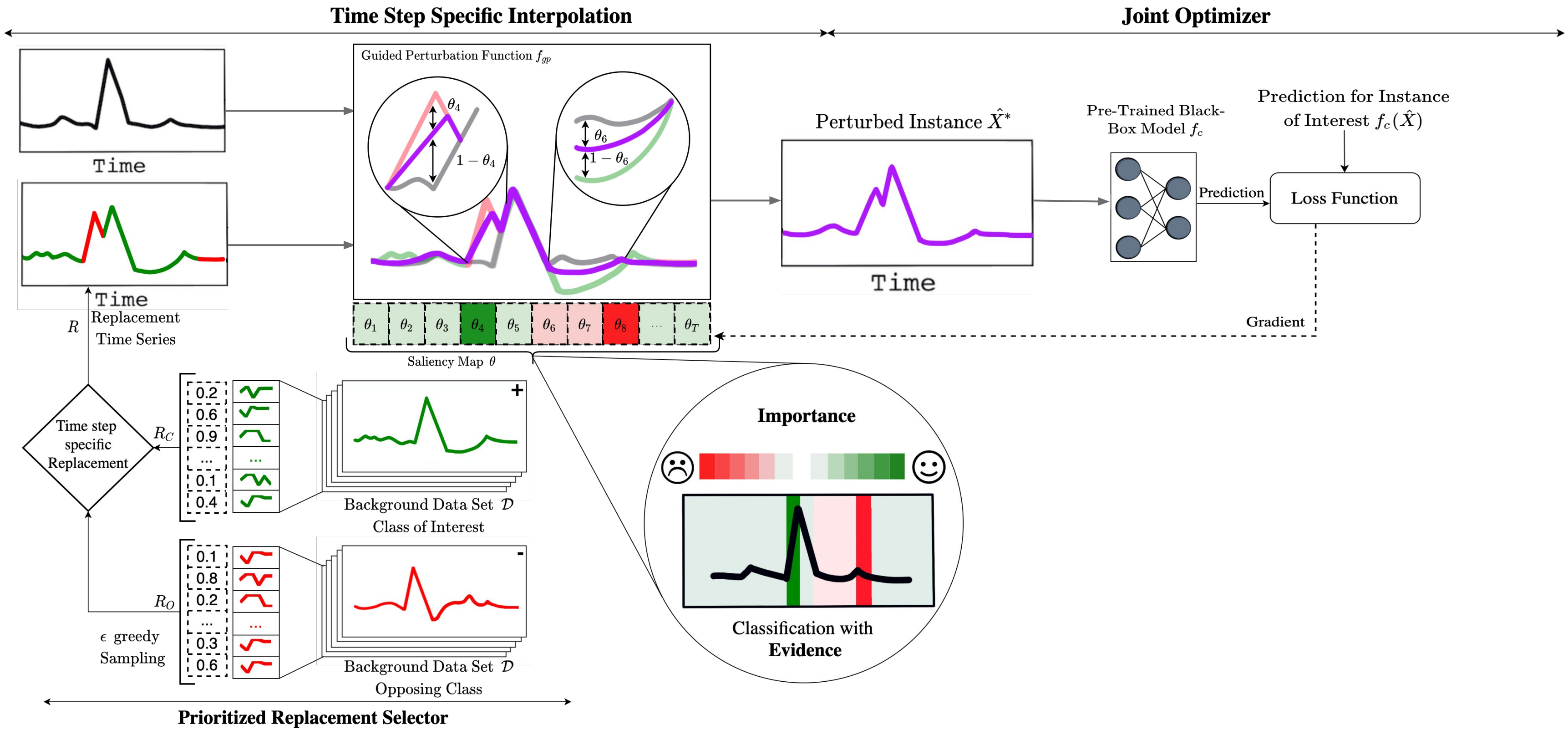
PERT: A High-level View



Prioritized Replacement Selector



Guided Perturbation Function



Learning Simple and Meaningful Local Explanations

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

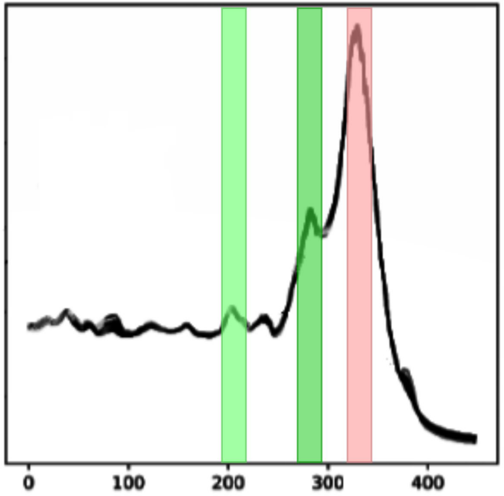
Component Coefficient

Prediction for Original Instance

Prediction for Perturbed Instance

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

Preserve the classifier's confidence



Learning Simple and Meaningful Local Explanations

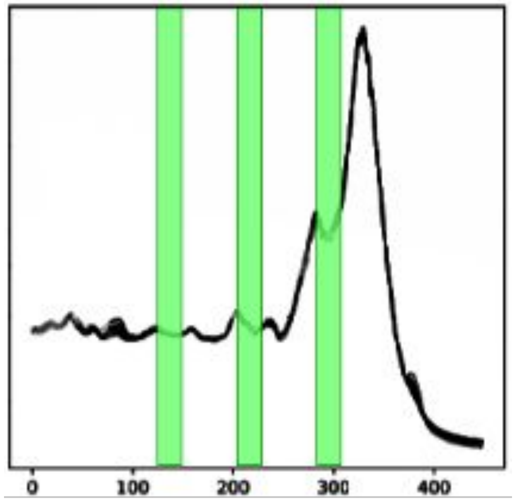
$$L(P(\hat{X});\theta) = (L_{\text{preservation}} + \boxed{L_{\text{budget}}} + L_{\text{TV}})$$

Component Coefficient

Time step specific importance value

$$L_{\text{budget}} = \lambda_2 \underbrace{\left(\frac{1}{\|\theta\|} \sum_{t=0}^{T} |\theta_t| \right)}_{\text{Minimize the sum of Saliency Map } (\theta)}$$

