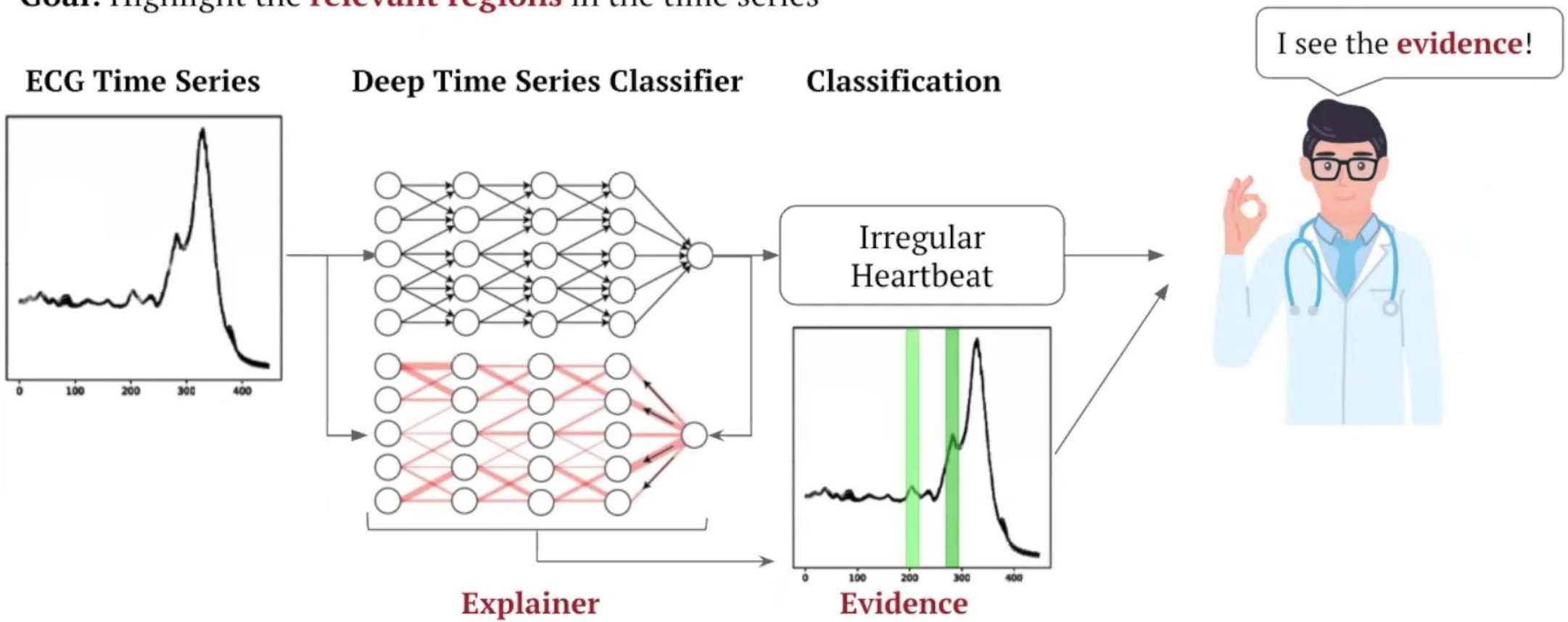
时间序列可解释性研究

汇报人：钟博

时间：2022-05-17

一元时间序列：时间点t只有一个维度的数据。（ECG）

多元时间序列：时间点t有多个维度与时间相关的数据。（EHR）



常见的时间序列可解释性方法

- Gradient-based:将梯度作为输入重要性的近似值。
- Perturbation-based:研究不同输入扰动下的输出变化。
- Attention-based:基于注意力分数。
- other: Shapley、Lime ……

Explaining Time Series Predictions with Dynamic Masks

Jonathan Crabbé [1] **Mihaela van der Schaar** [1 2 3]

挑战与研究动机

Contributions By building on the challenges that have been described, our work is, to our knowledge, the first saliency method to rigorously address the following questions in a time series setting.

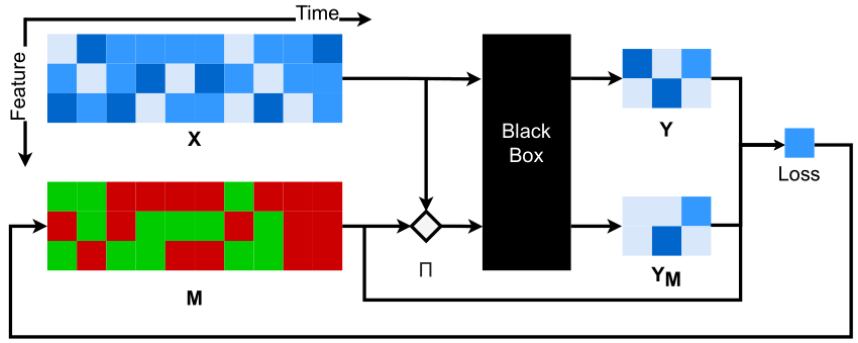


Figure 2. Diagram for Dynamask. An input matrix **X**, extracted from a multivariate time series, is fed to a black-box to produce a prediction **Y**. The objective is to give a saliency score for each component of **X**. In Dynamask, these saliency scores are stored in a mask **M** of the same shape as the input **X**. To detect the salient information in the input **X**, the mask produces a perturbed version of **X** via a perturbation operator Π. This perturbed **X** is fed to the black-box to produce a perturbed prediction $\mathbf{Y}_\mathbf{M}$. The perturbed prediction is compared to the original prediction and the error is backpropagated to adapt the saliency scores contained in the mask.

如何建立有意义的扰动？ → 利用该特征在相邻时间的值，对每个特征在每个时间建立一个扰动。

如何以给定精度重建黑盒预测的最小输入数? → extremal mask selects

如何清晰易读? → encourage mask to be almost binary

问题定义

$$\mathbf{Y} = f(\mathbf{X})$$

$$\mathbf{X} \in \mathbb{R}^{T \times d_X}$$

$$\mathbf{M} = (m_{t,i}) \in [0,1]^{T \times d_X}$$

当系数接近1时表示特征显著，当系数接近0时表示特征不显著。

表示特征i在t时刻对f产生预测的重要性。

Perturbation operators

$$\Pi_{\mathbf{M}} : \mathbb{R}^{T \times d_X} \rightarrow \mathbb{R}^{T \times d_X}$$

1. The perturbation for $x_{t,i}$ is dictated by $m_{t,i}$:

$$[\Pi_{\mathbf{M}}(\mathbf{X})]_{t,i} = \pi(\mathbf{X}, m_{t,i}; t, i),$$

where π is differentiable for $m \in (0,1)$ and continuous at $m = 0, 1$.

确保扰动独立地应用于所有输入，并且适用于基于梯度的优化。

2. The action of the perturbation operator is trivial when the mask coefficient is set to one :

$$\pi(\mathbf{X}, 1; t, i) = x_{t,i}.$$

扰动对显著输入的影响较小。

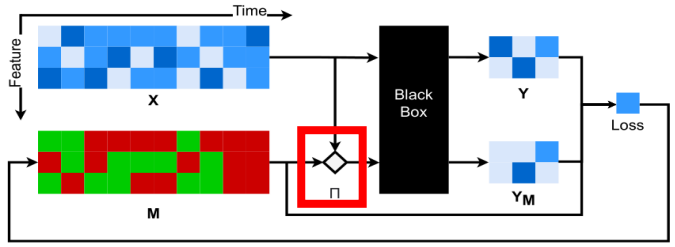


Figure 2. Diagram for Dynamask. An input matrix $\mathbf{X}$, extracted from a multivariate time series, is fed to a black-box to produce a prediction $\mathbf{Y}$. The objective is to give a saliency score for each component of $\mathbf{X}$. In Dynamask, these saliency scores are stored in a mask $\mathbf{M}$ of the same shape as the input $\mathbf{X}$. To detect the salient information in the input $\mathbf{X}$, the mask produces a perturbed version of $\mathbf{X}$ via a perturbation operator Π. This perturbed $\mathbf{X}$ is fed to the black-box to produce a perturbed prediction $\mathbf{Y}_{\mathbf{M}}$. The perturbed prediction is compared to the original prediction and the error is backpropagated to adapt the saliency scores contained in the mask.

