A Model-Agnostic Approach to Quantifying the Informativeness of Explanation Methods for Time Series Classification

Thu Trang Nguyen, Thach Le Nguyen, Georgiana Ifrim

School of Computer Science, University College Dublin, Ireland

Science Foundation Ireland Centre for Research Training in Machine Learning

Introduction

- Time Series Classification (TSC)
    - Prediction task common in many real-life applications, especially Human Activity Recognition tasks; often requires **explanation** for the algorithm's prediction

Landing Classes:

- Normal
- Bending
- Stumble

Science Foundation Ireland Centre for Research Training in Machine Learning

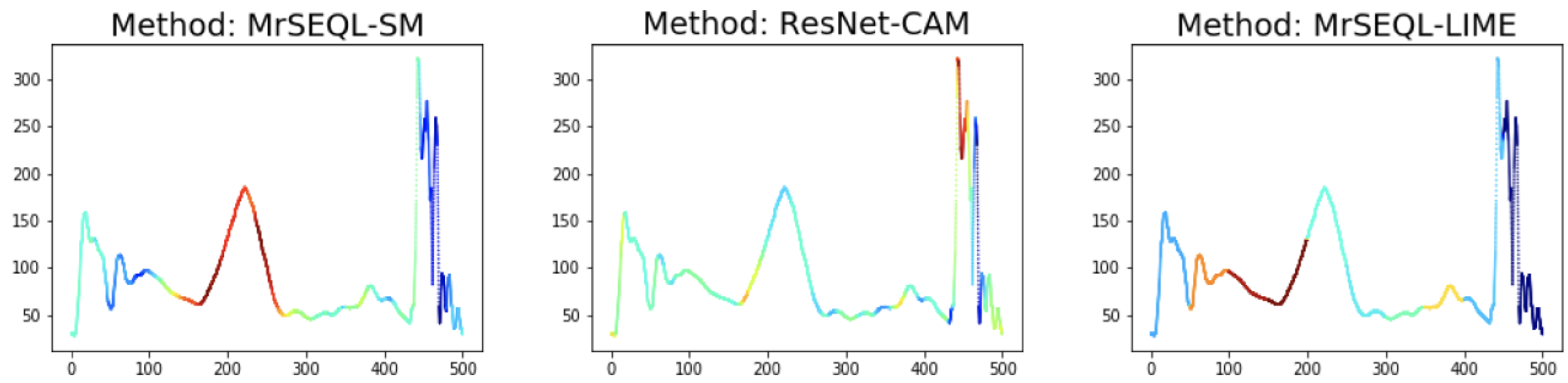


Fig. 1: Saliency map explanations for a motion time series obtained using different explanation methods. In this figure, the most discriminative parts are colored in deep red and the most non-discriminative parts are colored in deep blue.

⇒ Challenge: ***How to assess and objectively compare TSC explanation methods?***

Related Work

- We focus on **quantitative assessment of explanations** for TSC

- We use saliency-based explanations produced by the following methods:
  - MrSEQL-SM: Saliency Map computed from MrSEQL linear classifier weights [2]
  - CAM: Class Activation Map (explaining FCN/ResNet models) [3]
  - LIME: Local Interpretable Model-Agnostic Explanations (explaining any models) [1]

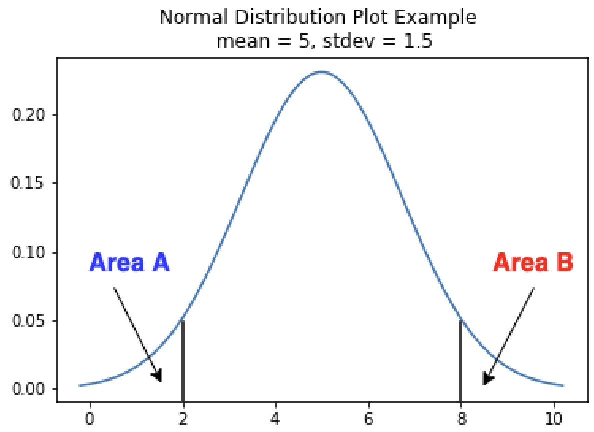
Research Methods

- **Key Concepts:**
  - **Explanation as a Saliency Map**: produced by matching a time series with a vector of weights (explanation) using a heatmap → **highlight the discriminative parts** of the time series
  - **Referee Classifiers**: independent TS classifiers to evaluate the explanation
  - **Explanation Informativeness:** via explanation-based data perturbation, a more informative explanation can more effectively impact the referee classifiers predictions

