Aalen University

Invited talk:

Explainable AI for time series classification and anomaly detection: Current state and open issues

XAI-TS Workshop @ ECML/PKDD 2023
Turin, Italy

Andreas Theissler
Aalen University of Applied Sciences
Germany

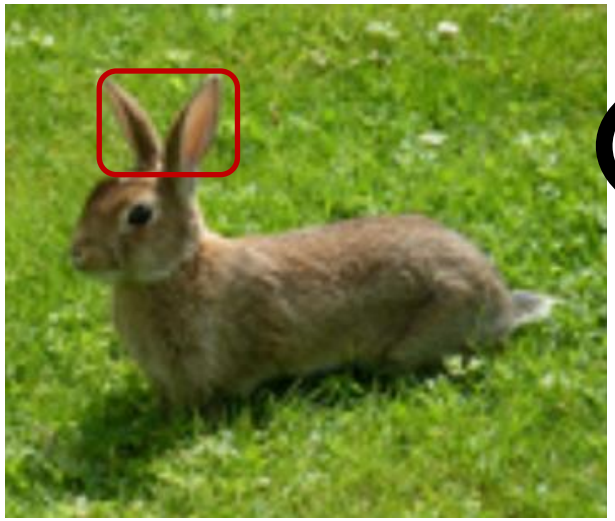
andreas.theissler@hs-aalen.de

https://ml-and-vis.org

https://www.researchgate.net/profile/Andreas_Theissler

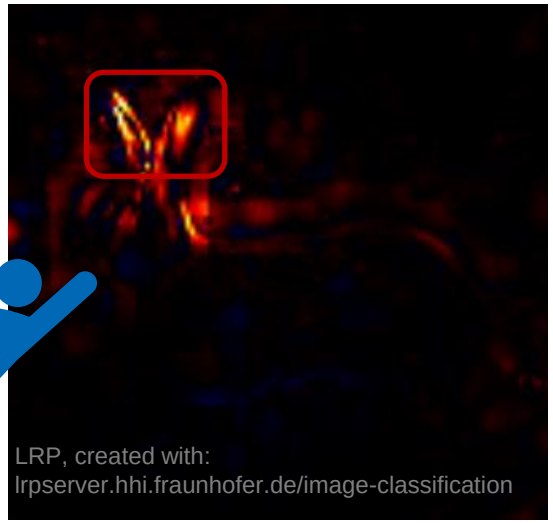
partially funded by research grant:
„AI Factory SME“
(Regionale Innovationszentren)

funded by:
EU (EFRE) and
Ministeriums für Wissenschaft, Forschung und Kunst Baden-Württemberg



What is that?
ML model says **rabbit**...
… why?

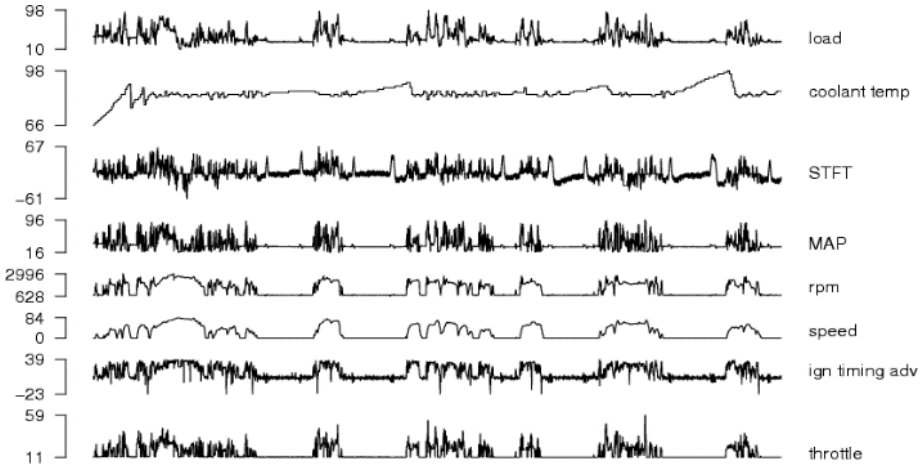
Ah, the ears…
Ok, makes
sense to me.



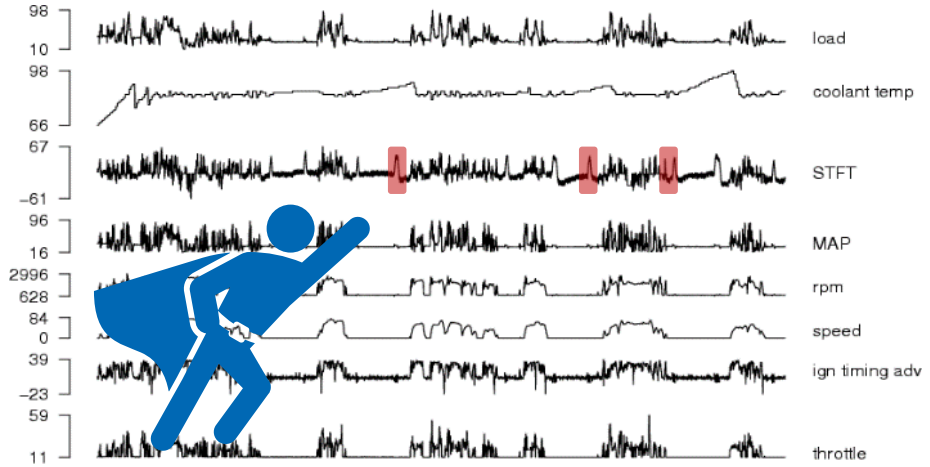
XAI super hero

And what is that?
ML model says **fault**...
… why?

Hmm… still don‘t get it.
Signals are jumping up and down all over the place…



recording of vehicle signals, taken from (Theissler, 2013)



XAI super hero

Let’s call this:
The non-inherent semantics problem.