$$\pi^g(\mathbf{X}, m_{t,i}; t, i) = \frac{\sum_{t'=1}^{T} x_{t',i} \cdot g_{\sigma(m_{t,i})}(t - t')}{\sum_{t'=1}^{T} g_{\sigma(m_{t,i})}(t - t')}$$

$$g_\sigma(t) = \exp\left(-\frac{t^2}{2\sigma^2}\right); \ \sigma(m) = \sigma_{max} \cdot (1 - m).$$

$$\pi^m(\mathbf{X}, m_{t,i}; t, i) = m_{t,i} \cdot x_{t,i} + (1 - m_{t,i}) \cdot \mu_{t,i}$$

$$\mu_{t,i} = \frac{1}{2W+1} \sum_{t'=t-W}^{t+W} x_{t',i}$$

$$\pi^p(\mathbf{X}, m_{t,i}; t, i) = m_{t,i} \cdot x_{t,i} + (1 - m_{t,i}) \cdot \mu^p_{t,i},$$

$$\mu^p_{t,i} = \frac{1}{W+1} \sum_{t'=t-W}^{t} x_{t',i}$$

Mask optimization

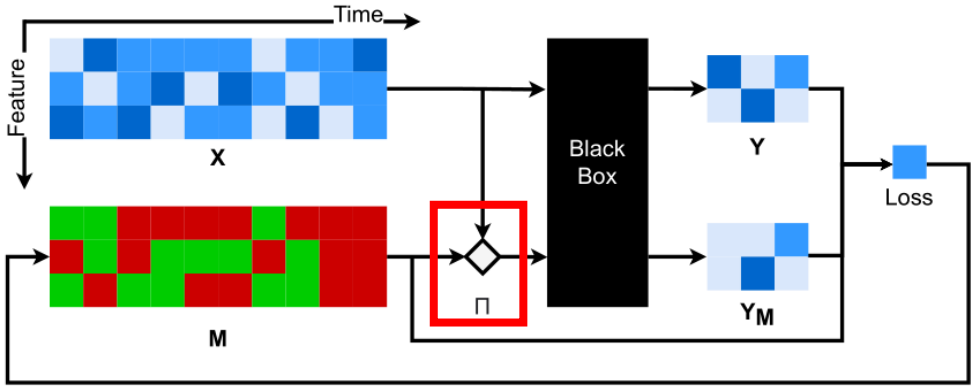


Figure 2. Diagram for Dynamask. An input matrix $\mathbf{X}$, extracted from a multivariate time series, is fed to a black-box to produce a prediction $\mathbf{Y}$. The objective is to give a saliency score for each component of $\mathbf{X}$. In Dynamask, these saliency scores are stored in a mask $\mathbf{M}$ of the same shape as the input $\mathbf{X}$. To detect the salient information in the input $\mathbf{X}$, the mask produces a perturbed version of $\mathbf{X}$ via a perturbation operator Π. This perturbed $\mathbf{X}$ is fed to the black-box to produce a perturbed prediction $\mathbf{Y_M}$. The perturbed prediction is compared to the original prediction and the error is backpropagated to adapt the saliency scores contained in the mask.

1. 应该保持黑盒预测中的变化很小

$$\mathcal{L}_e\left(\mathbf{M}\right) = \sum_{t=t_y}^{T}\sum_{i=1}^{d_Y}\left(\left[\left(f\circ\Pi_{\mathbf{M}}\right)\left(\mathbf{X}\right)\right]_{t,i} - \left[f(\mathbf{X})\right]_{t,i}\right)^2$$

$$\mathcal{L}_e\left(\mathbf{M}\right) = -\sum^{T}\sum^{d_Y}\left[f(\mathbf{X})\right]_{t,c}\log\left[\left(f\circ\Pi_{\mathbf{M}}\right)\left(\mathbf{X}\right)\right]_{t,c}$$

2. 鼓励掩码突出显示输入的一小部分

$$\mathcal{L}_a\left(\mathbf{M}\right) = \left\|\text{vecsort}(\mathbf{M}) - \mathbf{r}_a\right\|^2$$

vector in ascending order. The vector $\mathbf{r}_a$ contains $(1 - a)\cdot d_X \cdot T$ zeros followed by $a\cdot d_X\cdot T$ ones, where $a\in[0,1]$. In short, this regularization term encourages the mask

3. 避免显著性随时间快速变化

$$\mathcal{L}_c(\mathbf{M}) = \sum_{t=1}^{T-1}\sum_{i=1}^{d_X}\left|\, m_{t+1,i} - m_{t,i}\,\right|$$

$$\mathbf{M}_a^* = \underset{\mathbf{M}\in[0,1]^{T\times d_X}}{\arg\min}\ \mathcal{L}_e\left(\mathbf{M}\right) + \lambda_a\cdot\mathcal{L}_a\left(\mathbf{M}\right) + \lambda_c\cdot\mathcal{L}_c\left(\mathbf{M}\right)$$

Masks and information theory

将mask系数解释为相关特征对黑盒发出预测是显著的概率。

1. 测量从时间序列中提取的子序列中包含的信息量

Definition 2.3 (Mask information)**.** The mask information associated to a mask $\mathbf{M}$ and a subsequence $(x_{t,i})_{(t,i)\in A}$ of the input $\mathbf{X}$ with $A \subseteq [1:T] \times [1:d_X]$ is

$$I_{\mathbf{M}}(A) = -\sum_{(t,i)\in A} \ln\left(1 - m_{t,i}\right).$$

As in traditional information theory, the information content is not entirely informative on its own. For instance, consider two subsequences indexed by, respectively, A, B with $|A| = |B| = 10$. We assume that the submask extracted from $\mathbf{M}$ with A contains 3 coefficients $m = 0.9$ and 7 coefficients $m = 0$ so that the information content of A is $I_{\mathbf{M}}(A) \approx 6.9$. Now we consider that all the coefficient extracted from $\mathbf{M}$ with B are equal to 0.5 so that $I_{\mathbf{M}}(B) \approx 6.9$ and hence $I_{\mathbf{M}}(A) \approx I_{\mathbf{M}}(B)$. In this example, A clearly identifies 3 important features while B gives a mixed score for the 10 features. Intuitively, it is pretty clear that the information provided by A is sharper. Unfortunately,

2. 并且需要提供低熵的解释

Definition 2.4 (Mask entropy)**.** The mask entropy associated to a mask $\mathbf{M}$ and a subsequence $(x_{t,i})_{(t,i)\in A}$ of the input $\mathbf{X}$ with $A \subseteq [1:T] \times [1:d_X]$ is

$$S_{\mathbf{M}}(A) = -\sum_{(t,i)\in A} m_{t,i} \ln m_{t,i} + (1 - m_{t,i}) \ln\left(1 - m_{t,i}\right)$$

maximized when mask coefficients $m_{t,i}$ are close to 0.5. In this case, given our probabilistic interpretation, the mask coefficient is ambiguous as it does not really indicate whether the feature is salient. This is consistent with our previous example where $S_{\mathbf{M}}(A) \approx 0.98$ while $S_{\mathbf{M}}(B) \approx 6.93$ so that $S_{\mathbf{M}}(A) \ll S_{\mathbf{M}}(B)$. Since masks coefficients take various values in practice, masks with higher entropy appear less legible, as illustrated in Figure 3. Therefore, we use the

Experiments

提出了三个不同的实验，按照难度的升序排列。

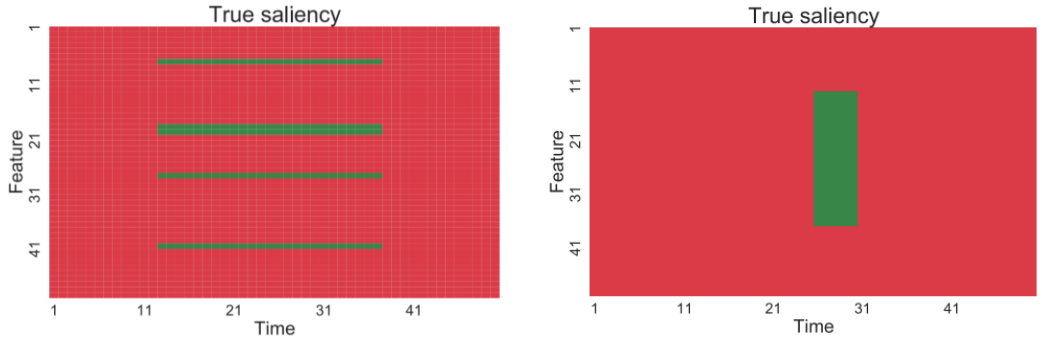
1. 使用了一个具有已知特征重要性的白盒。

2. 使用了在具有已知特征重要性的数据集上训练的黑盒。

3. 使用了在真实临床数据集上训练的黑盒.

