Aalen University

Invited talk:

Explainable AI: how far we have come and what's left for us to do

XKDD 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

Aalen University

Motivation: true story from a research project

The task

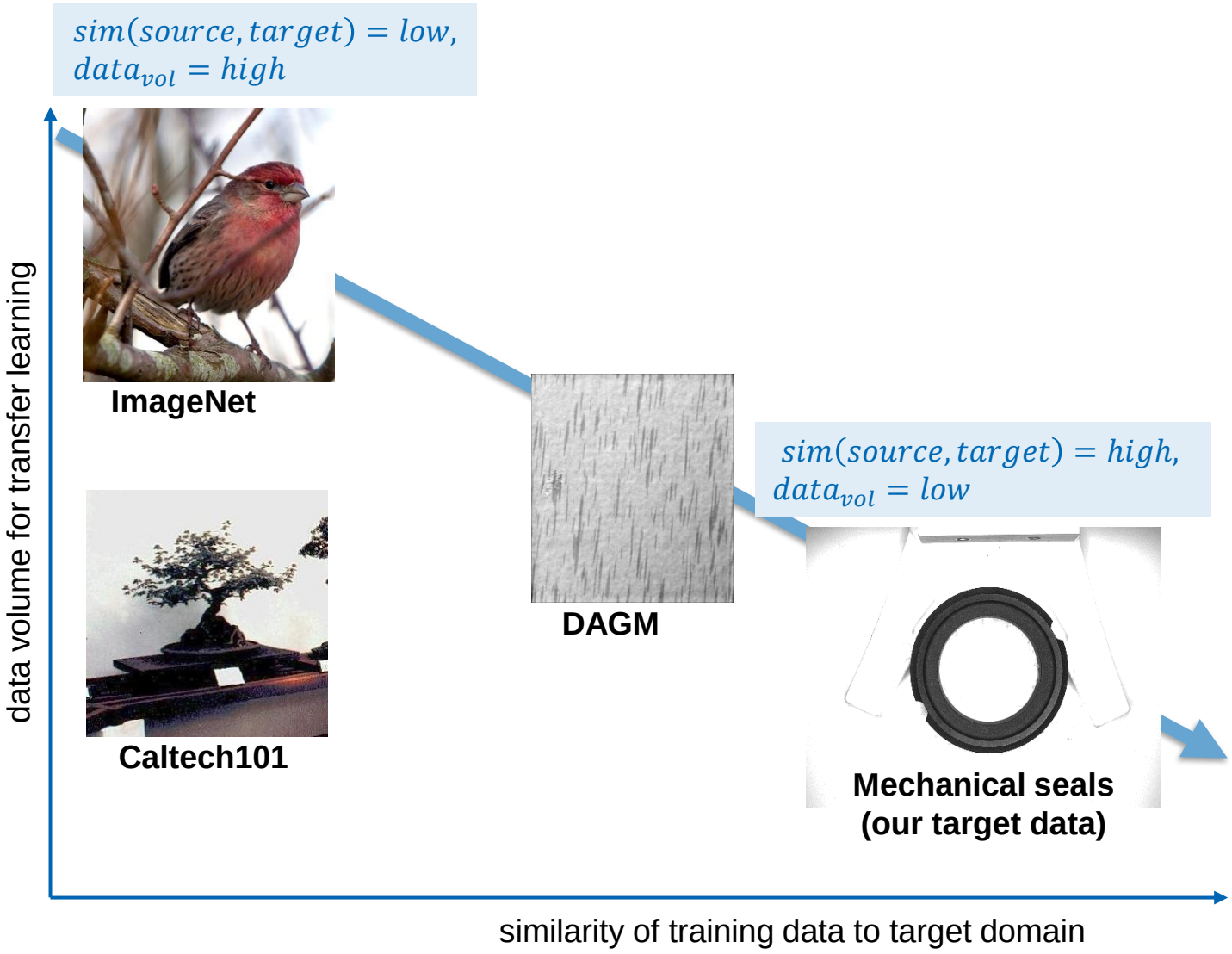
- **surface defect detection** in industrial settings
- binary classification of input images
  (see the black, circular workpiece)

The problem

- limited number of images available

The idea

- **transfer learning** with very few samples (**few-shot learning**)
- pre-trained models on data with differing similarity btw. source and target domain



source: Domain Transfer for Surface Defect Detection using Few-Shot Learning on Scarce Data
(should appear on IEEE Xplore soon)