→ **Key idea:** If the explanation is informative, knocking-off (perturbing) the discriminative parts of the time series leads to lower accuracy for the referee classifier

Research Methods

- **Discriminative vs. Non-discriminative parts of the Time Series**
  - Each time series index has a corresponding saliency weight
  - Discriminative/Non-discriminative parts: indices of the TS with weights in the top/bottom **k%** of the entire weight profile.
    - Example: with k = 20, discriminative parts are the parts of the TS in the top 20% of the weights (index 5 and 6), non-discriminative parts are those that belong in the bottom 20% of the weights (index 1 and 4). The perturbation threshold k varies, eg 0%, 10%, 20%,...,100%.



Index (0-10)	**1**	2	3	**4**	**5**	**6**	7	8	9	10
Weight (in range [0,100])	9	58	46	15	75	78	57	48	36	17

Science Foundation Ireland For what's next

DCU

TU Dublin

UCD Dublin

Science Foundation Ireland Centre for Research Training in Machine Learning

ML Labs

Research Methods

- **Explanation-driven Data Perturbation**
  - **Type1:** noise added to only discriminative parts (with different perturbation level k)
  - **Type2:** noise added to only non-discriminative parts (with different perturbation level k)

- Perturbation: adding Gaussian noise to the original signal

$$x_{perturbed} = x + \mathcal{N}(\mu, \sigma^2)$$

If a time series is normalized, the distribution for the Gaussian noise would be sampled from $\mathcal{N}(0, \sigma_1)$. The parameter σ_1 controls the magnitude of the noise.

For k = 20, we perturb 20% of the time series

Index	**1**	2	3	**4**	**5**	**6**	7	8	9	10
Weight	9	58	46	15	75	78	57	48	36	17
x	224	420	465	222	257	405	383	439	350	450
$x_{perturbed}$ (**type1**)	224	420	465	222	**258**	**400**	383	439	350	450
$x_{perturbed}$ (**type2**)	**220**	420	465	**229**	257	405	383	439	350	450

Research Methods

Quantifying the Informativeness of Explanation Methods

Process:

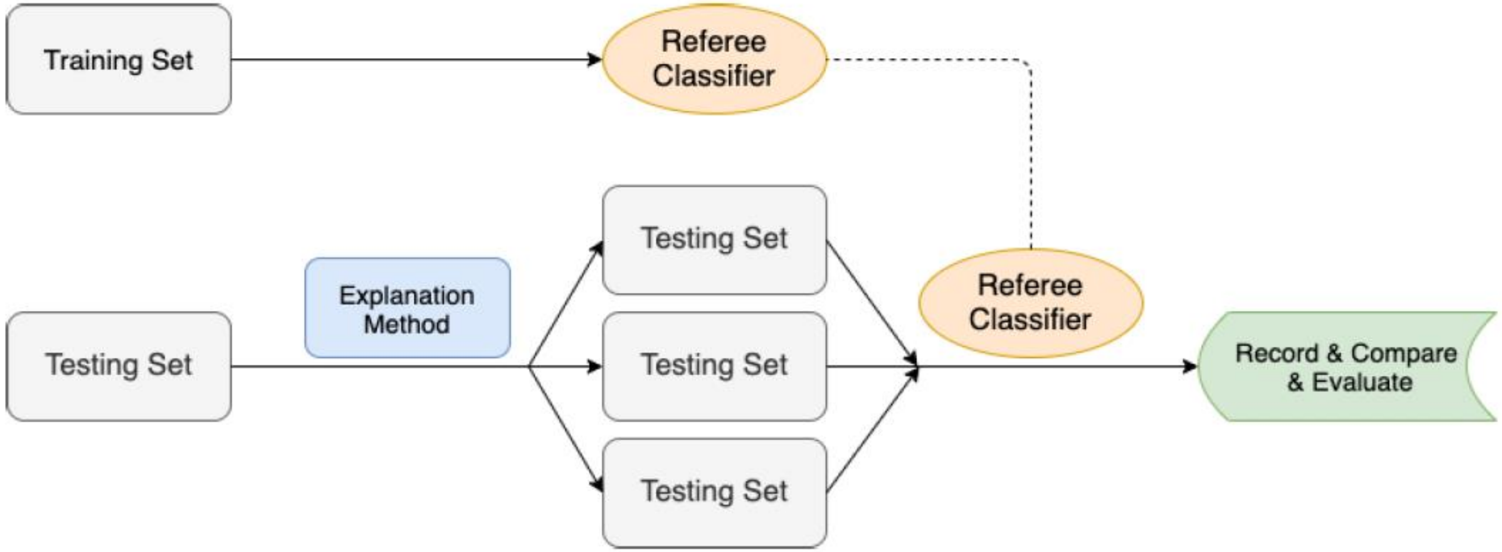
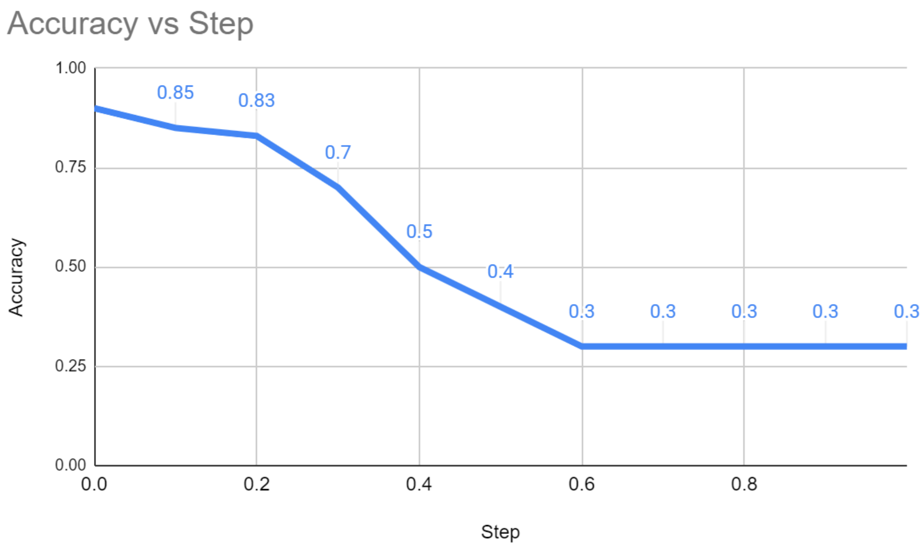


Fig. 4: Process of creating explanation-driven perturbed test sets and evaluating the explanation method using a referee classifier.

Research Methods

Threshold for discriminative weights	0%	10%	20%	...	90%	100%
Classification accuracy by referee classifier	0.90	0.85	0.83	...	0.30	0.30



Research Methods

Quantifying the Informativeness of Explanation Methods

- **Evaluation Measure**
  - Measure the impact of the accuracy reduction induced by different explanation methods by estimating the area under the (explanation-driven) accuracy curve
  - Method: use trapezoidal rule
  - Proposed Metric: ***eLoss***