XAI for time series: challenging and a bit „underresearched“

Observation:

- large part of work in XAI done on tabular data and computer vision
- (expect to see explosion of work done on text thanks to the rise of LLMs)

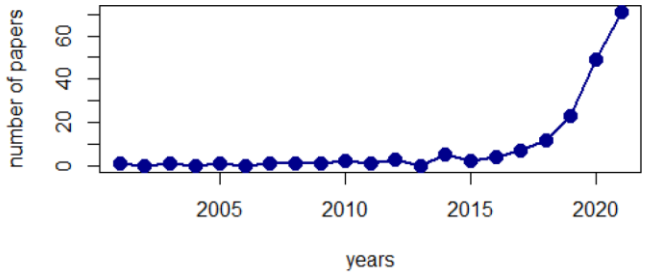
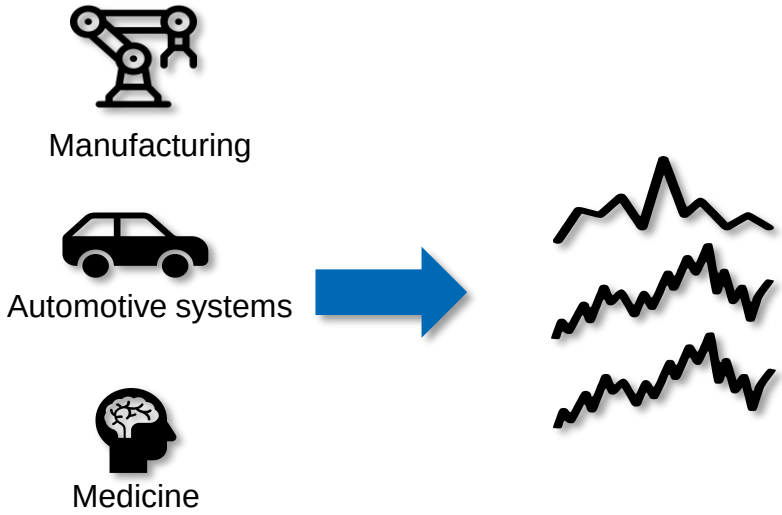


FIGURE 1. The number of papers published per year on XAI for time series classification started to increase significantly in 2019, suggesting an increase in the topic's relevance. The search was performed on Scopus

Less work done on time series, possible reasons:

- non-inherent semantics problem making it harder to develop and evaluate explanations
- availability of data (e.g. Imagenet, CIFAR, MNIST, etc.)

But: time series are omnipresent



and many more…

So we can conclude:

- XAI for time series is particularly challenging, but is worth the effort.
- It is great to have this workshop and a community of researchers.

XAI terminology in a nutshell

- intrinsically interpretable model: $\boldsymbol{understand}\left(\boldsymbol{M_j(x)}\right)$ or $\boldsymbol{understand}\left(\boldsymbol{M_j}\right)$

  vs.

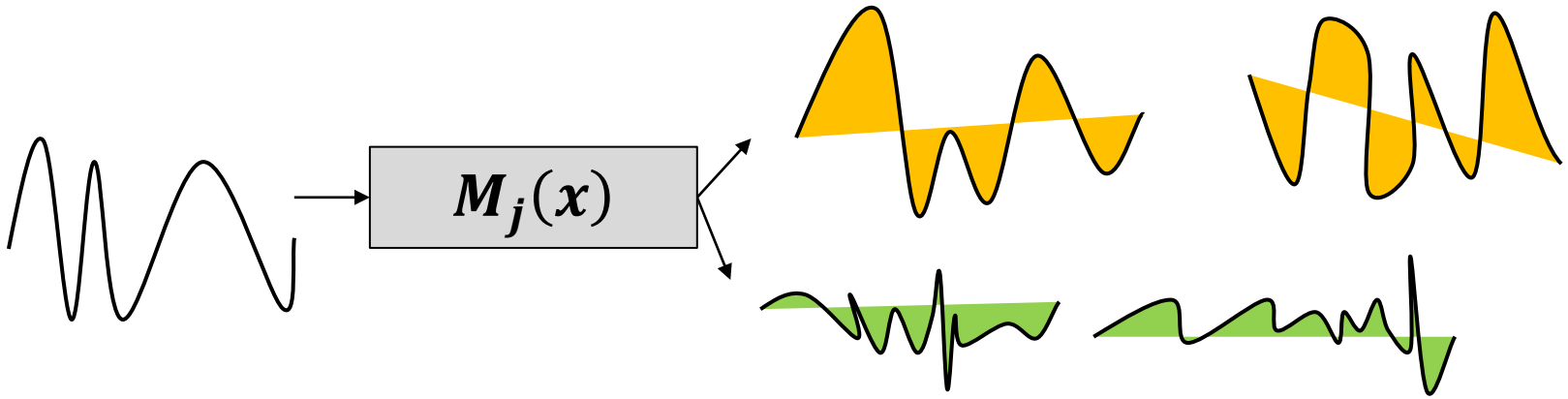
- explanations: $\boldsymbol{\varepsilon_i}\left(\boldsymbol{M_j(x)}\right)$ or $\boldsymbol{\varepsilon_i}\left(\boldsymbol{M_j}\right)$

Notation:
$\boldsymbol{\varepsilon_i}(\dots)$: explanation
$\boldsymbol{understand}(\dots)$: understanding an explanation or an instriniscally interpretable model
$\boldsymbol{M_j}$: machine learning model
$\boldsymbol{x}$: data item (i.e. time series)
$\boldsymbol{M_j(x)}$: prediction of machine learning model

(for non-intrinsically interpretable models)	**model-agnostic**	**model-specific**
local (outcome explanation)	$\boldsymbol{\varepsilon_i}\left(\boldsymbol{M_j(x)}\right) \quad i = const\ ;\ \forall\, j$	$\boldsymbol{\varepsilon_i}\left(\boldsymbol{M_j(x)}\right) \quad i = j\ ;\ \forall\, i,j$
global (model explanation)	$\boldsymbol{\varepsilon_i}\left(\boldsymbol{M_j}\right) \quad i = const\ ;\ \forall\, j$	$\boldsymbol{\varepsilon_i}\left(\boldsymbol{M_j}\right) \quad i = j;\ \forall\, i,j$

We should not forget:
Also for post-hoc explanations, the following is true: $\boldsymbol{understand}\left(\boldsymbol{\varepsilon_i}(\dots)\right)$

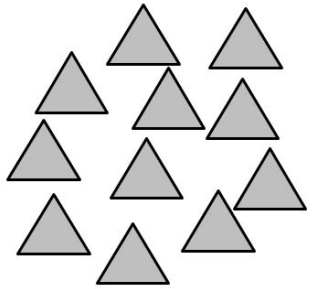
Time series classification



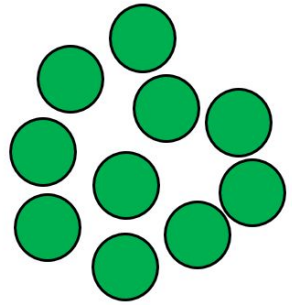
Definition: Given a dataset, **time series classification** is the task of training a function or mapping M_j from the space of possible inputs to a probability distribution over the class values such that an entire time series x is mapped to class C_i.