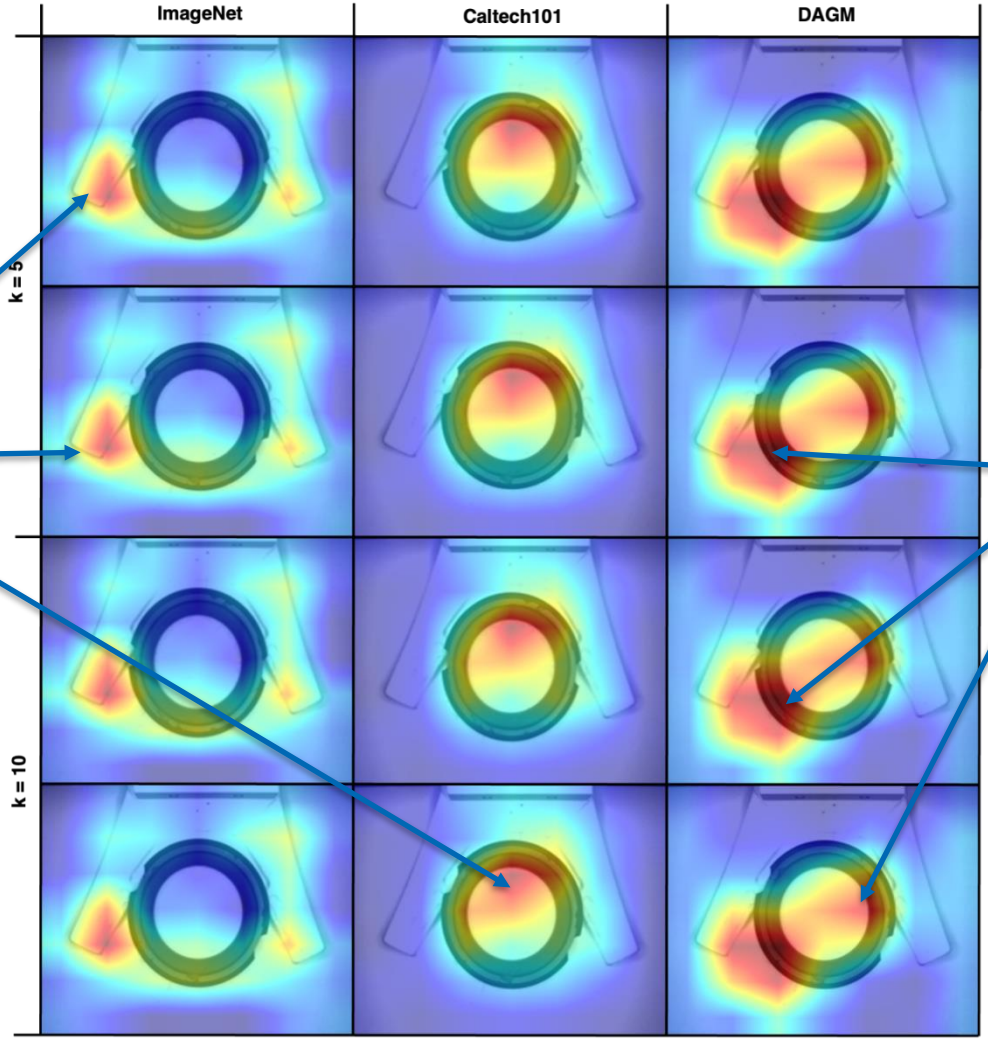
Aalen University

Motivation: true story from a research project

Models pre-trained on ImageNet or Caltech101:

- classified majority of images correctly
- but: did not focus on the workpiece
- **did the right thing, but for the wrong reason**

(such a setting was coined as the “Clever-Hans Effect” in [28])



Grad-CAM, consistent behaviour over the images in the test set

Models pre-trained on smaller data set with higher similarity to target domain (DAGM)

- identified damaged regions on the workpiece and based decision on these
- **did the right thing (almost as good) and for the right reason**