$$eLoss = \frac{1}{2}k\sum_{i=1}^{t}(acc_{i-1} + acc_i)$$

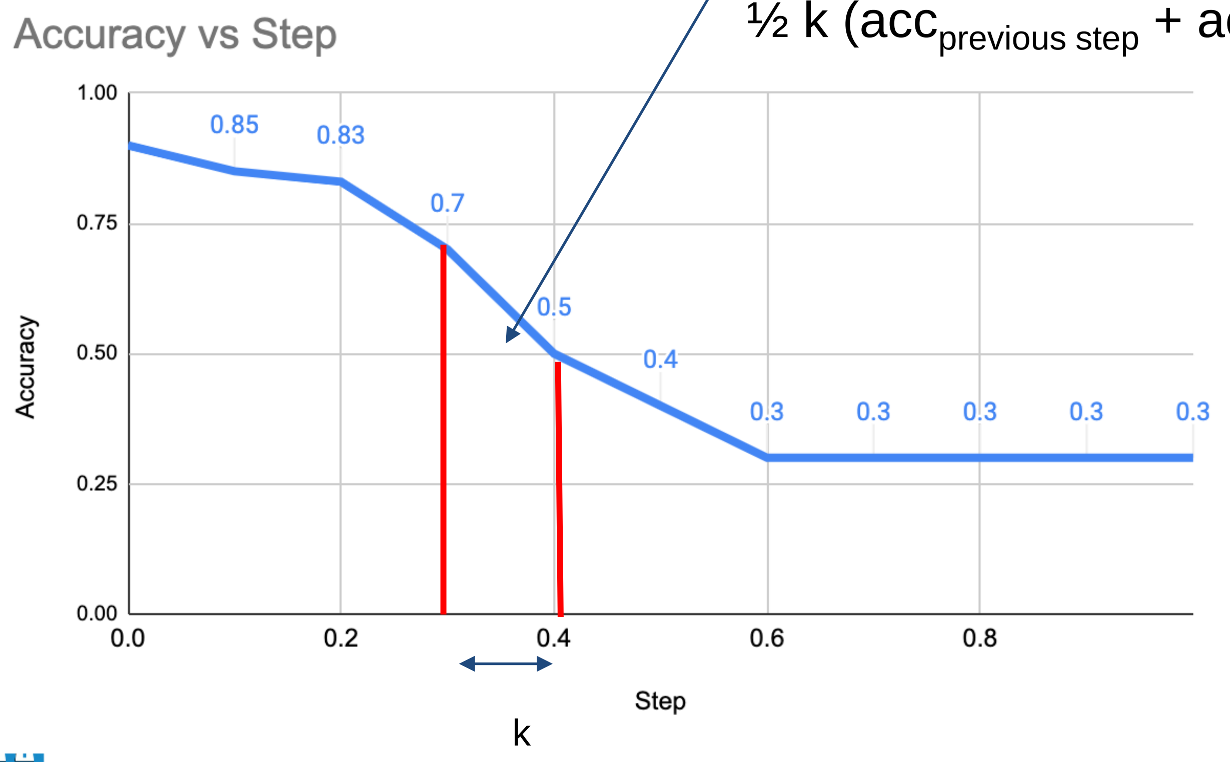
k	- value of each step between the 0-1 range
t	- number of steps (*100/k*)
acc_i	- the accuracy at step i, measured by a referee classifier

Research Methods

Quantifying the Informativeness of Explanation Methods

- **Evaluation Measure**

Area of the red-bounded trapezoid:
$\frac{1}{2}$ k (acc$_{\text{previous step}}$ + acc$_{\text{this step}}$)



Science Foundation Ireland Centre for Research Training in Machine Learning

Research Methods

- Evaluating Explanations:
  - **One Explanation:**
    - For a set of thresholds from 0-100, identify the **discriminative** and **non-discriminative** parts. Perturb these parts of the test time series.
    - If the explanation method is informative, the accuracy (measured by a referee classifier) drops more when the discriminative parts are perturbed.
    - Method is informative when Type1 eLoss ($eLoss_1$) is less than Type2 eLoss ($eLoss_2$)
    - ΔeLoss >0: $\Delta_{eLoss} = eLoss_2 - eLoss_1.$

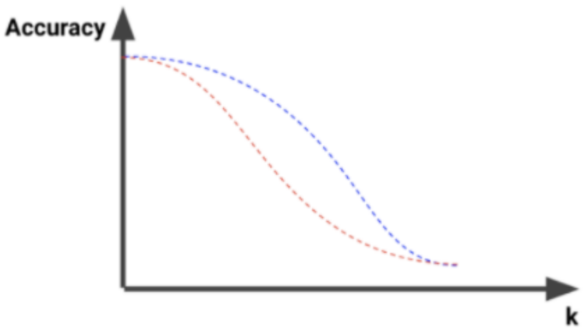


Figure: Change of accuracy when the test set is perturbed with a threshold k.