Time series anomaly detection

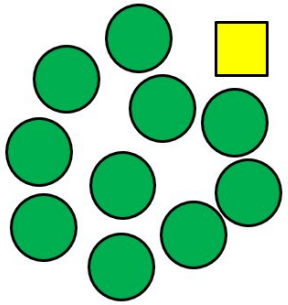
unsupervised
no class labels



semi-supervised
training set with one class

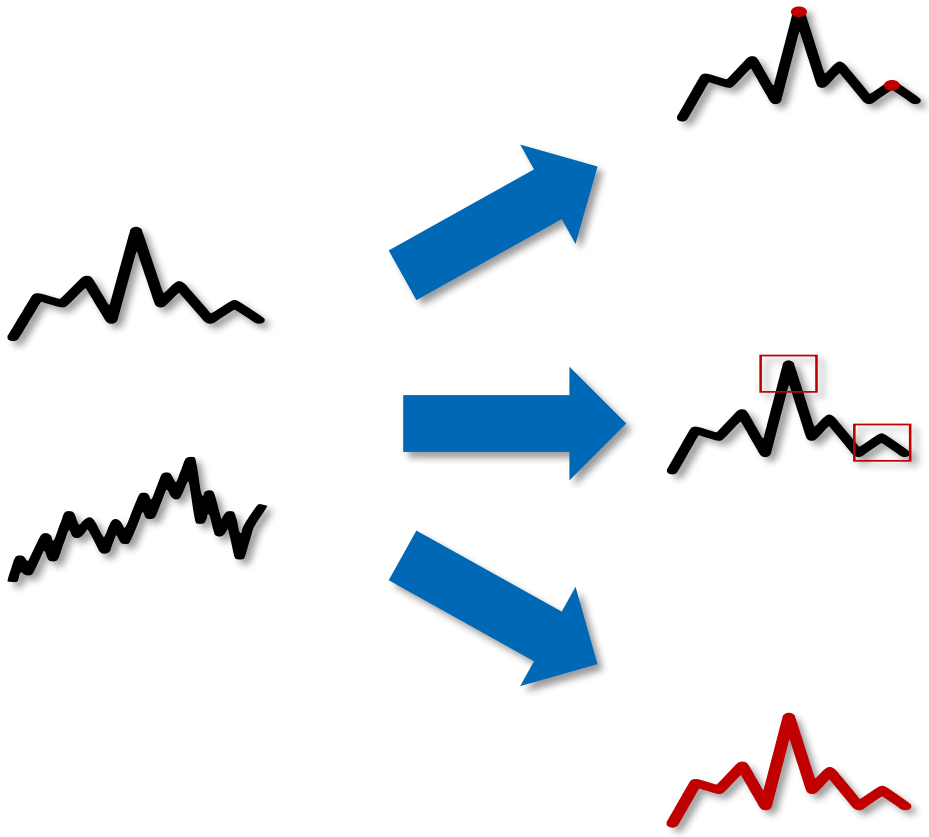


supervised
training set with ≥ 2 classes
(normal and anomalies)



categorization following (Chandola, 2009)

Time series anomaly detection



detecting anomaly time points in $\boldsymbol{x}$

detecting anomaly subsequences in $\boldsymbol{x}$

whole series anomaly detection

(resembles imbalanced time series classification)

Definition: Time anomaly detection is the task of training a function or mapping $\boldsymbol{M_j}$ such that either parts of or the entire time series $\boldsymbol{x}$ is classified as „normal“ or „anomaly“.

source: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions
https://doi.org/10.1109/ACCESS.2022.3207765

Categorization of XAI for time series classification

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 August 2022, accepted 13 September 2022, date of publication 19 September 2022,
date of current version 28 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3207765

SURVEY

Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions

ANDREAS THEISSLER[1], (Member, IEEE), FRANCESCO SPINNATO[2], UDO SCHLEGEL[3], AND RICCARDO GUIDOTTI[4]
[1]Information Systems, Aalen University of Applied Sciences, 73430 Aalen, Germany
[2]Computer Science, Scuola Normale Superiore, 56126 Pisa, Italy
[3]Department of Computer and Information Science, University of Konstanz, 78457 Konstanz, Germany
[4]Department of Computer Science, University of Pisa, 56126 Pisa, Italy

Corresponding author: Andreas Theissler (andreas.theissler@hs-aalen.de)

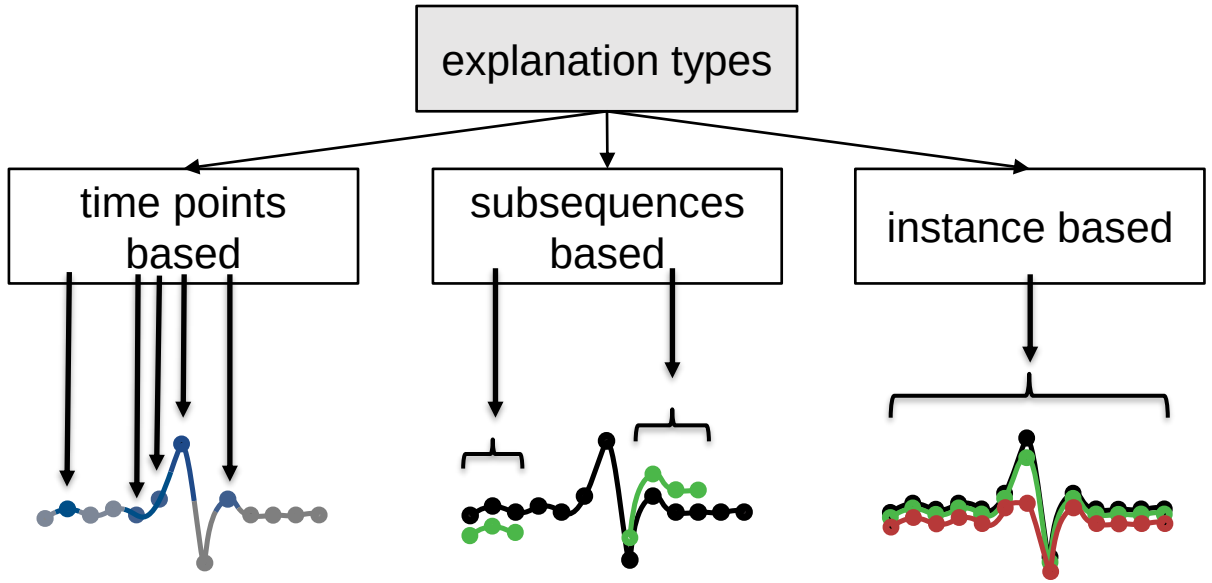
This work was supported in part by the European Community H 2020 Programme under the following funding schemes: G.A. 871042 SoBigData++, G.A. 952026 HumanE AI Net, ERC-2018-ADG G.A. 834756 XAI—Science and Technology for the eXplanation of AI Decision Making (xai), G.A. 952215 TAILOR (tailor), European coordinated research on long-term ICT and ICT-based scientific challenges (CHIST-ERA) Grant CHIST-ERA-19-XAI-010, by the Italian Ministry of University and Research (MUR) (N. not yet available), Austrian Science Fund (FWF) (N. I 5205), Engineering and Physical Sciences Research Council (EPSRC) (N. EP/V055712/1), National Science Center (NCN) (N. 2020/02/Y/ST6/00064), Estonian Research Council (ETAg) (N. SLTAT21096), and Bulgarian National Science Fund (BNSF) (N. KP-06-AOO2/5); in part by the Federal Ministry of Education and Research [Bundesministerium für Bildung und Forschung (BMBF)] under the VIKING (13N16242) Project, EXPLOR-20AT; in part by Stiftung Kessler + CO für Bildung und Kultur; and in part by the Aalen University of Applied Sciences.