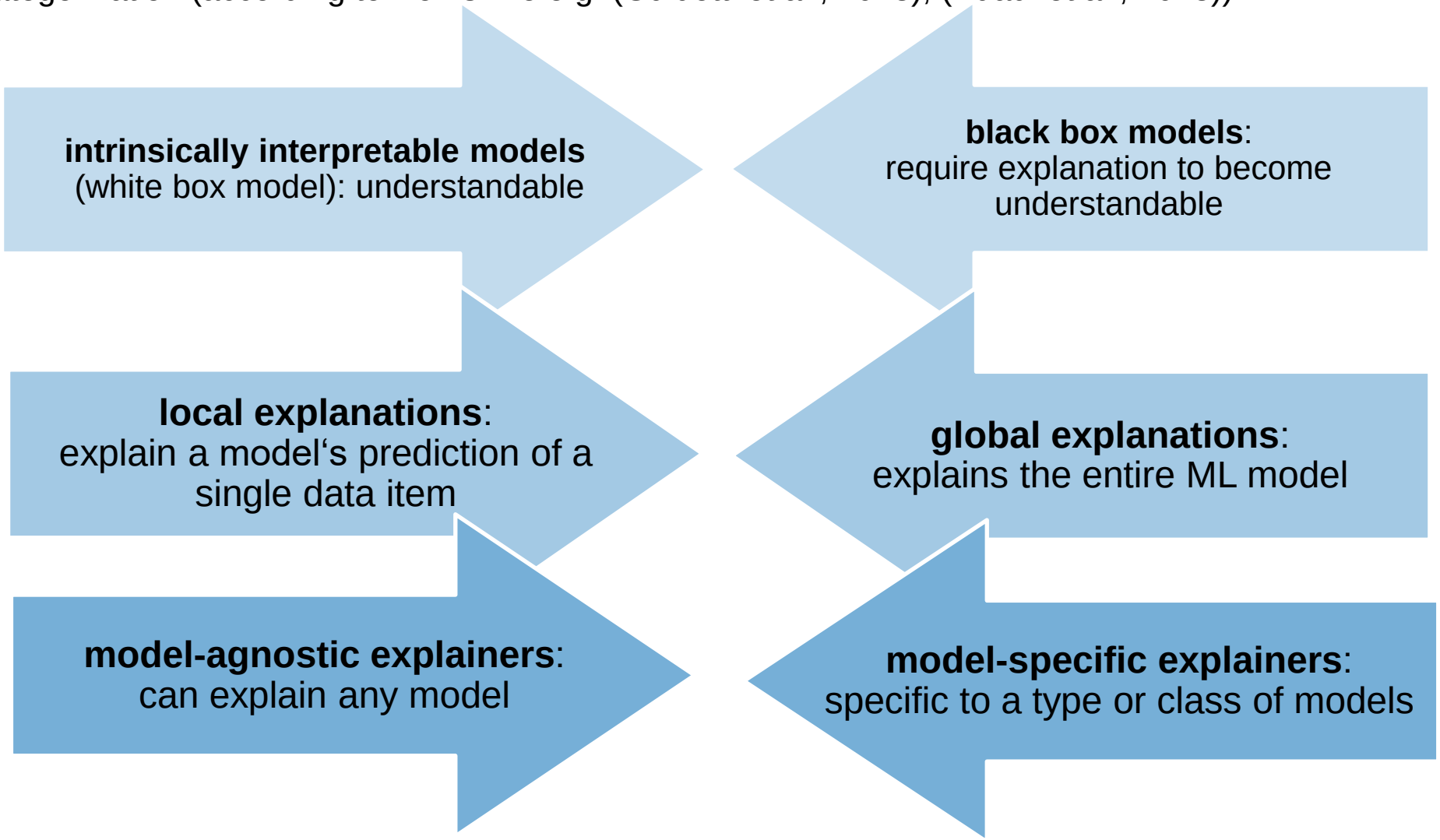
Findings:
- damaged workpieces were redorded at a later point in time (6 weeks) by industry partner
- hidden effects in the data

Prequesites to understand this talk

- some basic experience with XAI
- common sense

XAI terminology in a nutshell

A first rough categorization (according to works like e.g. (Guidotti et al., 2018), (Adadi et al., 2018))



XAI terminology in a nutshell

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

  vs.

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

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

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

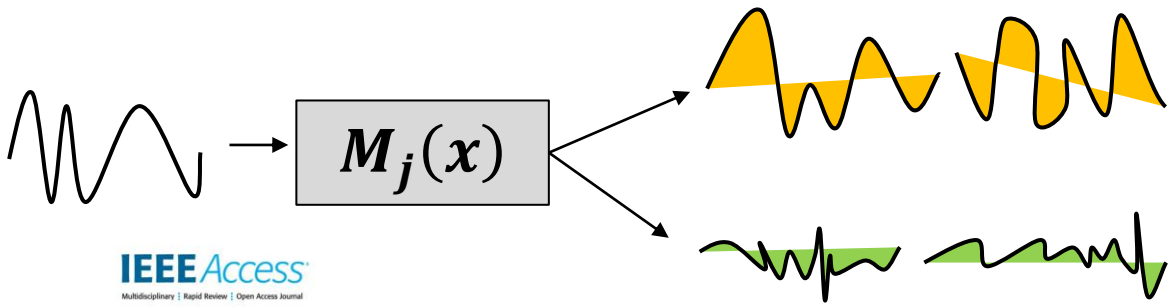
We should not forget:
Also for (post-hoc) explanations, the following is true: $\boldsymbol{understand}(\boldsymbol{\varepsilon_i}(\ldots))$