Experiments

已知特征重要性的白盒



$$A = A_T \times A_X \subset [1:T] \times [1:d_X]$$

$$[f(\mathbf{X})]_t = \begin{cases} \sum_{i \in A_X} (x_{t,i})^2 & \text{if } t \in A_T \\ 0 & \text{else.} \end{cases}$$

Metrics

1. 测量确实显著的已识别特征的比例:AUP

2. 测量确实已识别的显著特征部分：AUR

3. 测量显著性方法为显著区域预测的信息量:Im

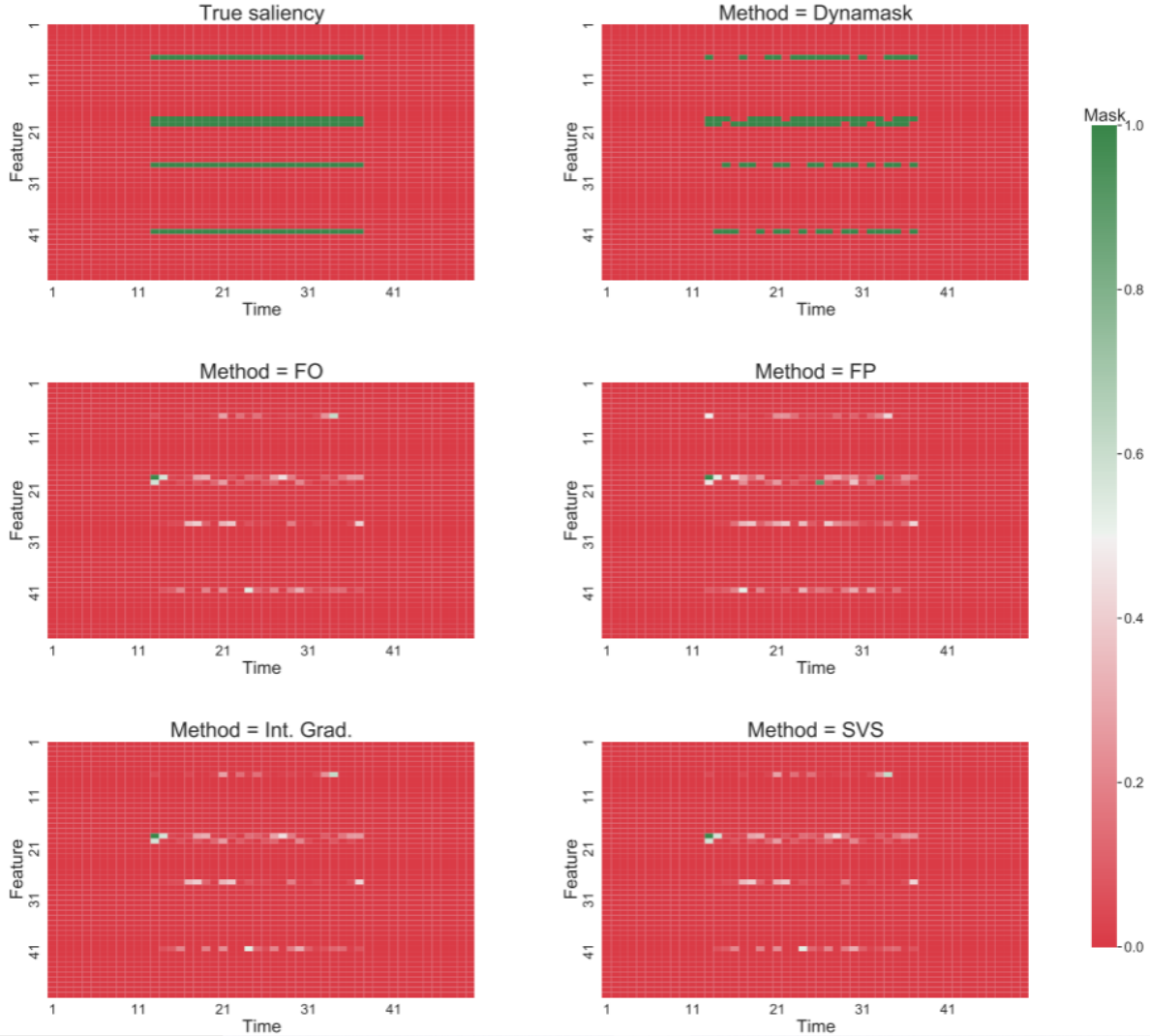
4. 衡量这识别解释区域的混乱程度：Sm

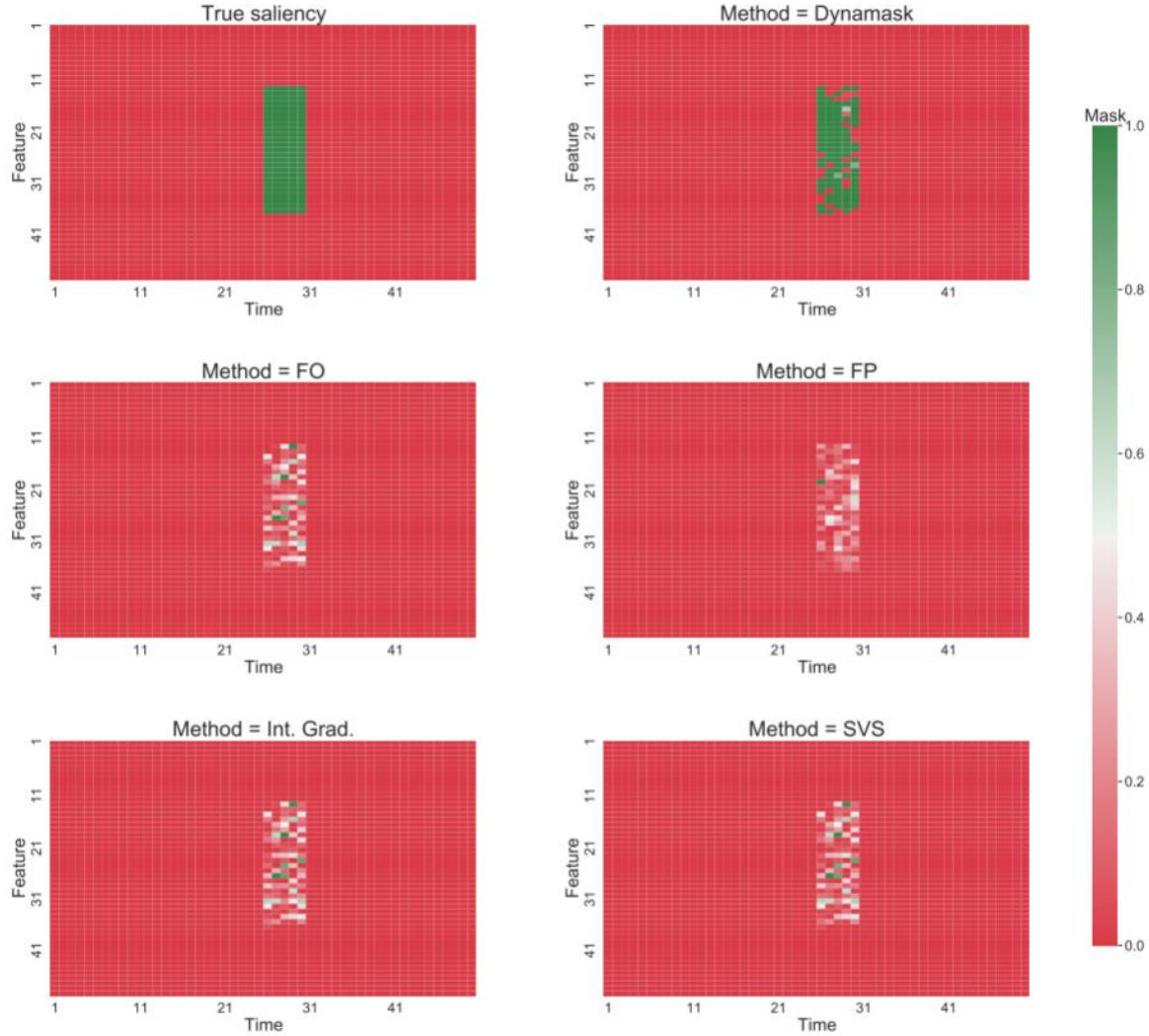
Table 1. Scores for the rare feature experiment.

	AUP	AUR	$I_{\mathbf{M}}(A)$	$S_{\mathbf{M}}(A)$
MASK	0.99 ± 0.01	$\mathbf{0.58 \pm 0.03}$	$\mathbf{252 \pm 69}$	$\mathbf{0.7 \pm 0.7}$
FO	$\mathbf{1.00 \pm 0.00}$	0.14 ± 0.03	9 ± 6	11.0 ± 2.5
FP	$\mathbf{1.00 \pm 0.00}$	0.16 ± 0.04	13 ± 7	12.6 ± 3.3
IG	0.99 ± 0.00	0.14 ± 0.03	8 ± 4	11.1 ± 2.5
SVS	$\mathbf{1.00 \pm 0.00}$	0.14 ± 0.04	9 ± 6	11.0 ± 2.5

Table 2. Scores for the rare time experiment.