ABSTRACT Time series data is increasingly used in a wide range of fields, and it is often relied on in crucial applications and high-stakes decision-making. For instance, sensors generate time series data to recognize different types of anomalies through automatic decision-making systems. Typically, these systems are realized with machine learning models that achieve top-tier performance on time series classification tasks. Unfortunately, the logic behind their prediction is opaque and hard to understand from a human

time points-based explanation:
a relevance score for every input data point $t_{i,j}$ of time series x where j refers to the dimension in case of a multivariate time series

subsequences based explanation:
based on subsequences which can be original (“proper”) subsequences or deduced (“improper”) subsequences from x

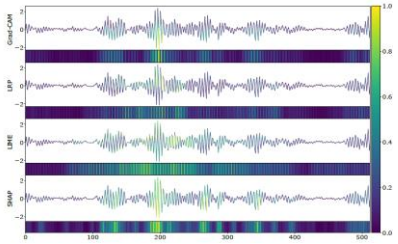
instance based explanation:
entire instance is used for the explanation

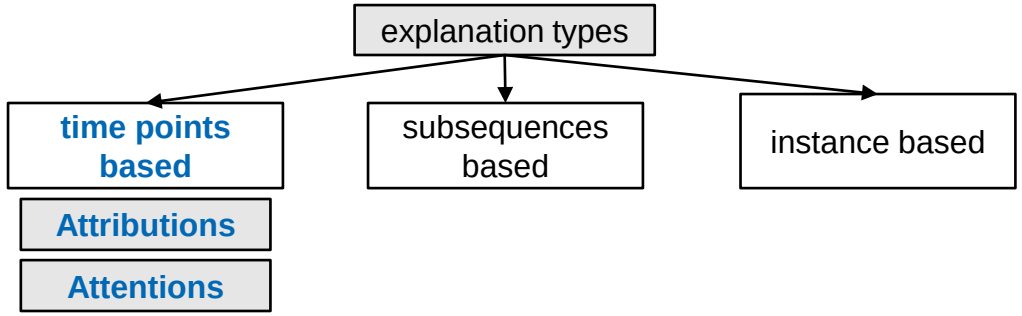


source: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions
https://doi.org/10.1109/ACCESS.2022.3207765

Aalen University

Time points based explanations





Attributions:

- use of external method to attribute model prediction $M_j(x)$ to time points of x
- many of the approaches originally proposed for tabular data or images and transferred to time series (SHAP, LIME, Grad-CAM, etc.)

Attentions:

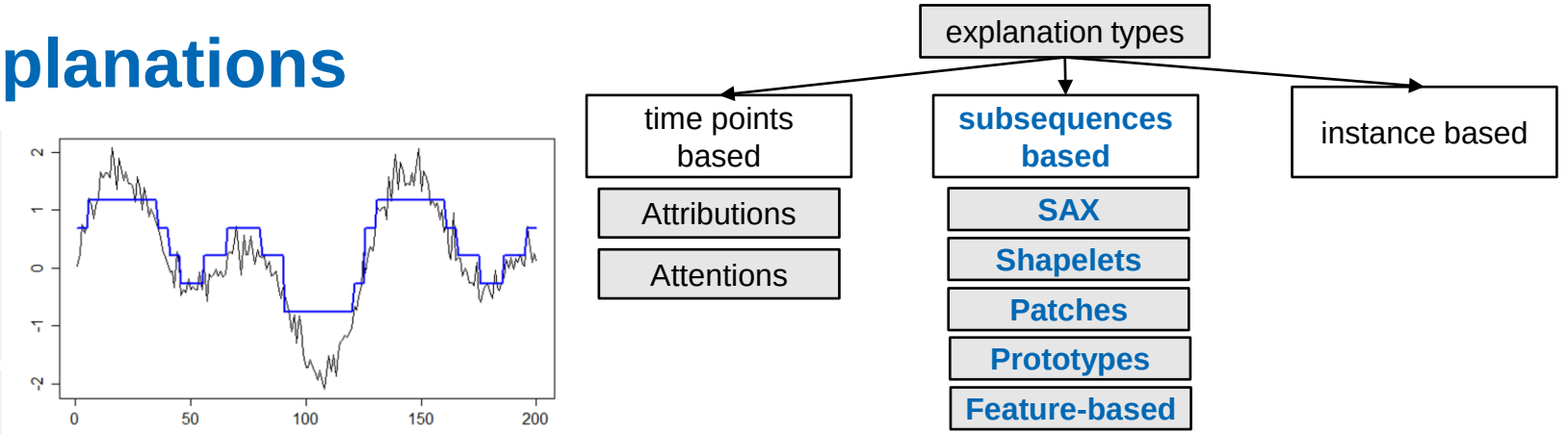
- use of model-internas to show which parts of x the model focusses on
- typically: attention layer as part of some neural network structure – e.g. combination of LSTM and CNN