Just briefly,

some own work…

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

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

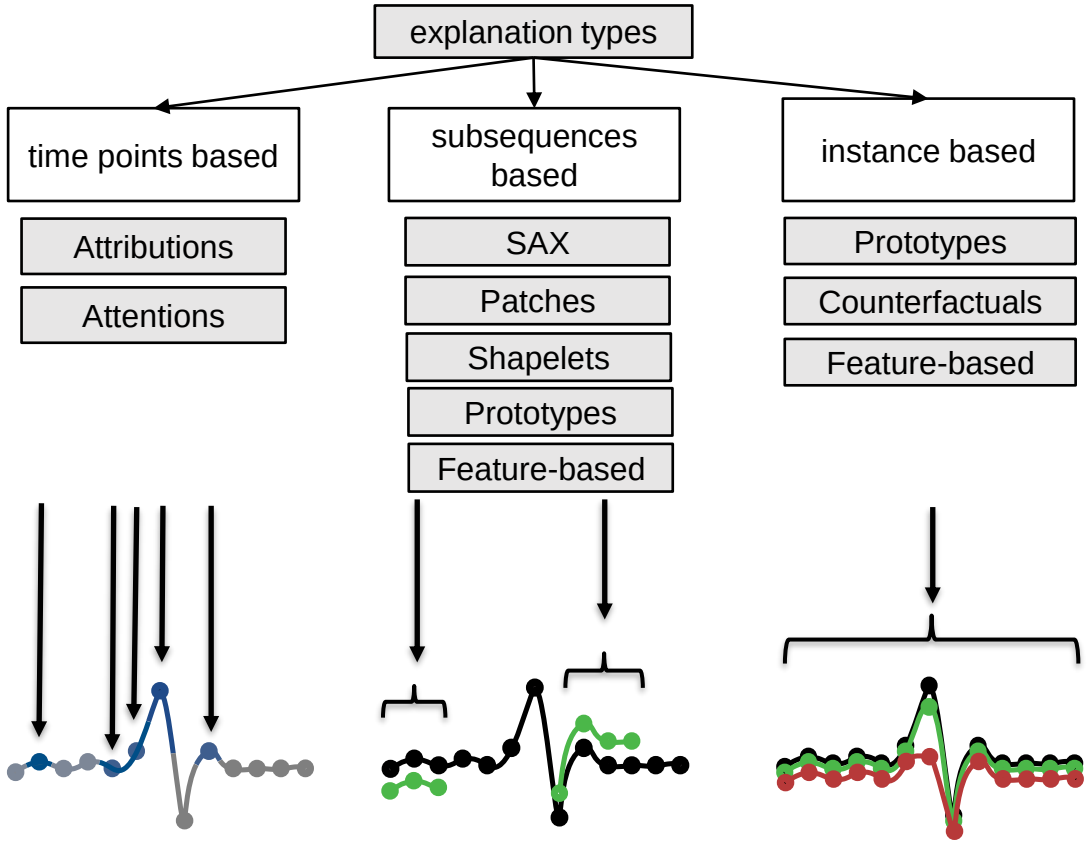
ANDREAS THEISSLER[1], (Member, IEEE), FRANCESCO SPINNATO[2], UDO SCHLEGEL[3], AND RICCARDO GUIDOTTI[4]

[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

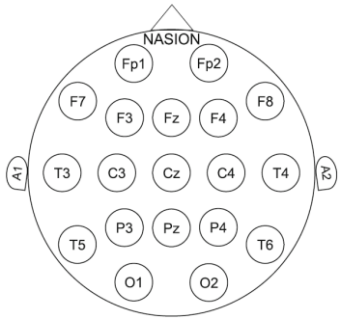


Results:
- vast majority of approaches not evaluated with users
- most approaches are model-specific
- a bit underresearched, lacking maybe 2-3 years behind XAI for computer vision