	AUP	AUR	$I_{\mathbf{M}}(A)$	$S_{\mathbf{M}}(A)$
MASK	0.99 ± 0.01	$\mathbf{0.68 \pm 0.04}$	$\mathbf{1290 \pm 106}$	$\mathbf{7.1 \pm 2.5}$
FO	$\mathbf{1.00 \pm 0.00}$	0.14 ± 0.04	49 ± 14	48.3 ± 6.5
FP	$\mathbf{1.00 \pm 0.00}$	0.16 ± 0.03	53 ± 8	54.7 ± 5.8
IG	0.99 ± 0.00	0.14 ± 0.04	38 ± 12	48.7 ± 6.7
SVS	$\mathbf{1.00 \pm 0.00}$	0.14 ± 0.04	49 ± 14	48.3 ± 6.5


True saliency
Method = Dynamask
Method = FO
Method = FP
Method = Int. Grad.
Method = SVS
Feature
Time
Mask
1.0
0.8
0.6
0.4
0.2
0.0

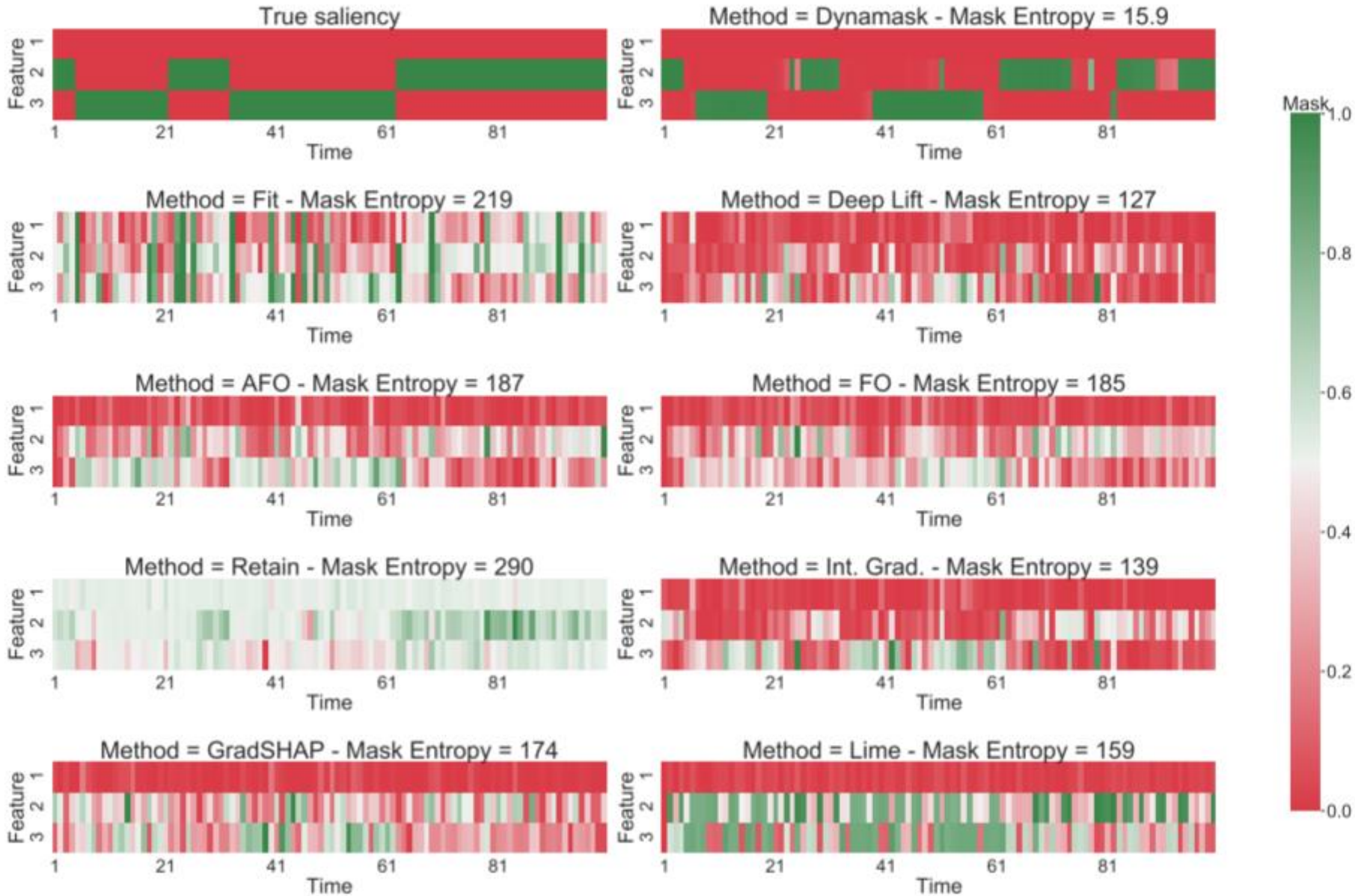

True saliency
Method = Dynamask
Method = FO
Method = FP
Method = Int. Grad.
Method = SVS
Feature
Time
Mask
1.0
0.8
0.6
0.4
0.2
0.0

Experiments

在具有已知特征重要性的数据集上训练的黑盒

Table 3. Scores for the state experiment.

	AUP	AUR	$I_{\mathbf{M}}(A) \div 10^5$	$S_{\mathbf{M}}(A) \div 10^4$
MASK	$\mathbf{0.88 \pm 0.01}$	0.70 ± 0.00	$\mathbf{2.24 \pm 0.01}$	$\mathbf{0.04 \pm 0.00}$
FO	0.63 ± 0.01	0.45 ± 0.01	0.21 ± 0.00	1.79 ± 0.00
AFO	0.63 ± 0.01	0.42 ± 0.01	0.19 ± 0.00	1.76 ± 0.00
IG	0.56 ± 0.00	$\mathbf{0.78 \pm 0.00}$	0.05 ± 0.00	1.39 ± 0.00
GS	0.49 ± 0.00	0.62 ± 0.00	0.33 ± 0.00	1.73 ± 0.00
LIME	0.49 ± 0.01	0.50 ± 0.01	0.04 ± 0.00	1.11 ± 0.00
DL	0.57 ± 0.01	0.20 ± 0.00	0.09 ± 0.00	1.18 ± 0.00
RT	0.42 ± 0.03	0.51 ± 0.01	0.03 ± 0.00	1.75 ± 0.00
FIT	0.44 ± 0.01	0.60 ± 0.02	0.47 ± 0.02	1.57 ± 0.00


True saliency
Method = Dynamask - Mask Entropy = 15.9
Method = Fit - Mask Entropy = 219
Method = Deep Lift - Mask Entropy = 127
Method = AFO - Mask Entropy = 187
Method = FO - Mask Entropy = 185
Method = Retain - Mask Entropy = 290
Method = Int. Grad. - Mask Entropy = 139
Method = GradSHAP - Mask Entropy = 174
Method = Lime - Mask Entropy = 159
Feature
Time
1
21
41
61
81
3
2
1
Mask
1.0
0.8
0.6
0.4
0.2
0.0

Experiments

$$x_{t,i} \mapsto \tilde{x}_{t,i} = \frac{1}{T} \sum_{t=1}^{T} x_{t,i}$$

临床数据

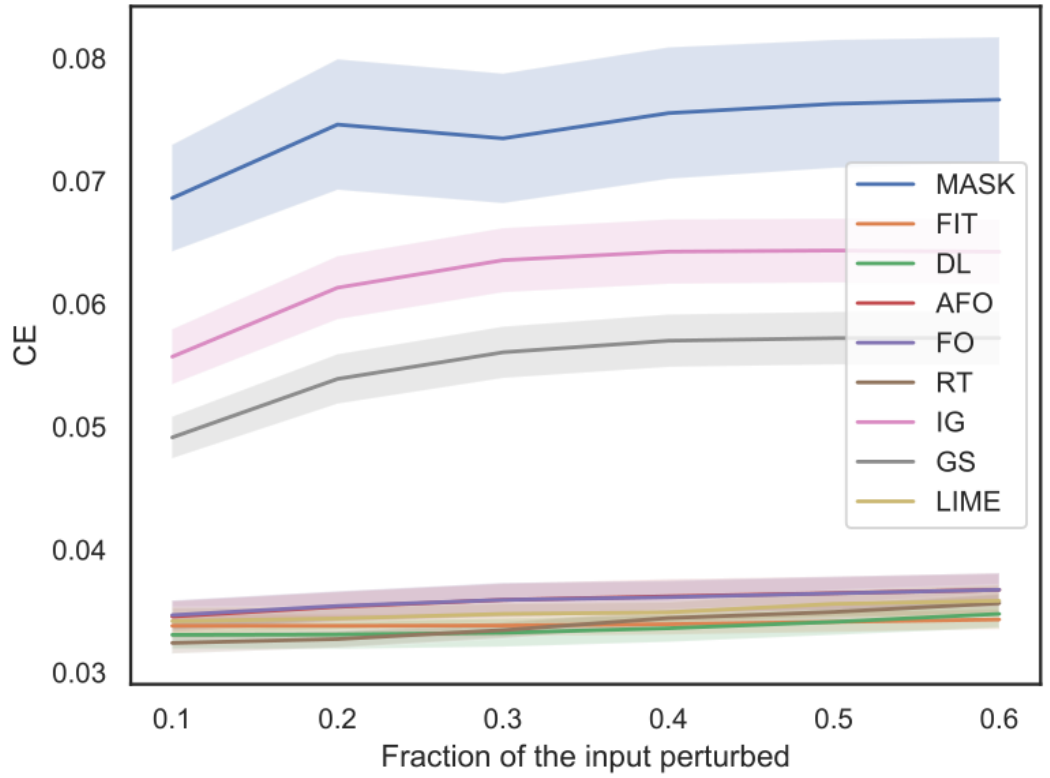


Figure 4. CE as a function of a for the MIMIC experiment.