Name	Ref	Year	Explanation Method	Post/Ante -hoc	Model-Agnostic/Specific	Global/ Local	TS-Specific	Uni/Multi -variate	Codi/Multi (URlariate	Code (URL)
Integrated Gradients	[53]	2017	Attributions	P	S	L	✗	-	P -	P
FCN	[54]	2017	Attributions	P	S	L	✓	U	P U	P
LIME	[36]	2016	Attributions	P	A	L	✗	-	P -	P
LRP	[55]	2015	Attributions	P	S	L	✗	-	P -	P
ExcitationBP	[56]	2016	Attributions	P	S	L	✗	-	P -	P
Occlusion	[57]	2014	Attributions	P	S	L	✗	-	P -	P
SHAP	[37]	2017	Attributions	P	A	L	✗	-	P -	P
SmoothGrad	[58]	2017	Attributions	P	S	L	✗	-	P -	P
DeepLIFT	[39]	2019	Attributions	P	S	L	✗	-	P -	P
Salience-CAM	[59]	2021	Attributions	P	S	L	✓	U	- U	-
Grad-CAM	[60]	2020	Attributions	P	S	L	✗	-	P -	P
TSViz	[61]	2019	Attributions	P	S	L	✓	U	P U	P
TSXplain	[62]	2019	Attributions	P	S	L	✓	M	- M	-
TSInsight	[63]	2021	Attributions	P	S	L	✓	M	- M	-
SoundLIME	[64]	2017	Attributions	P	A	L	✓	U	P U	P
MTEX-CNN	[65]	2019	Attributions	P	S	L	✓	M	- M	-
PERT	[66]	2021	Attributions	P	A	L	✓	U	P U	P
FIT	[67]	2020	Attributions	P	S	L	✓	M	P M	P
WinIT	[68]	2021	Attributions	P	S	L	✓	M	P M	P
CEFEs	[69]	2021	Attributions	P	S	L	✓	U	- U	-
XTF-CNN	[70]	2021	Attributions	P	S	L	✓	U	- U	-
LEFTIST	[71]	2019	Attributions	P	A	L	✓	U	P U	P
ALSTM-FCN	[12]	2017	Attentions	P	S	L	✓	U	P U	P
GCRNN	[72]	2018	Attentions	A	S	L	✓	U	P U	P
-	[73]	2018	Attentions	P	S	L	✓	U	P U	P
ETSCM	[74]	2019	Attentions	A	S	L	✓	U	- U	-
DACNN	[75]	2020	Attentions	A	S	L	✓	M	- M	-
LAXCAT	[76]	2021	Attentions	A	S	G	✓	M	- M	-
DeepVix	[77]	2020	Attentions	A	S	G	✓	M	JS M	JS
VixLSTM	[78]	2021	Attentions	P	S	G	✓	M	- M	-
-	[79]	2021	Attentions	A	S	L	✓	M	P M	P

source: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions
https://doi.org/10.1109/ACCESS.2022.3207765

Aalen University

Subsequence based explanations



SAX:

- symbolic representation of “proper“ subsequences
- for example used to form bag-of-patterns (SAX-VSM, MR-SEQL)

Shapelets:

- “improper” subsequences that are most representative of class membership
- most research focusses on *efficient* shapelet generation, assuming shapelets to be interpretable by definition
  - XCNN enhances interpretability by learning shapelets similar to real subsequences
  - LASTS creates a decision tree of factual and counterfactual shapelets

Patches / Prototypes:

- representative “improper” subsequences
  - e.g. P2ExNet

Feature-based:

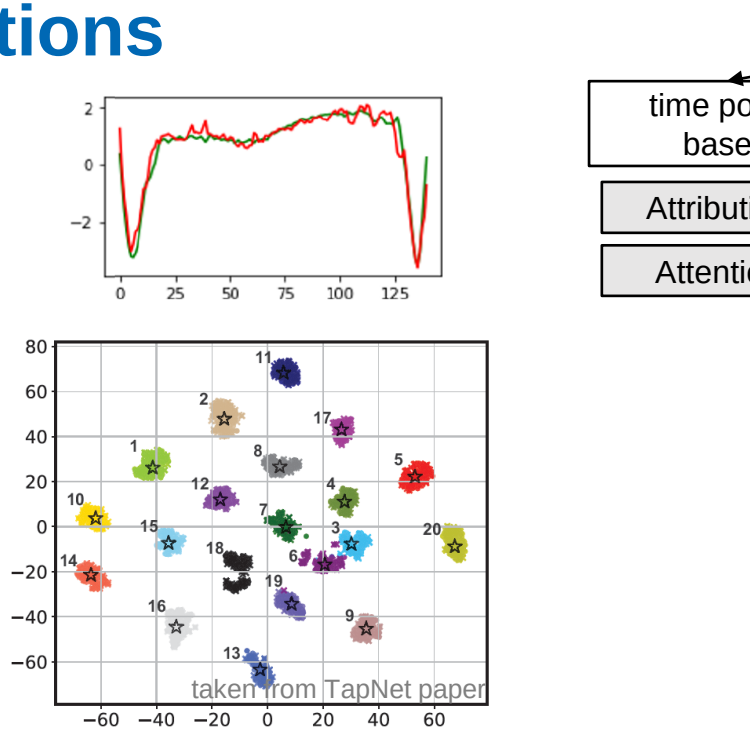
- operate on features extracted from subsequences
  - e.g. transformation to frequency domain

Name	Ref	Year	Explanation Method	Post/Ante -hoc	Model-Agnostic/Specific	Global/ Local	TS-Specific	Uni/Multi -variate	Code (URL)
-	[10]	2014	SAX	A	S	G	✓	U	-
SAX-VSM	[80]	2013	SAX	A	S	G	✓	U	-
-	[81]	2020	SAX	A	S	G	✓	M	P
MR-SEQL	[82]	2020	SAX	A	S	G	✓	U	P
CPHAP	[83]	2021	Shapelets	P	S	L	✓	U	-
Shapelets	[84]	2011	Shapelets	A	S	G	✓	U	J
ShapeletTransform	[85]	2012	Shapelets	A	S	G	✓	U	-
MSD	[86]	2012	Shapelets	A	S	G	✓	M	-
LS	[87]	2011	Shapelets	A	S	G	✓	U	-
LTS	[88]	2014	Shapelets	A	S	G	✓	U	P
LCTS	[89]	2016	Shapelets	A	S	G	✓	U	J
-	[90]	2018	Shapelets	A	S	G	✓	U	C++
-	[91]	2018	Shapelets	A	S	G	✓	U	-
LRS	[92]	2019	Shapelets	A	S	G	✓	U	P
ADSNs	[93]	2020	Shapelets	A	S	G	✓	U	P
XCNN	[94]	2020	Shapelets	A	S	L	✓	U	-
GENDIS	[95]	2021	Shapelets	A	S	G	✓	U	P
GMSM	[96]	2021	Shapelets	A	S	G	✓	M	-
MAPIC	[97]	2021	Shapelets	A	S	G	✓	U	P
DASH	[98]	2021	Shapelets	A	S	G	✓	U	P
TORRENT	[99]	2021	Shapelets	P	S	G	✓	U	-
LASTS	[100]	2020	Shapelets	P	A	L	✓	U	P
PatchX	[101]	2021	Patches	A	S	L	✓	M	-
P2ExNet	[102]	2020	Prototypes	A	S	L	✓	M	-
ProtoFac	[103]	2020	Prototypes	P	S	G	✗	-	P
mWDN	[104]	2018	Feature-based	A	S	G	✓	U	P

source: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions
https://doi.org/10.1109/ACCESS.2022.3207765