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

Spectral and spatio-temporal explanation for multivariate EEG time series

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



EEG channels on scalp

- explanation in domain-specific terminology
- hybrid explanation covering several aspects of the time series

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

S.I.: INTERPRETATION OF DEEP LEARNING

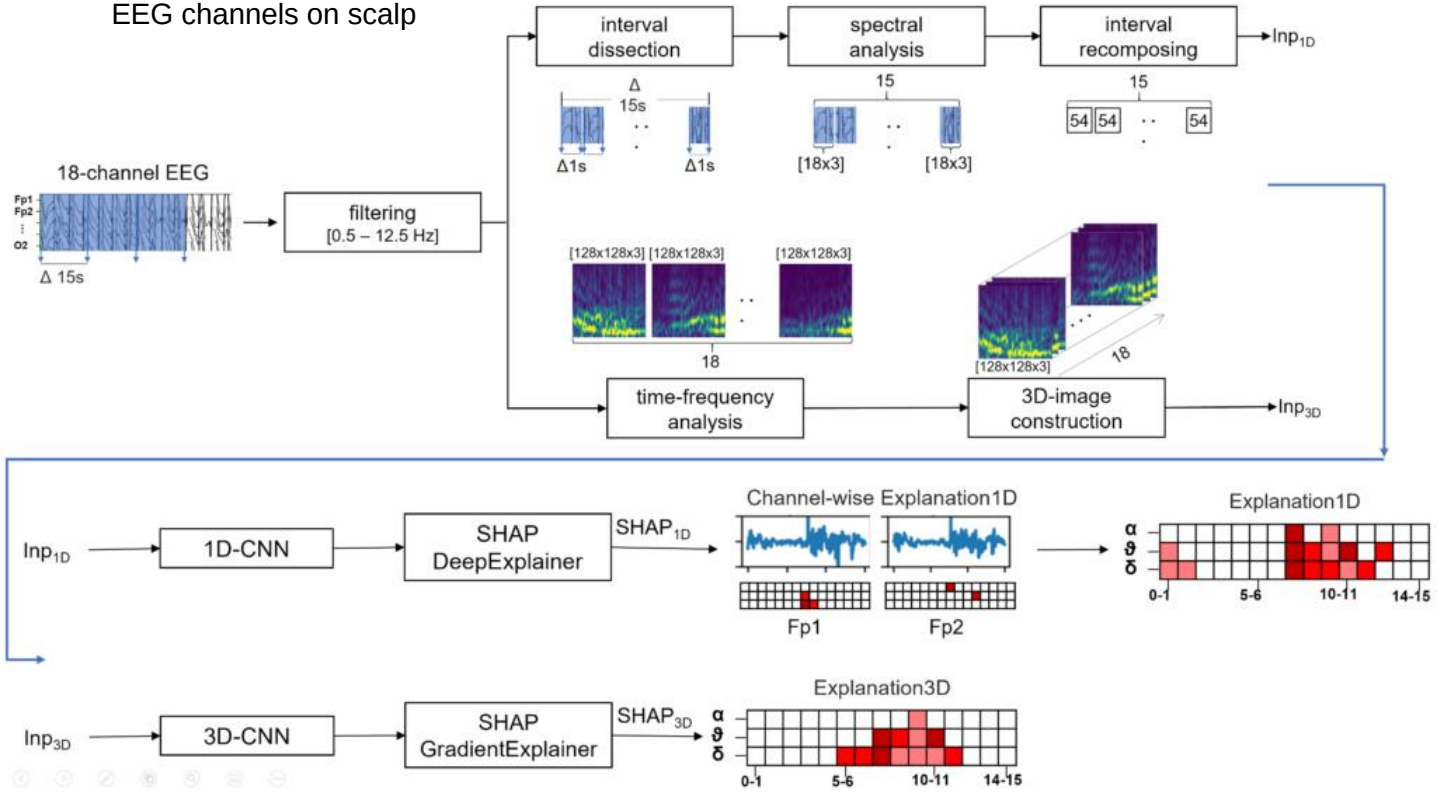
XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series