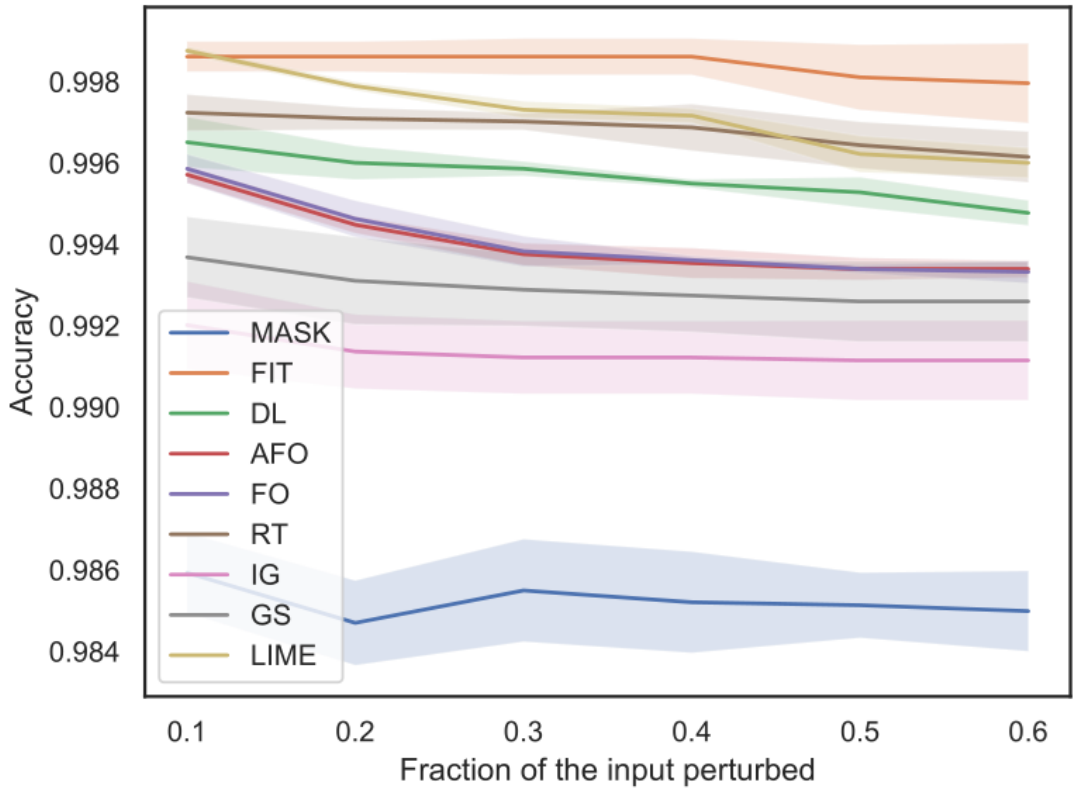


Figure 5. ACC as a function of a for the MIMIC experiment.

replacing the most important observations, we use the *cross-entropy* between $f(\mathbf{X})$ and $f(\tilde{\mathbf{X}})$ (CE, higher is better). To

accuracy of $f(\tilde{\mathbf{X}})$

Learning Saliency Maps to Explain Deep Time Series Classifiers

Prathyush S. Parvatharaju*
Worcester Polytechnic Institute
Worcester, MA, USA
psparvatharaju@wpi.edu

Ramesh Doddaiah*
Worcester Polytechnic Institute
Worcester, MA, USA
rdoddaiah@wpi.edu

Thomas Hartvigsen
Worcester Polytechnic Institute
Worcester, MA, USA
twhartvigsen@wpi.edu

Elke A. Rundensteiner
Worcester Polytechnic Institute
Worcester, MA, USA
rundenst@wpi.edu

挑战与研究动机

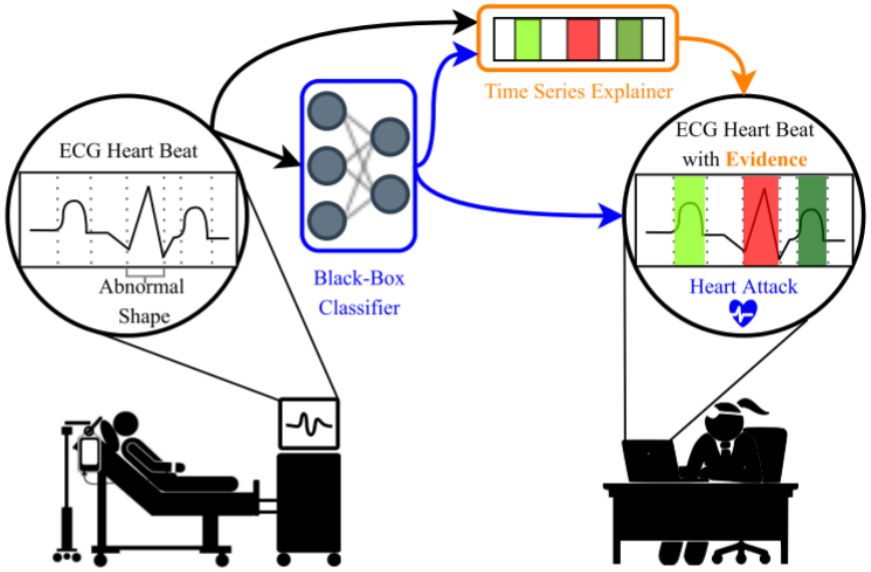


Figure 1: Example of Time Series Saliency Map highlighting input time steps contributing to classification result.