Research Methods

- Evaluating Explanations:
  - **Multiple Explanations:**
    - For set of thresholds from 0-100, identify **only the discriminative** parts. Perturb these parts of the test time series.
    - Most informative explanation leads to most accuracy drop (measured by a referee classifier), when the discriminative parts are perturbed.
    - Most informative method has lowest $eLoss_1$
    - Compare $eLoss_1$ of the methods under investigation

Experiments

- Datasets

Table 2: Summary of TSC datasets used to evaluate explanation methods.

Dataset	Train Size	Test Size	Length	Type	No. Classes
CBF	30	900	128	Simulated	3
CMJ	419	179	500	Motion	3
Coffee	28	28	286	SPECTRO	2
ECG200	100	100	96	ECG	2
GunPoint	50	50	150	Motion	2

- Explanation Methods:
  - MrSEQL-SM
  - ResNet-CAM
  - MrSEQL-LIME
- Referee Classifiers:
  - MrSEQL
  - ROCKET
  - WEASEL

Experiments

Evaluate Single Method:

- Type1 curve in red, Type2 curve in blue
- Each row shows an explanation method and the accuracy of 3 referee classifiers for different levels of Type1 and Type2 noise
- Explanation method is informative when the red curve is ***below*** the blue curve (loss in accuracy due to the explanation)