Aalen University

Instance based explanations



Prototypes:

- time series x' exemplifying the main aspects responsible for a model's prediction $M_j(x)$
- prototypes are e.g. classified with nearest neighbours methods
  - DPSN combines instance-based prototypes with shapelets
  - TapNet and is applicable to multivariate time series

Counterfactuals:

- given a time series x, a counterfactual is a generated time series $\tilde{x}$, such that $M_j(x) \neq M_j(\tilde{x})$ and the difference between x and $\tilde{x}$ is minimal
  - NativeGuide yields counterfactuals with proximity, sparsity, plausibility, and diversity
  - CoMTE obtains a distractor time series from the training set and substitutes parts of x to obtain $\tilde{x}$

Feature-based:

- operate on feature sets extracted from the entire time series
- interpretability is offered for example using
  - decision trees (MTDT) or
  - textual explanations (MODL-TSC)

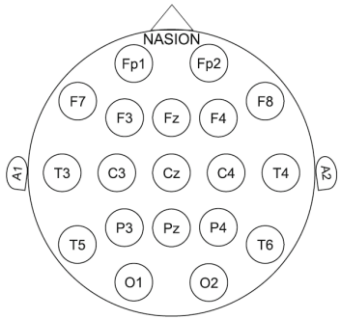
Name	Ref	Year	Explanation Method	Post/Ante -hoc	Model-Agnostic/Specific	Global/ Local	TS-Specific	Uni/Multi -variate	Code (URL)
ProSeNet	[105]	2019	Prototypes	A	S	G	✓	U	-
-	[52]	2019	Prototypes	A	S	G	✗	U	JS
DPSN	[106]	2020	Prototypes	P	S	G	✓	U	P
TapNet	[107]	2020	Prototypes	P	S	G	✓	M	P
MODL-TSC	[108]	2013	Feature-based	A	S	G	✓	U	-
MTDT	[109]	2018	Feature-based	A	S	G	✓	U	R
FC/TAA/LTAA	[110]	2019	Feature-based	A	S	G	✓	M	-
Conceptual	[111]	2020	Feature-based	P	A	G	✓	U	-
-	[112]	2021	Feature-based	A	S	G	✓	U	-
Native Guide	[113]	2021	Counterfactuals	P	A	L	✓	U	P
τ_{RT}/τ_{IRT}	[114]	2020	Counterfactuals	P	S	L	✓	U	P
CoMTE	[115]	2021	Counterfactuals	P	A	L	✓	M	P
CEM	[116]	2020	Counterfactuals	P	S	L	✗	-	-

XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series
https://doi.org/10.1007/s00521-022-07809-x

Aalen University

An application: Spectral and spatio-temporal explanation for multivariate EEG time series

Approach: Classification with 1D-CNN and 3D-CNN. Novel hybrid SHAP-based explanation in spectral, spatial and temporal dimension.



EEG channels on scalp

Neural Computing and Applications (2023) 35:10051–10068
https://doi.org/10.1007/s00521-022-07809-x

S.I.: INTERPRETATION OF DEEP LEARNING

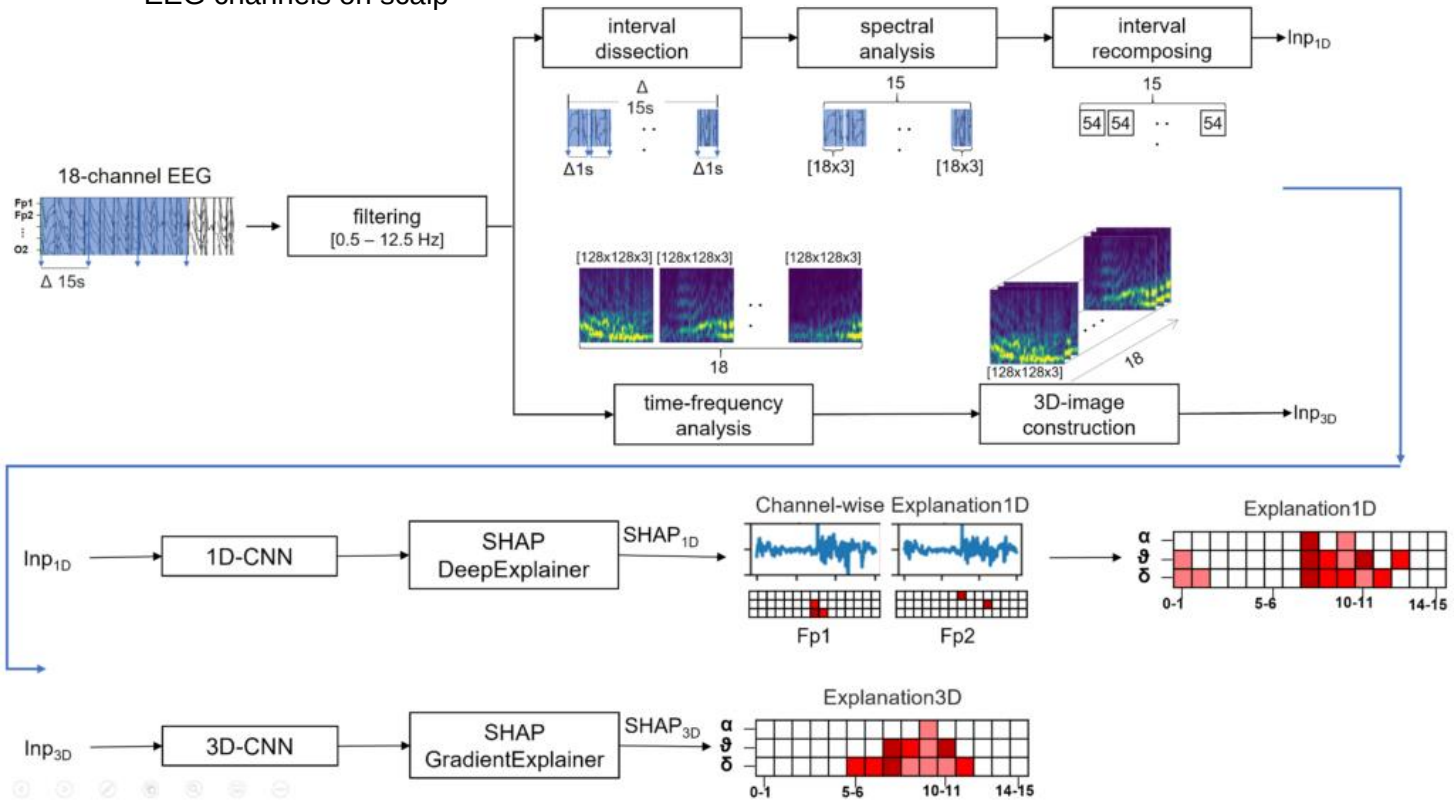
XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series