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

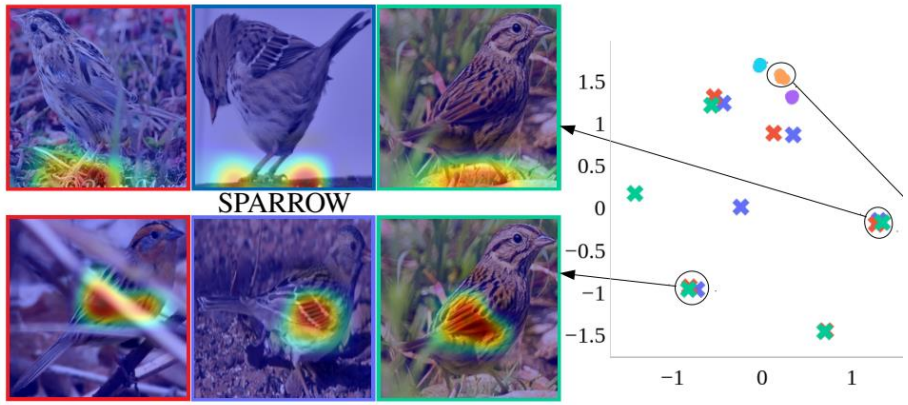


Stefan Kraft, Klaus Brölemann, Andreas Theissler, Gjergji Kasneci (2021).
SPARROW: Semantically Coherent Prototypes for Image Classification
https://www.bmvc2021-virtualconference.com/assets/papers/0896.pdf

Aalen University

SPARROW: Semantically coherent prototypes for image classification

- A high number of prototypes with semantic overlap do not add to interpretability, since this violates the principle of sparsity.
- we proposed a SPARROW to obtain semantically coherent prototypes
  (mainly conducted by a PhD student at an industry partner and at Tuebingen University)
- bases on ProtoPNet by Chen et al.
- approach requires a ground truth data set with image patches



SPARROW: Semantically Coherent Prototypes for Image Classification

Stefan Kraft[1,4]
stefan.kraft@stz-softwaretechnik.de
Klaus Broelemann[2]
klaus.broelemann@schufa.de
Andreas Theissler[3]
https://orcid.org/0000-0003-0746-0424
Gjergji Kasneci[4,2]
gjergji.kasneci@uni-tuebingen.de

[1] IT-Designers Group
Esslingen am Neckar, GER
[2] SCHUFA Holding AG
Wiesbaden, GER
[3] Aalen University of Applied Sciences
Aalen, GER
[4] Data Science & Analytics Research
The University of Tübingen
Tübingen, GER

Abstract

Current prototype-based classification often leads to prototypes with overlapping semantics where several prototypes are similar to the same image parts. Also, single prototypes tend to activate highly on a mixture of semantically different image parts. This impedes interpretability since the nature of the connections between the parts is unknown. We propose a framework that is comprised of two key elements: (i) A novel method which leads to semantically coherent prototypes and (ii) an evaluation protocol which is based on part annotations and allows to quantitatively compare the explanatory capacity of prototypes from different methods. We demonstrate the viability of our framework by comparing our method to a standard prototype-based classification method and show that our method is capable of producing prototypes of superior interpretability.

Theissler, A., Thomas, M., Burch, M., Gerschner, F. (2022),
ConfusionVis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices.
Elsevier Knowledge-Based Systems

Aalen University

Model selection based on outputs: loosely related to XAI

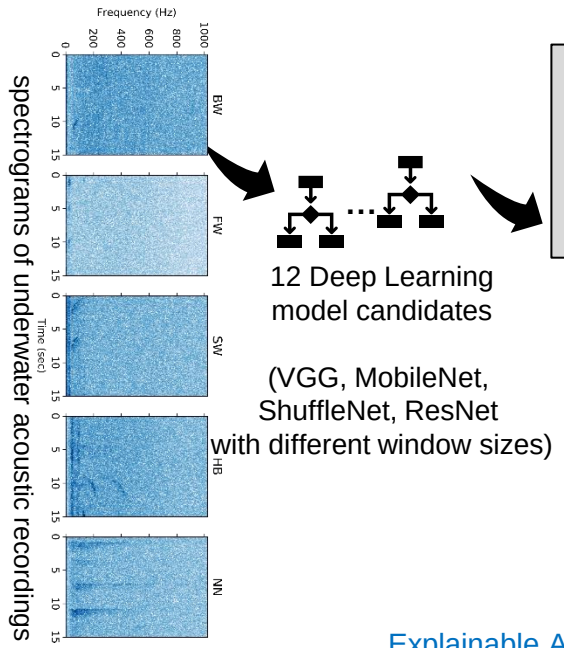
Problem setting:

A large number of ML model candidates is generated during training. Ranking by one single metric does not reflect all aspects of the models.