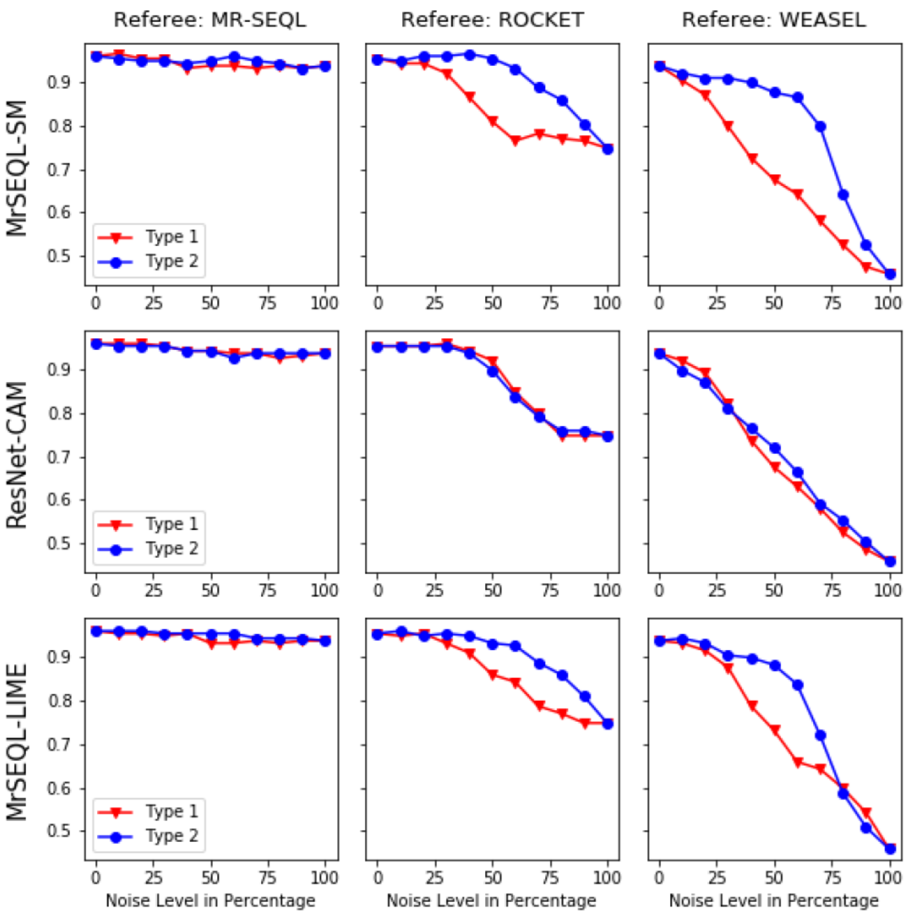


Fig. 7: Comparison of accuracy for *Type 1* (red) and *Type 2* (blue) perturbation for each explanation method and referee classifier for the CMJ dataset.

Experiments

Evaluate Single Method:

Table 3: Summary of Δ_{eLoss} of three explanation methods on five different TSC problems. Positive values suggest the findings of the explanation method are informative according to the referee classifier. Negative values suggest otherwise.

Dataset	Explanation Method	Referee Classifier		
		Mr-SEQL	ROCKET	WEASEL
CBF	MrSEQL-SM	0.0001	0.002	0.0126
	ResNet-CAM	**-0.0005**	0.0007	0.0141
CMJ	MrSEQL-SM	0.0045	0.0709	0.1151
	ResNet-CAM	**-0.0006**	**-0.0028**	0.0106
	MrSEQL-LIME	0.0084	0.0475	0.0531
Coffee	MrSEQL-SM	0.0286	0.0	0.0
	ResNet-CAM	0.0179	0.0	0.0143
ECG200	MrSEQL-SM	0.033	**-0.001**	0.024
	ResNet-CAM	**-0.011**	**-0.003**	0.038
GunPoint	MrSEQL-SM	0.0026	0.1373	0.0273
	ResNet-CAM	0.0067	0.0967	**-0.002**
	MrSEQL-LIME	0.002	0.0714	0.0007

Method is informative when ΔeLoss >0

Experiments

Evaluate Multiple Methods:

- Most informative method has the lowest (most impacted) explanation curve plotted by accuracy of referee classifier