Dominik Raab[1] · Andreas Theissler[1] · Myra Spiliopoulou[2]

Received: 31 March 2022 / Accepted: 6 September 2022 / Published online: 29 September 2022

Abstract
In clinical practice, algorithmic predictions may seriously jeopardise patients' health and thus are required to be validated by medical experts before a final clinical decision is met. Towards that aim, there is need to incorporate explainable artificial intelligence techniques into medical research. In the specific field of epileptic seizure detection there are several machine learning algorithms but less methods on explaining them in an interpretable way. Therefore, we introduce XAI4EEG: an application-aware approach for an explainable and hybrid deep learning-based detection of seizures in multivariate EEG time series. In XAI4EEG, we combine deep learning models and domain knowledge on seizure detection, namely (a) frequency bands, (b) location of EEG leads and (c) temporal characteristics. XAI4EEG encompasses EEG data preparation, two deep learning models and our proposed explanation module visualizing feature contributions that are obtained by two SHAP explainers, each explaining the predictions of one of the two models. The resulting visual explanations provide an intuitive identification of decision-relevant regions in the spectral, spatial and temporal EEG dimensions. To evaluate XAI4EEG, we conducted a user study, where users were asked to assess the outputs of XAI4EEG, while working under time constraints, in order to emulate the fact that clinical diagnosis is done - more often than not - under time pressure. We found that the visualizations of our explanation module (1) lead to a substantially lower time for validating the predictions and (2) leverage an increase in interpretability, trust and confidence compared to selected SHAP feature contribution plots.

Keywords Explainable AI · SHAP · Deep learning · Machine learning · Epileptic seizures · EEG time series



A number of open issues we might want to work on

more can be found e.g. in:

Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions

https://doi.org/10.1109/ACCESS.2022.3207765

see: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions
https://doi.org/10.1109/ACCESS.2022.3207765

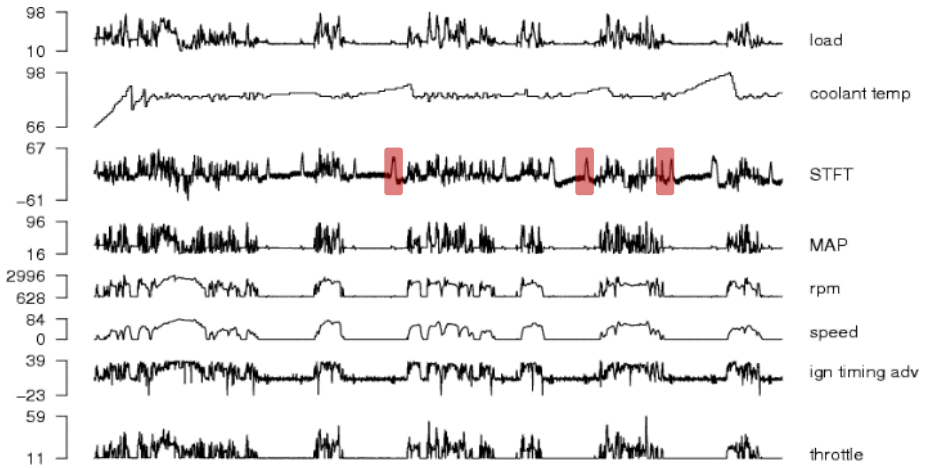
Aalen University

Open issue 1: Base explanations on higher-order representations

THE PROBLEM:

Explanations based on time points or proper subsequences might not be sufficient due to the non-inherent semantics problem.

Remember: $\boldsymbol{understand}(\boldsymbol{\varepsilon_i}(\ldots))$



Aalen University

Open issue 1: Base explanations on higher-order representations

RESEARCH DIRECTION:

Can we base our explanations on "higher-order representations" of the time series?

Using representations not constrained to the time domain (e.g. time, frequency, time-frequency, recurring patterns, missing patterns, statistical features, domain-specific terminology, etc. – all-in-one).

Some work in that direction:

- (Nguyen et al., 2019) uses multiple resolutions to work with subsequences
- SoundLIME (Mishra et al., 2017) uses time, time-frequency and frequency domain
- CEM, transferred to time series in (Labaien et al., 2020), uses existing and missing patterns
- XAI4EEG (Raab et al., 2023) incorporates temporal, spectral and spatial information in a domain-specific explanation
- Shapelet-based explanations, e.g. LASTS (Guidotti et al., 2020)

see: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions
https://doi.org/10.1109/ACCESS.2022.3207765

Aalen University

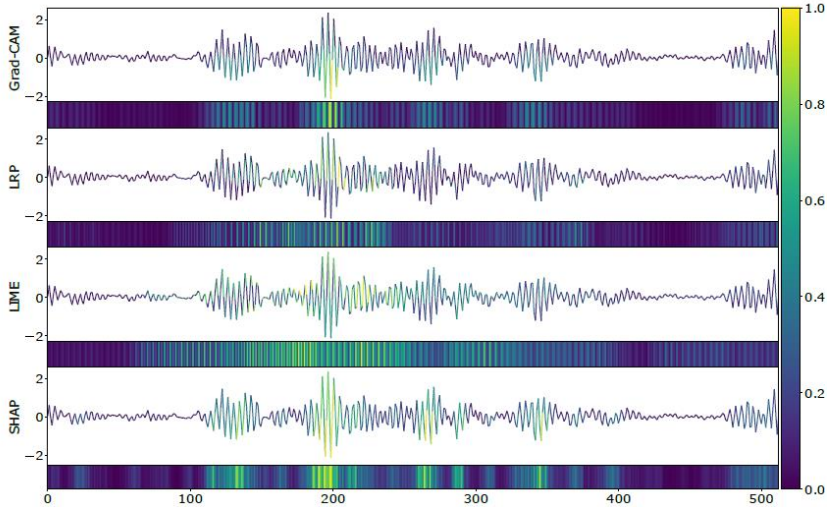
Open issue 2: Higher-order explanations

THE PROBLEM:

Commonly saliency maps (heatmaps) on a time point or subsequence basis are used.