Approach:
Comparative evaluation and selection of of ML classifiers based on their high-level outputs (confusion matrices)

Result: in a case study, the marine biologist identified a model that would not have been selected by considering the accuracy.

www.ml-and-vis.org/mex
www.ml-and-vis.org/confusionvis



12 Deep Learning model candidates

(VGG, MobileNet, ShuffleNet, ResNet with different window sizes)

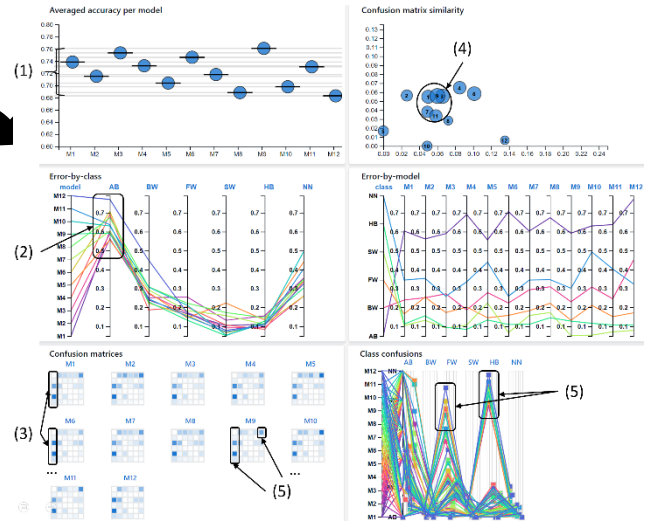
Class labels

Class labels

Prediction	AB	BW	FW	SW	HB	NN
AB	995	75	250	24	192	101
BW	65	566	49	7	5	14
FW	505	66	2554	4	155	25
SW	39	11	10	353	0	3
HB	790	9	175	1	2833	10
NN	117	17	23	5	5	291

class	source	label
0	Ambient noise	AB
1	Blue whale	BW
2	Fin whale	FW
3	Sei whale	SW
4	Humpback whale	HB
5	Non-biological noise	NN

ConfusionVis:



What‘s left for us to do …

… some (preliminary) thoughts

5 problems

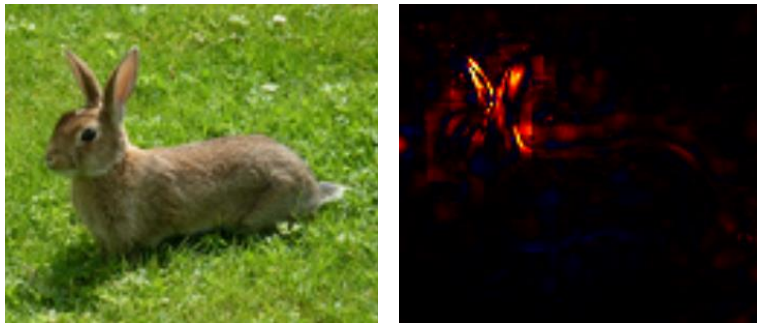
(open or partially solved)

we might want to work on

XAI: What‘s left for us to do… some thoughts (Problem 1)

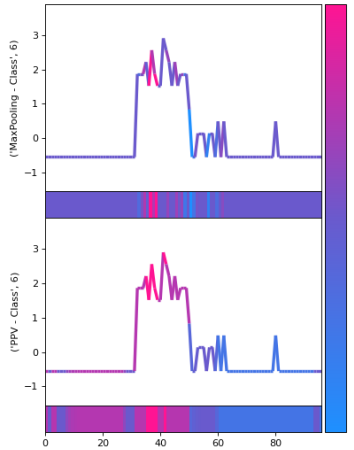
Many of our explainers show **where** in $\boldsymbol{x}$ data points influenced the prediction $\boldsymbol{M_j(x)}$,
but **not why** the prediction was made.

works fine here:



LRP, created with: lrpserver.hhi.fraunhofer.de/image-classification

not so clear here:



We tend to forget:

$$\boldsymbol{understand}\left(\boldsymbol{\varepsilon_i(\ldots)}\right)$$

Let‘s call this:

The non-inherent semantics problem

XAI: What‘s left for us to do… some thoughts (Problem 2)

We build our explainers $\boldsymbol{\varepsilon_i}(\ldots)$ with the target users in mind.