Retain only most-relevant timesteps



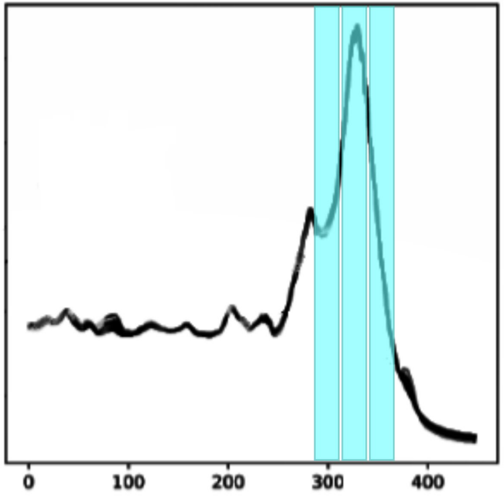
Learning Simple and Meaningful Local Explanations

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

Component
Coefficient

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

Find discriminative subsequences



Experiments

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

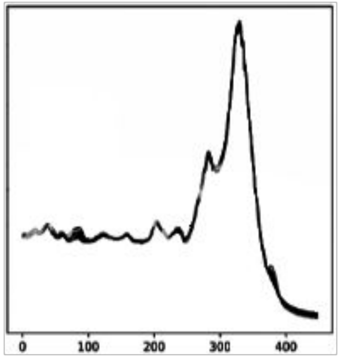
Dataset	WAFER	GUNPOINT	COMPUTERS	EARTHQUAKES	FORDA	FORDB	CRICKETX	PTB	ECG
Num. Train Instances	1000	50	250	322	3601	3636	390	1456	100
Num. Test Instances	6164	150	250	139	1320	810	390	1456	100
Num. Timesteps	152	150	720	512	500	500	300	187	96
FCN Accuracy (%)	99	99	80	75	96	92	81	98	98
RNN Accuracy (%)	99	99	79	75	96	92	80	98	98

Table 1: Summary statistics for the real-world datasets and the Accuracy of our corresponding FCN and RNN models.