多目标优化问题：一个好的显著性矩阵必须准确地突出最相关的时间步，同时通过找到尽可能少的时间步来保持对最终用户的直观性。

同一数据集下，时间序列是高度可变的。 → 将扰动规则化，以确保扰动与模型之前见过的其它序列相似。

长时间序列的扰动。 → 进行特定兴趣的扰动。

更加直观的解释。 → 将焦点集中。

问题定义

given a set of N time series $\mathcal{D} = \{X^1, \ldots, X^N\}$

a black-box classifier $f_c : \mathbb{X} \rightarrow \mathcal{Y}$

$\theta = [\theta_1, \ldots, \theta_T]$

perturbation function $f_p(X)$

Proposed Method: PERT

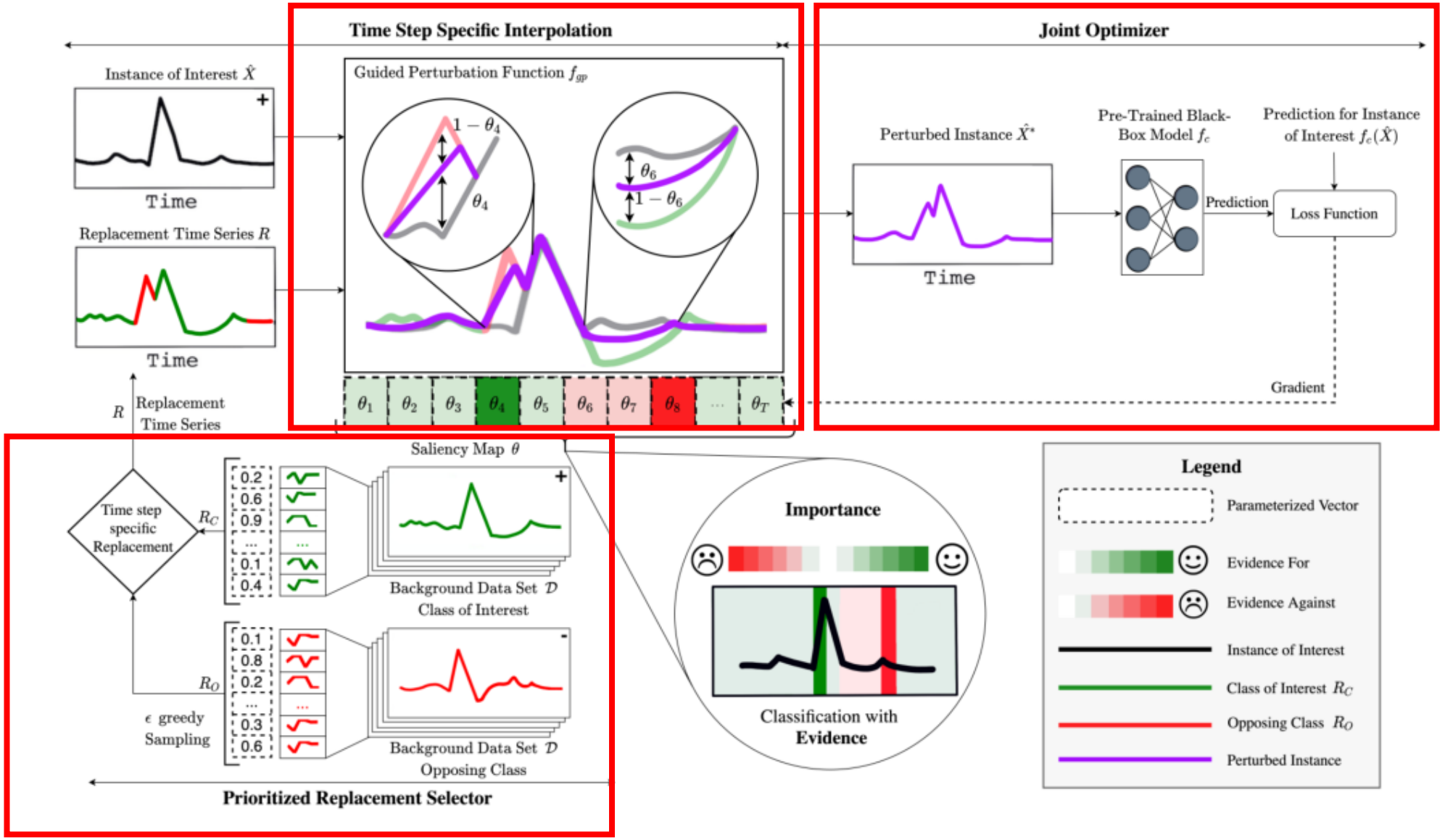


Figure 2: PERT Architecture

Prioritized Replacement Selector

为了扰动时间序列的值，必须选择新的值来替换它们。然而，这种替换在时间序列中是非常有影响的，因为信号的形状和趋势可能会发生巨大的变化。

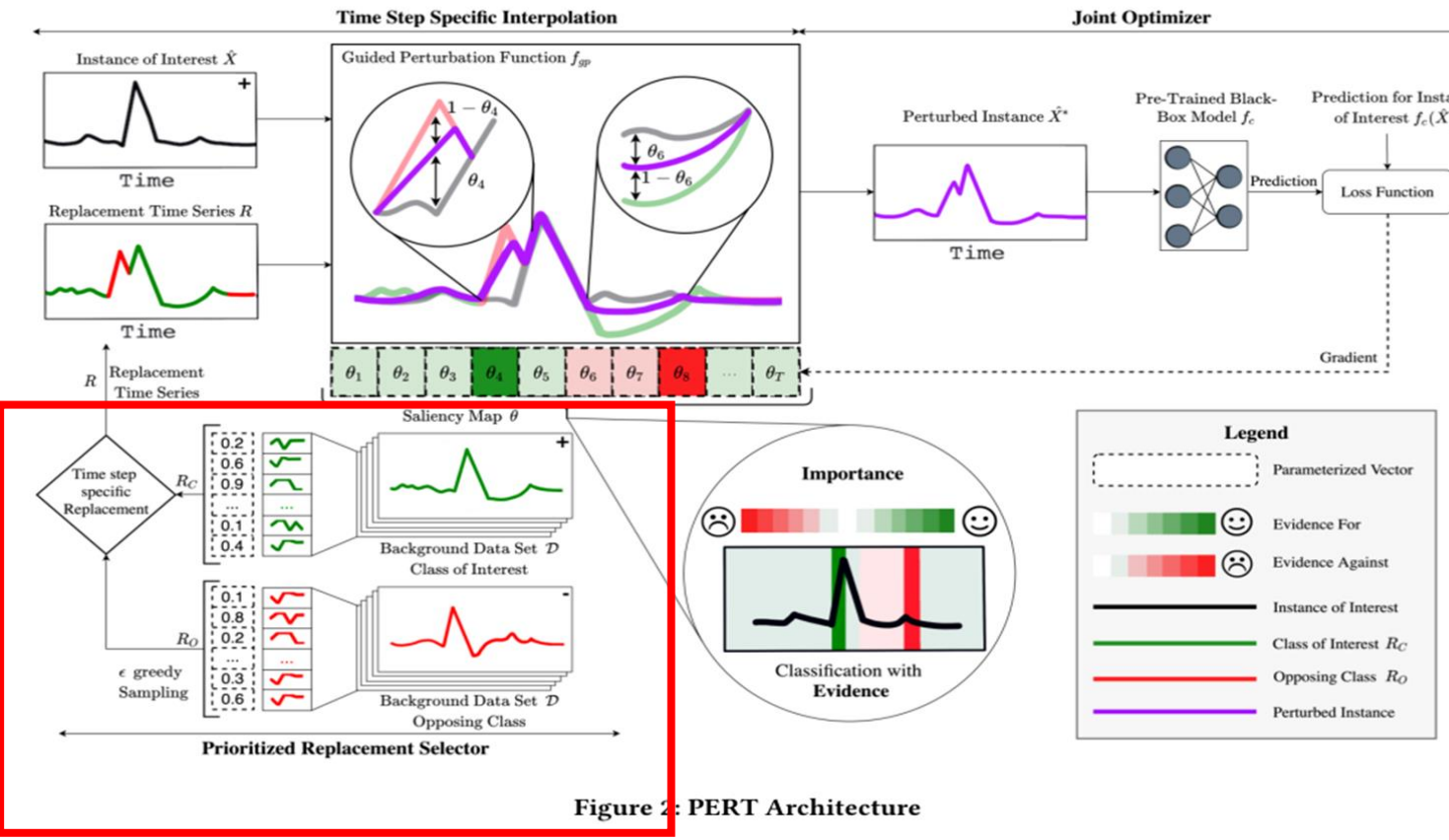


Figure 2: PERT Architecture

To provide evidence both for and against C。

we split $\mathcal{D}$ into two subsets

$$\mathcal{D}_C \qquad \mathcal{D}_O = \mathcal{D} - \mathcal{D}_C$$

$$w_C \qquad w_O$$

$$\mathrm{P}(R_C^i) = w_C^i$$

$$\mathrm{P}(R_O^i) = w_O^i$$

$$R_C = \begin{cases} R_C \text{ (itself)} & \text{with probability } 1-\epsilon \\ \text{random } R_C \in \mathcal{D}_C & \text{with probability } \epsilon \end{cases}$$

Guided Perturbation Function

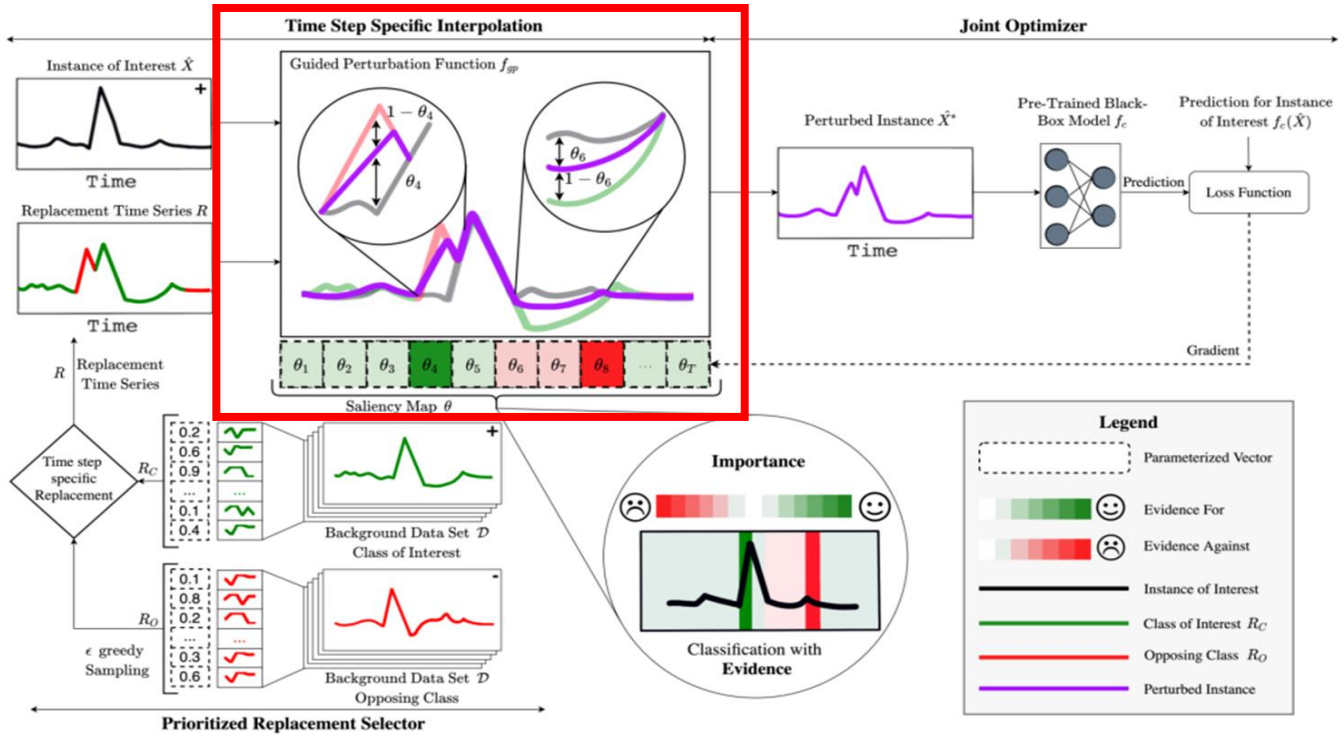


Figure 2: PERT Architecture

$\theta_t < 0$ indicates that its corresponding time step X_t is evidence *against* class C, we replace the time step with R_C, the representative of class C. Similarly, when $\theta_t \geq 0$, X_t is replaced with the corresponding timestep of R_O. This way, PERT learns the degree of sensitivity of each time step. The function f_p generates its final perturbation $\tilde{X}$ by performing *timestep-specific interpolation*:

$$\tilde{X} = \theta \odot X + (1 - \theta) \odot (\mathbb{1}_{\theta<0} \odot R_C + \mathbb{1}_{\theta\geq 0} \odot R_O) + g \qquad (2)$$

prediction. Higher values of θ indicate stronger evidence *for* class C while lower values are evidence *against* class C