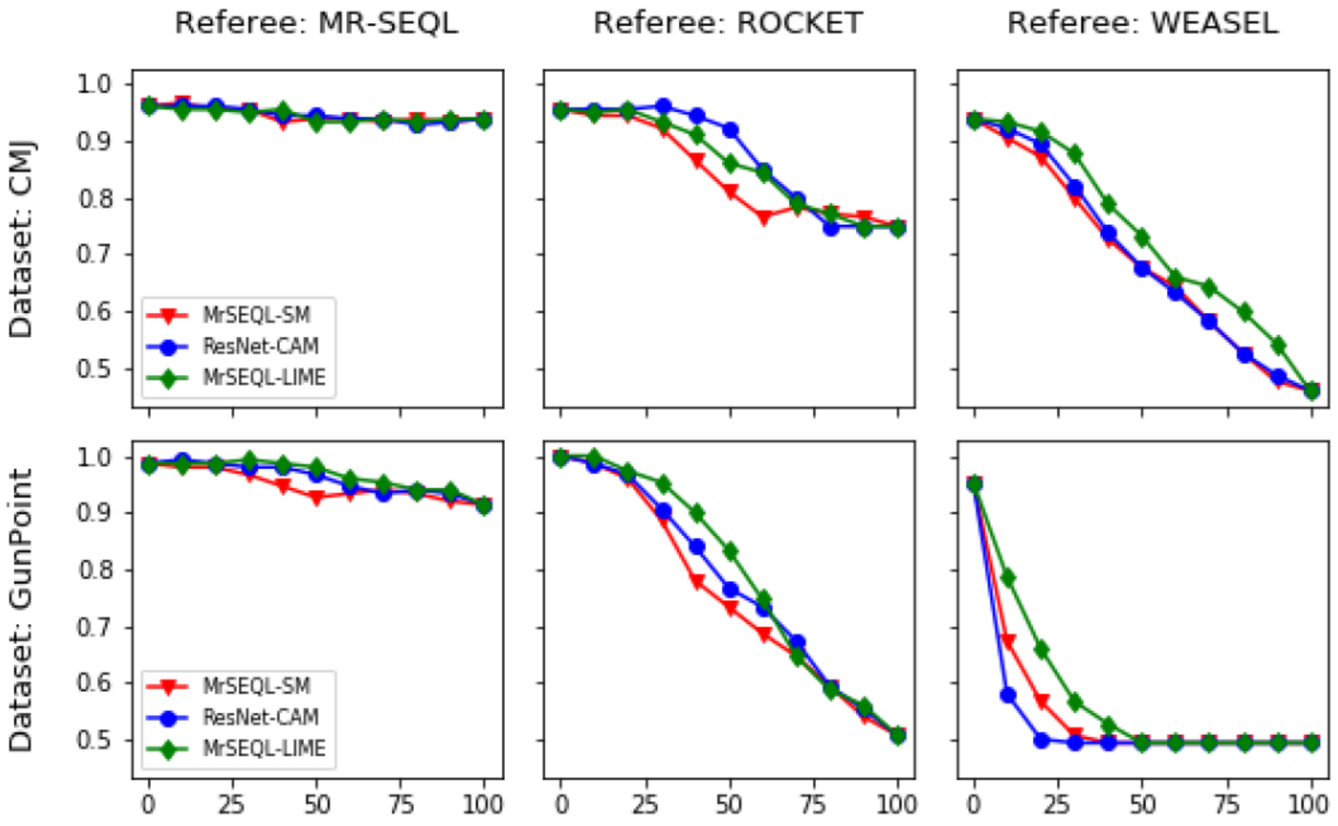


Fig. 8: Comparison of accuracy for *Type 1* perturbation based on three explanation methods (MrSEQL-SM, ResNet-CAM and Mr-SEQL-LIME) for GunPoint and CMJ datasets and three referee classifiers. Lower curve is better.

Experiments

Evaluate Multiple Explanation Methods:

Table 4: Summary of $eLoss_1$ of three explanation methods on five different problems. Lower value (column-wise) suggests the explanation method is better in explaining the problem according to the referee classifier.

Dataset	Explanation Method	Referee Classifier		
		Mr-SEQL	ROCKET	WEASEL
CBF	MrSEQL-SM	**0.9991**	**0.9941**	**0.6018**
	ResNet-CAM	0.9993	0.9945	0.6041
CMJ	MrSEQL-SM	**0.9441**	**0.8422**	**0.6899**
	ResNet-CAM	0.9453	0.8735	0.6972
	MrSEQL-LIME	**0.9441**	0.8612	0.7385
Coffee	MrSEQL-SM	**0.9625**	1.0	**0.9786**
	ResNet-CAM	0.9696	1.0	0.9821
ECG200	MrSEQL-SM	**0.811**	0.9065	0.7565
	ResNet-CAM	0.838	**0.9035**	**0.7385**
GunPoint	MrSEQL-SM	**0.9477**	**0.7567**	0.543
	ResNet-CAM	0.961	0.7773	**0.5257**
	MrSEQL-LIME	0.9677	0.7953	0.573

Experiments

Sanity Check for Experiment Result

- MrSEQL-SM explanation is the most informative according to the quantitative estimation and also the qualitative sanity check. The qualitative result is confirmed by a domain expert in sports science.