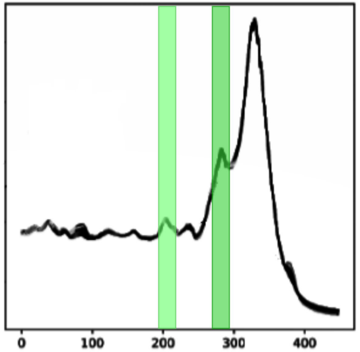
Metrics - AUC Difference

Saliency maps evaluated by "inserting" or "deleting" timesteps from time-series

ECG Time Series

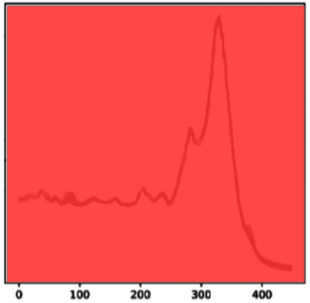


Evidence

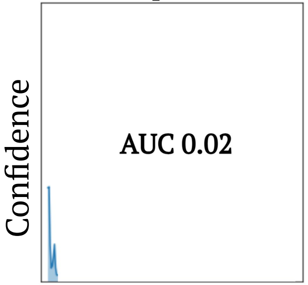


Insertion

Perturbations



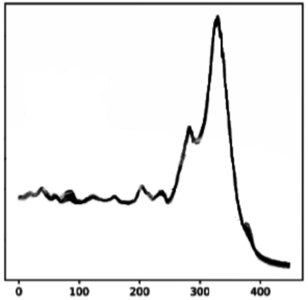
% timesteps inserted



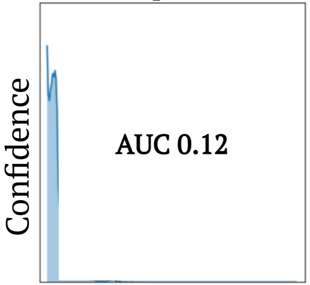
0.02 - 0.12 = -0.1
(AUC Difference)

Deletion

Perturbations



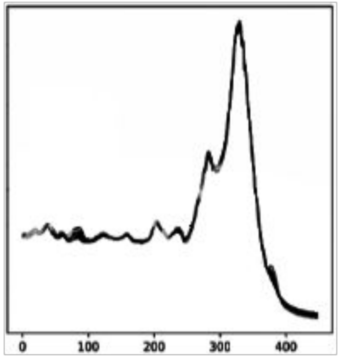
% timesteps deleted



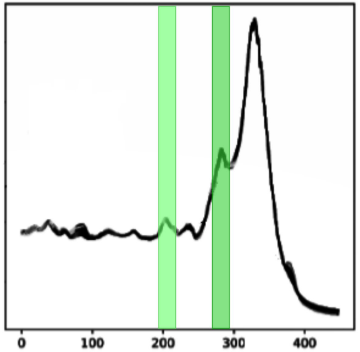
Metrics - AUC Difference

Saliency maps evaluated by “inserting” or “deleting” timesteps from time-series

ECG Time Series

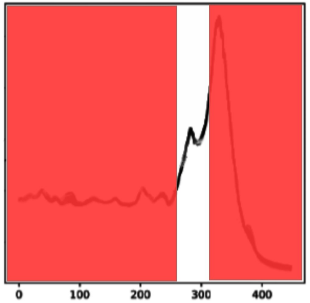


Evidence

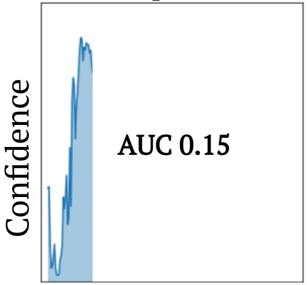


Insertion

Perturbations



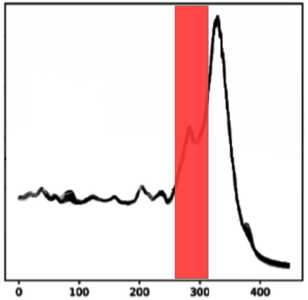
% timesteps inserted



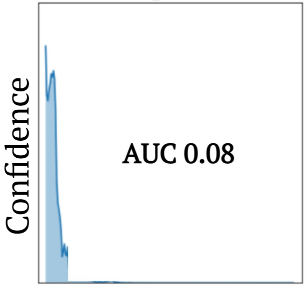
0.15 - 0.08 = 0.07
(AUC Difference)