Optimizing PERT

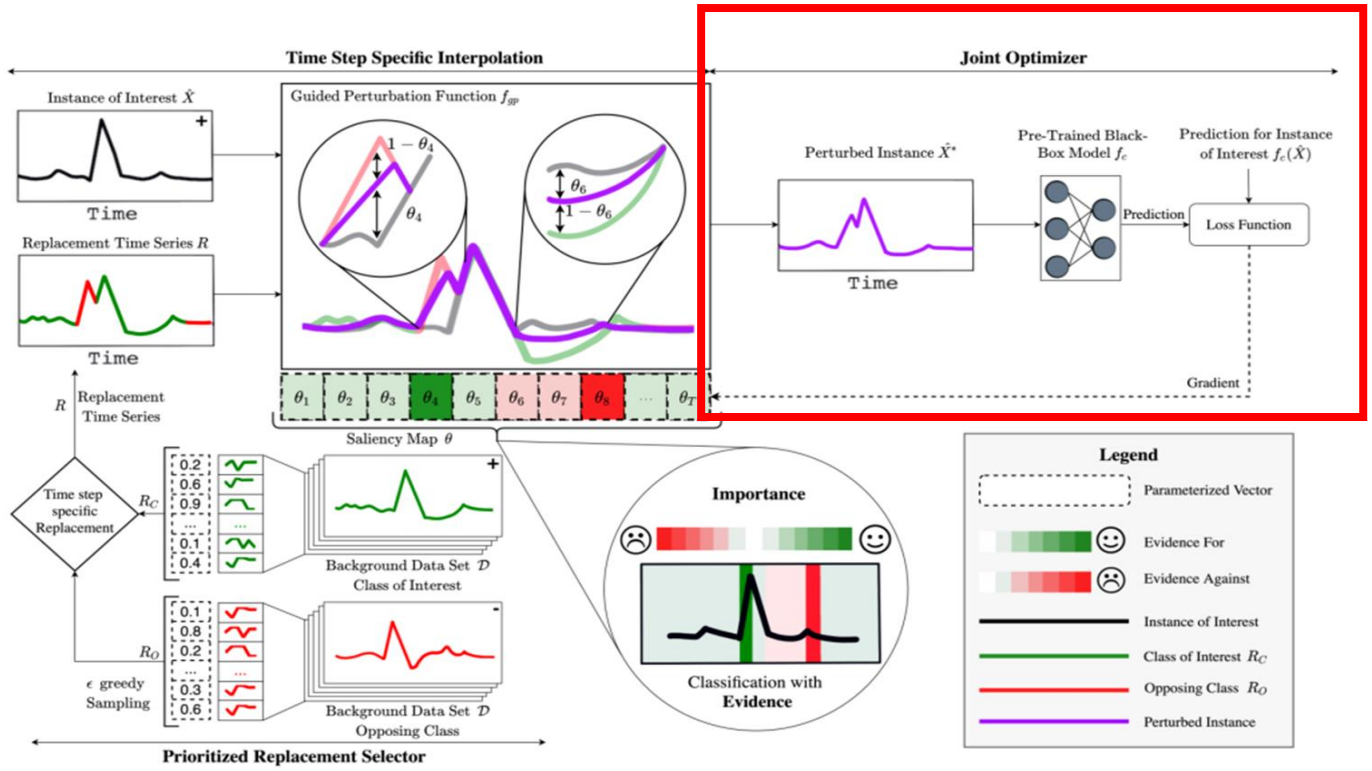


Figure 2: PERT Architecture

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

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

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

The final loss for PERT is shown in Equation 6.

$$L(P(\hat{X}); \theta) = (L_{\text{preservation}} + L_{\text{budget}} + L_{\text{TV}}) * \frac{1}{2}(w_O + w_C)$$

Datasets

We evaluate our method on nine real-world time-series datasets: Wafer [26], GunPoint [30], Computers [5], Earthquakes [5], FordA [5], FordB [5], CricketX [5], PTB [11], ECG [26].

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.

Baseline

- *Random.* Each time step is assigned a random saliency value between -1 to 1 from a uniform distribution. This approach serves as a baseline for all methods. 随机生成
- *RISE* [29]. The partial derivative of the black-box model's prediction $P(C|\hat{X})$ with respect to each time step is estimated empirically by randomly setting timesteps to zero and summarizing its impact on the $P(C|\hat{X})$. 导数
- *LEFTIST* [12]. The partial derivative of $P(C|\hat{X})$ with respect to each time step is estimated empirically by randomly *replacing* the timestep with a corresponding value from a random instance from the background dataset and summarizing its impact on $P(C|\hat{X})$. 随机扰动
- *LIME* [32]. Saliency values are derived from the coefficients of a linear surrogate model, trained to mimic the behavior of the black-box model in the feature space surrounding $\hat{X}$. The success of this approach relies on the model behaving linear locally, which is rarely guaranteed in practice.
- *SHAP* [21]. SHAP assigns Shapley values [4] to each time step, thus computing their contributions to $P(C|\hat{X})$. Each time step is replaced by every value observed at the corresponding time step across all other instances in the background dataset.
- *Meaningful Perturbation (MP)* [9]. MP learns to perturb each time step such that $P(C|\hat{X})$ decreases. Perturbation is achieved by combining squared exponential smoothing with additive gaussian noise. Saliency values are then learned iteratively and are ultimately used as the final explanation. 扰动

Metrics

AUC-Difference. Saliency maps can be evaluated by "inserting" or "deleting" timesteps from the time-series instance based on the derived importance map and observing the changes in the black-box model's predictions [29].

根据重要性图，通过从时间序列实例中"插入"或"删除"时间步长，并观察黑盒模型预测中的变化

Confidence Suppression Game. Another popular approach to evaluating saliency maps is to find the smallest number of timesteps required to suppress the confidence of the black-box model by a given percentage [9]. We thus ask each explanation to play this

找到将黑盒模型的置信度抑制到给定百分比所需的最小时间步数

Minimality of Saliency Maps. A good explanation returns *only* the most important timesteps, thus allowing a user to consider only a few regions of the input. Thus we compute the *sum* of a saliency map, under the intuition that it should be small, as the goals of explainability and simplicity are jointly optimized, we

一个好的解释只会返回最重要的时间步，因此用户只需考虑输入的几个区域---》更好地直观性（高方差、低sum）

Experimental Results

Methods	Datasets								
	Wafer	GunPoint	Computers	Earthquakes	FordA	FordB	CricketX	PTB	ECG
Random	-0.02 (.01)	0.02 (.01)	0.01 (.01)	-0.01 (.01)	-0.03 (.01)	-0.01(.01)	-0.01 (.01)	0.06 (.04)	0.01 (.06)
RISE [29]	0.22 (.01)	0.16 (.01)	-0.01 (.02)	0.23 (.05)	0.15 (.02)	0.11 (.01)	0.42 (.01)	0.07 (.05)	0.14 (.07)
LEFTIST [12]	0.53 (.02)	0.15 (.03)	-0.16 (.01)	0.15 (.03)	0.16 (.01)	0.15 (.01)	-0.01 (.01)	0.51 (.01)	0.55 (.01)
LIME [32]	0.06 (.01)	0.10 (.01)	0.06 (.03)	-0.01 (.01)	0.01 (.02)	0.01 (.01)	-0.01 (.01)	0.18 (.07)	0.09 (.06)
SHAP [21]	-0.19 (.01)	-0.01 (.01)	0.10 (.01)	0.71 (.03)	0.23 (.01)	-0.17 (.01)	0.09 (.01)	-0.15 (.01)	-0.11 (.09)
MP [9]	0.49 (.01)	0.03 (.01)	0.15 (.01)	0.33 (.01)	0.48 (.01)	0.37 (.01)	0.43 (.01)	0.33 (.01)	-0.16 (.00)
PERT	**0.74 (.01)**	**0.56 (.01)**	**0.95 (.01)**	**0.93 (.01)**	**0.83 (.01)**	**0.82 (.01)**	**0.69 (.01)**	**0.58 (.01)**	**0.60 (.01)**

Table 2: Performance of the AUC-difference metric with the FCN black-box model. Parentheses indicate σ. Compared methods are separated into four groups: Random perturbation, linear surrogate model, game theory method, and learned perturbations.