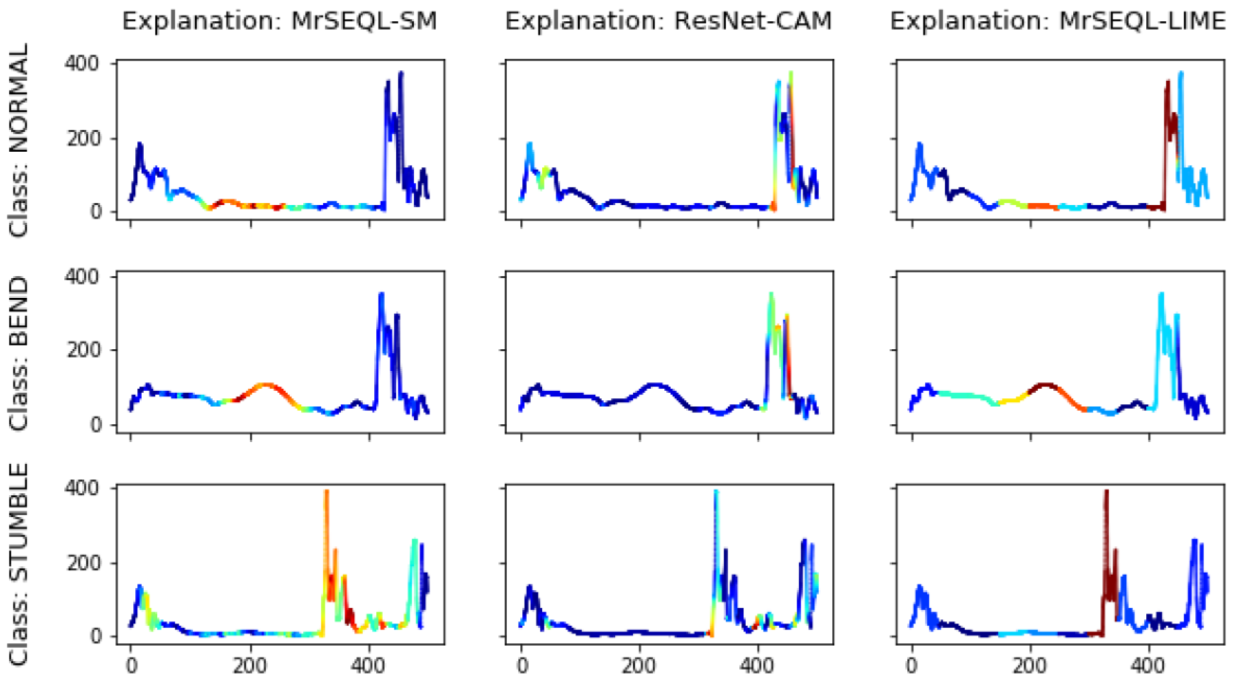


Fig. 9: Saliency maps produced by three explanation methods for example time series from the three classes of the CMJ dataset.

Conclusions

- It is possible to **quantitatively evaluate the informativeness** of explanation methods
  - **Key ingredients:** a set of explanation methods, explanation-driven perturbation, referee classifiers, explanation-driven loss in accuracy
  - The sanity check step (qualitative assessment) confirms the experiment result (quantitative assessment)
- Use cases
  - Our approach enables a user to assess an existing explanation method in the context of a given application or to evaluate different explanation methods and opt for one that works best for a specific use case.
  - Our method can be used to filter a set of potential explanation methods before conducting expensive user-studies.

Future Work

- Other perturbation approaches
  - Gaussian noise vs. Centroid-based
- Other comparison benchmarks - lower/upper bound on informativeness
  - Compare Type 1-2 vs. Compare Type 1-Random SM
- Use more/diverse referee classifiers
  - Detangle the robustness to noise from impact of explanation
- Quantify other XAI properties in the context of TSC
  - Coverage, stability, and more

Science Foundation Ireland
Centre for Research Training
in Machine Learning

Acknowledgements

This work was funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning (18/CRT/6183), the Insight Centre for Data Analytics (12/RC/2289 P2) and the VistaMilk SFI Research Centre (SFI/16/RC/3835).

Q&A

References

[1] Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classier. CoRR abs/1602.04938 (2016), http://arxiv.org/abs/1602.04938

[2] Le Nguyen, T., Gsponer, S., Ilie, I., O'Reilly, M., Ifrim, G.: Interpretable time series classification using linear models and multi-resolution multi-domain symbolic representations. Data Mining and Knowledge Discovery 33(4), 1183{1222 (Jul 2019). https://doi.org/10.1007/s10618-019-00633-3

[3] Zhou, B., Khosla, A., A., L., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization. CVPR (2016)

Science Foundation Ireland Centre for Research Training in Machine Learning