Saliency maps show where in x data points influenced the prediction $M_j(x)$,
but not necessarily why the prediction was made.

Again remember: $\boldsymbol{understand}(\boldsymbol{\varepsilon_i}(\ldots))$



see: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions
https://doi.org/10.1109/ACCESS.2022.3207765

Aalen University

Open issue 2: Higher-order explanations

RESEARCH DIRECTION:

Can we build “higher-order explanations” of the time series?

Using higher-level ways of explanations in addition to saliency maps.
Textual explanations, possibly in domain-specific terminology, seem to be a promising direction.

Example: *Signal1 indicates that sensorX is erroneous because signal1 increases while signal2 remains constant*

Some work in that direction:

- textual explanations in TSXplain (Munir et al., 2019), were found to be valuable in a user study
- temporal logic is used in (Mohammadinejad, 2022)

see: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions
https://doi.org/10.1109/ACCESS.2022.3207765

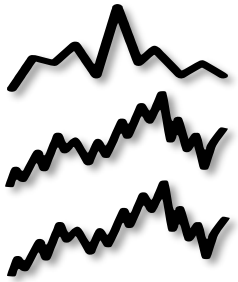
Open issue 3: Model-agnostic approaches desirable

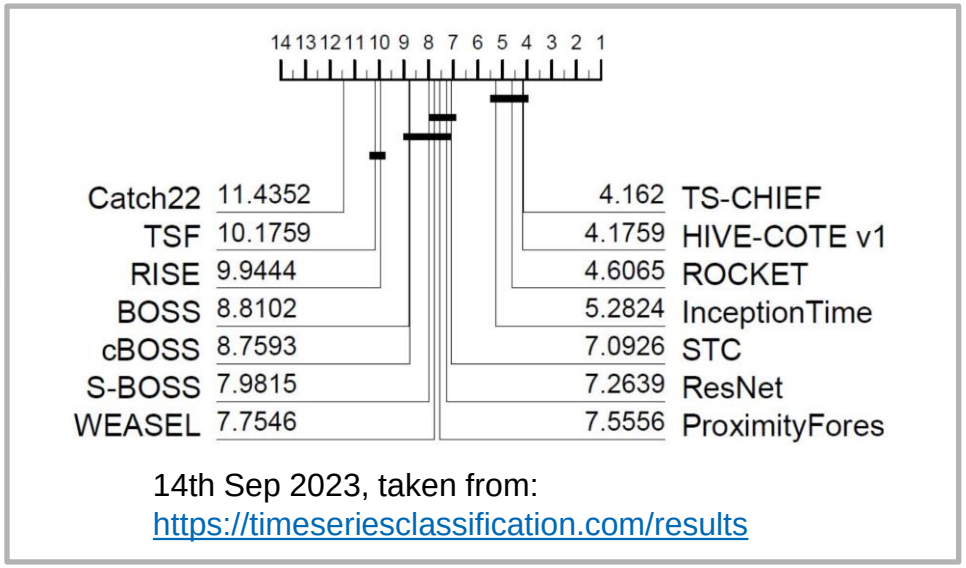
THE PROBLEM:

During ML model selection, entirely different model architectures are competitive.

- In contrast to, e.g., computer vision applications, deep learning methods do currently not clearly dominate the field. In the search for the best method, deep learning, ensembles, distance-based methods, shapelets and further methods are used.

- In order to compare the interpretability of these entirely different model architectures, model-agnostic methods are desirable.
- While there is a variety of model-agnostics approaches, the majority of the approaches is model-specific.





14th Sep 2023, taken from:
https://timeseriesclassification.com/results

$M_j(x)$

$\varepsilon_i\left(M_j(x)\right)$

$i = const\ ;\ \forall\ j$ or $i = j\ ;\ \forall\ i,\ j$?

see: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions
https://doi.org/10.1109/ACCESS.2022.3207765

Aalen University

Open issue 3: Model-agnostic approaches desirable

RESEARCH DIRECTION:

Can we build a set of model-agnostic explanations, at least for use during model selection?

Some of the work in that direction:

- time points based: PERT (Parvatharaju et al., 2021), LEFTIST(Guilleme et al., 2019)
- subsequence based: LASTS (Guidotti et al, 2020)
- instance based: NativeGuide (Delaney et al., 2021)

Fun fact: This very speaker is currently working on a **time point-based**, **model-specific** explanation.

Aalen University

Open issue 4: The evaluation problem

THE PROBLEM:

One strong motivation for our research field of XAI is:

We want users to be able to trust the decision of ML models.

However:

Can we trust our XAI methods?

A recent experimental study (Bodria et al., 2023) showed that XAI methods for computer vision may yield highly inconsistent results. A fact that corresponds with our experience using XAI models.

see: Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions
https://doi.org/10.1109/ACCESS.2022.3207765

Open issue 4: The evaluation problem

RESEARCH DIRECTIONS:

a) Quantitative evaluation of explanations should be developed further

b) Benchmark data sets with ground truth for evaluation are desirable

- comparable to the image data CLEVR-XAI (Arras et al., 2022)

c) Evaluation should also address human interpretability

- from the >60 time series-specific XAI approaches reviewed in (Theissler et al., 2022) only four were evaluated with a user study

Time series-specific work in that direction:

- evaluation methods: (Schlegel et al., 2019), (Nguyen et al., 2020), (Baric et al., 2021), (Schlegel and Keim, 2023) [preprint]
- benchmark data sets: none that I am aware of – we keep creating our own artificial test data

Aalen University

Thanks!

I am happy to have a discussion during the breaks…
Open for collaborations.

I am here all week.

Wish you a successful workshop and conference!

Andreas Theissler
Aalen University of Applied Sciences
Germany

andreas.theissler@hs-aalen.de

www.ml-and-vis.org

https://www.researchgate.net/profile/Andreas_Theissler

Slides available:

www.ml-and-vis.org/xai-ts