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 [29]	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 [12]	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 [32]	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 [21]	-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 [9]	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.

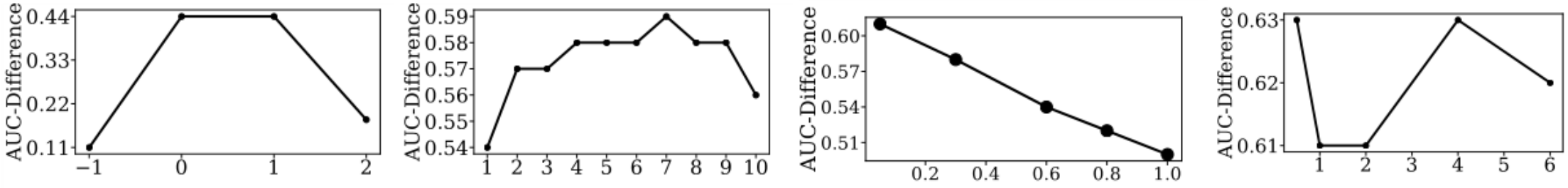
根据重要性图，通过从时间序列实例中“插入”或“删除”时间步长，并观察黑盒模型预测中的变化

找到将黑盒模型的置信度抑制到给定百分比所需的最小时间步数

		Confidence suppression game ↓										Minimality of Saliency	
Dataset	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	Sum (%) ↓	$\|\|Variance\|\|$ ↑
Wafer [26]	Random	0.79	0.82	0.86	0.87	0.87	0.88	0.89	0.91	0.91	0.92	50	0.0016
	RISE [29]	0.57	0.57	0.58	0.59	0.59	0.59	0.61	0.61	0.61	0.61	82	0.0001
	LEFTIST [12]	0.56	0.57	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.59	78	0.0001
	LIME [32]	0.69	0.73	0.82	0.82	0.83	0.84	0.84	0.86	0.86	0.88	42	0.0001
	SHAP [21]	0.78	0.78	0.78	0.78	0.78	0.78	0.79	0.79	0.79	0.79	52	0.0004
	MP [9]	0.41	0.47	0.53	0.54	0.55	0.56	0.57	0.58	0.68	0.73	6	0.0041
	PERT	**0.40**	**0.40**	**0.46**	**0.47**	**0.48**	**0.49**	**0.51**	**0.53**	**0.55**	**0.59**	**1**	**0.0153**
ECG [5]	Random	0.40	0.41	0.41	0.43	0.43	0.43	0.44	0.45	0.45	0.45	50	0.0001
	RISE [29]	0.40	0.40	0.40	0.40	0.40	0.41	0.41	0.41	0.41	0.41	76	0.0001
	LEFTIST [12]	0.30	0.30	0.30	0.30	0.30	0.31	0.31	0.31	0.31	0.31	60	0.0001
	LIME [32]	0.71	0.72	0.74	0.77	0.79	0.81	0.81	0.84	0.85	0.64	42	0.0001
	SHAP [21]	0.62	0.62	0.62	0.63	0.63	0.63	0.63	0.64	0.64	0.64	64	0.0001
	MP [9]	0.51	0.51	0.51	0.52	0.52	0.52	0.52	0.52	0.53	0.53	3	0.0005
	PERT	**0.21**	**0.21**	**0.21**	**0.21**	**0.21**	**0.22**	**0.22**	**0.22**	**0.22**	**0.22**	**2**	**0.0034**
Computers [5]	Random	0.69	0.73	0.75	0.77	0.79	0.81	0.83	0.85	0.90	0.95	50	0.0001
	RISE [29]	0.92	0.92	0.93	0.93	0.93	0.94	0.94	0.94	0.94	0.94	46	0.0001
	LEFTIST [12]	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	49	0.0001
	LIME [32]	0.60	0.62	0.65	0.68	0.7	0.71	0.73	0.75	0.78	0.82	23	0.0001
	SHAP [21]	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	51	0.0001
	MP [9]	0.75	0.76	0.78	0.81	0.81	0.82	0.82	0.83	0.84	0.88	4	0.0005
	PERT	**0.50**	**0.50**	**0.50**	**0.51**	**0.51**	**0.51**	**0.51**	**0.51**	**0.51**	**0.51**	**2**	**0.0017**
Earthquakes [5]	Random	0.97	0.97	0.98	0.98	0.99	0.99	0.99	0.99	1.00	1.00	50	0.0001
	RISE [29]	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	74	0.0001
	LEFTIST [12]	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	61	0.0001
	LIME [32]	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.99	43	0.0001
	SHAP [21]	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	51	0.0001
	MP [9]	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	4	0.0001
	PERT	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**2**	**0.0039**
Ford-A [5]	Random	0.75	0.78	0.82	0.84	0.86	0.88	0.89	0.91	0.92	0.95	50	0.0001
	RISE [29]	0.76	0.76	0.76	0.76	0.76	0.76	0.77	0.77	0.77	0.78	55	0.0001
	LEFTIST [12]	0.78	0.78	0.79	0.80	0.81	0.82	0.83	0.84	0.84	0.85	49	0.0001
	LIME [32]	0.71	0.75	0.78	0.81	0.83	0.84	0.86	0.87	0.89	0.92	44	0.0001
	SHAP [21]	0.93	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	47	0.0001
	MP [9]	0.66	0.67	0.70	0.72	0.75	0.77	0.78	0.81	0.83	0.87	3	0.0010
	PERT	**0.63**	**0.64**	**0.65**	**0.66**	**0.66**	**0.67**	**0.67**	**0.68**	**0.69**	**0.72**	**2**	**0.0012**
Ford-B [5]	Random	0.73	0.74	0.77	0.79	0.81	0.83	0.84	0.85	0.86	0.88	50	0.0001
	RISE [29]	0.73	0.81	0.85	0.86	0.86	0.86	0.86	0.86	0.87	0.87	48	0.0001
	LEFTIST [12]	0.72	0.73	0.75	0.77	0.80	0.82	0.83	0.84	0.85	0.86	49	0.0001
	LIME [32]	0.71	0.72	0.74	0.77	0.79	0.81	0.81	0.84	0.85	0.87	30	0.0001
	SHAP [21]	0.71	0.72	0.74	0.75	0.77	0.81	0.81	0.82	0.83	0.84	40	0.0009
	MP [9]	0.71	0.71	0.73	0.74	0.76	0.77	0.78	0.79	0.82	0.83	4	0.0008
	PERT	**0.70**	**0.71**	**0.71**	**0.72**	**0.73**	**0.73**	**0.74**	**0.74**	**0.76**	**0.77**	**2**	**0.0010**
CricketX [24]	Random	0.21	0.27	0.35	0.41	0.49	0.61	0.72	0.89	0.93	1.00	50	0.0001
	RISE [29]	0.15	0.15	0.26	0.35	0.45	0.55	0.63	0.91	0.91	0.95	74	0.0001
	LEFTIST [12]	0.20	0.26	0.29	0.29	0.30	0.50	0.50	0.50	0.50	0.50	51	0.0001
	LIME [32]	0.14	0.19	0.23	0.26	0.29	0.31	0.34	0.37	0.41	0.48	39	0.0001
	SHAP [21]	0.19	0.29	0.33	0.36	0.38	0.39	0.40	0.41	0.43	0.45	33	0.0005
	MP [9]	0.09	0.11	0.12	0.15	0.17	0.18	0.19	0.21	0.21	0.25	2	**0.0040**
	PERT	**0.06**	**0.07**	**0.07**	**0.08**	**0.08**	**0.09**	**0.10**	**0.11**	**0.12**	**0.15**	**1**	**0.0037**

Table 4: Confidence suppression metric performance for seven key datasets with the FCN black-box model. Lower values are better for Saliency sum and Confidence suppression game, Higher values are better for Saliency variance. ↓ indicates *the lower the better.* ↑ indicates *the higher the better.*

一个好的解释只会返回最重要的时间步，因此用户只需考虑输入的几个区域---》更好地直观性（高方差、低sum）



(a) Mean g_n. (b) $L_{\text{preservation}}$ coefficient (λ_1). (c) L_{budget} coefficient (λ_2). (d) L_{TV} coefficient (λ_3).

Figure 4: PERT Hyperparameter Study

Figure 5: CricketX "No-Ball" Class Case Study. The signal is shown in black. Important time steps for the class-of-interest are shown in bright green and while bright red indicates important time steps for the opposing class.

坏球(no ball)、偏球(wide)

左手信号的结束（一旦裁判的手抬起并稳定下来）是支持模型决策的最有力证据。
亮红色的条形图表明，如果这是一个偏球（相反的类），则表明存在相反类的有力证据。