Deletion

Perturbations



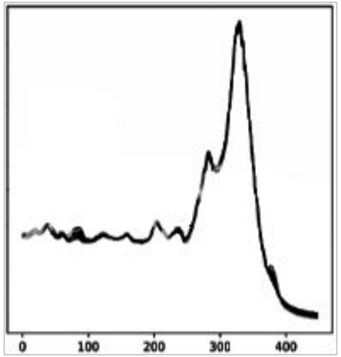
% timesteps deleted



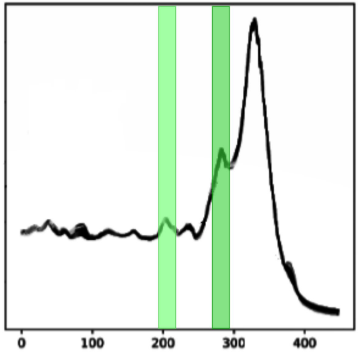
Metrics - AUC Difference

Saliency maps evaluated by “inserting” or “deleting” timesteps from time-series

ECG Time Series

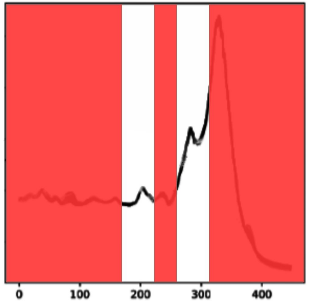


Evidence

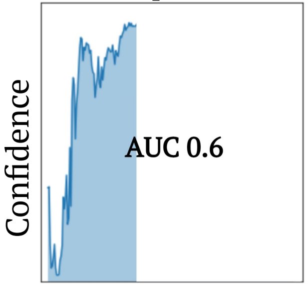


Insertion

Perturbations



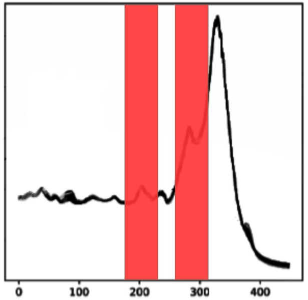
% timesteps inserted



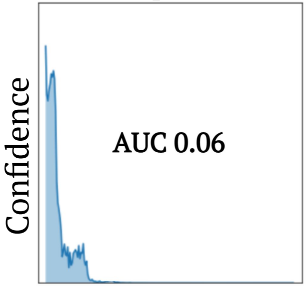
0.60 - 0.06 = 0.54
(AUC Difference)

Deletion

Perturbations

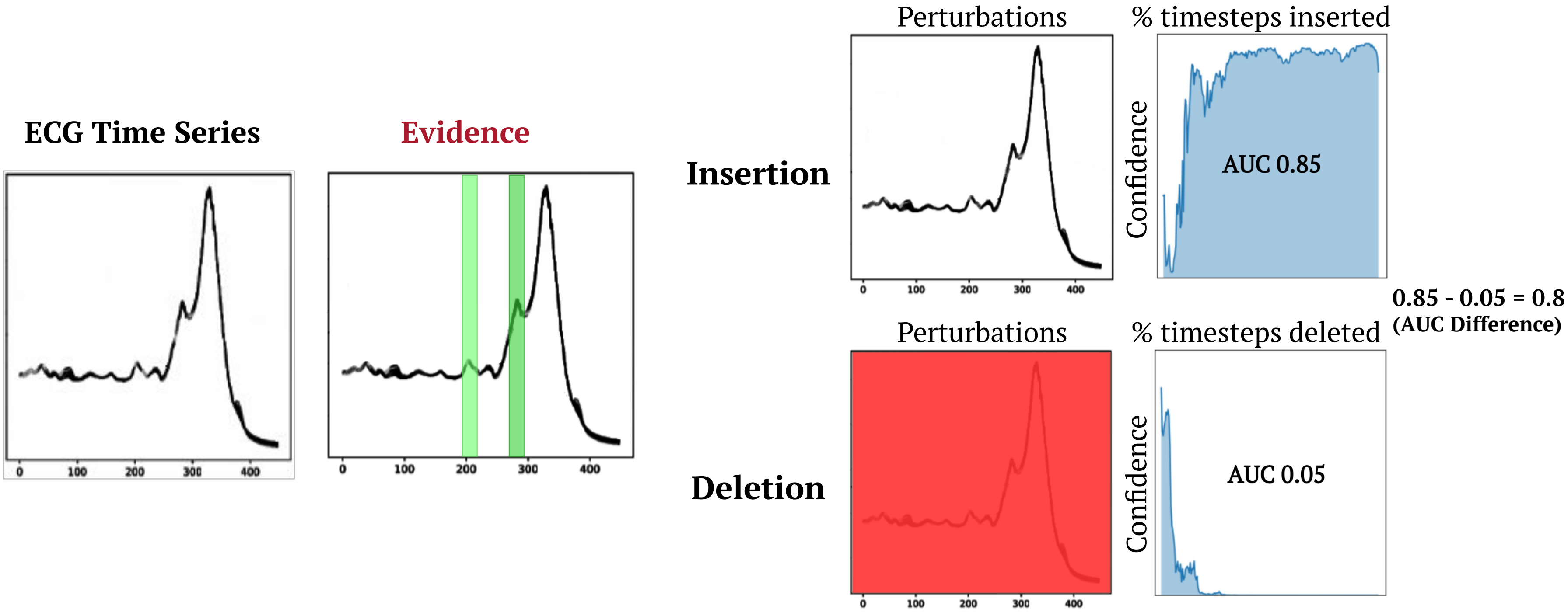


% timesteps deleted



Metrics - AUC Difference

Saliency maps evaluated by "inserting" or "deleting" timesteps from time-series



AUC-Difference Results - RNN

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 3: 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%**

Conclusion

- Identify the need for attribution-based explanations for deep time series classifiers
- We formalize Perturbation Learning for time series classifiers
- Propose **PERT**, a novel perturbation method specific to time series
- Demonstrate PERT achieves state-of-the-art performance on **9 datasets** and **3 metrics**
- Our code is publicly-available at **https://github.com/kingspp/timeseries-explain**

Thank You

DAISY Group at WPI

Academic Research Computing at WPI

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