So we tailor our explainers towards ML engineers, domain experts, or laymen, etc. (well, sometimes at least…) such that:

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

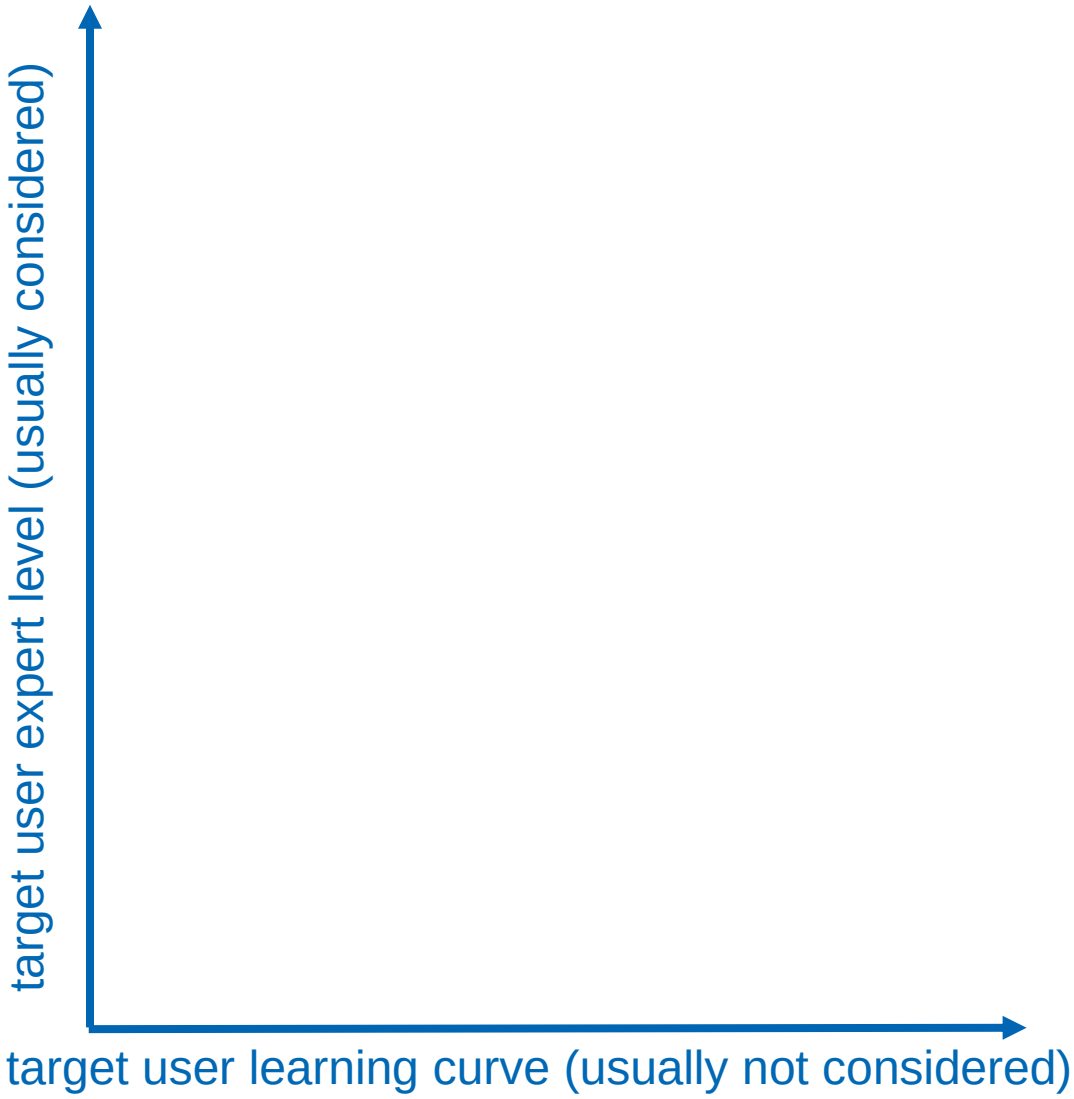
However: users have learning curves,

users „understand“ in hierarchical manners, i.e.:

$$\boldsymbol{understand} = \boldsymbol{f}(\boldsymbol{user\ and\ time})$$

Let‘s call this:

The dynamic-target-user problem



XAI: What‘s left for us to do… some thoughts (Problem 3)

From different ML models $\boldsymbol{M_i}$ we expect:

$$\boldsymbol{M_i}(\boldsymbol{X}) \approx \boldsymbol{M_j}(\boldsymbol{X});\ \forall\ i, j$$

So from different explainers $\boldsymbol{\varepsilon_i}$ we should intuitively desire:

$$\boldsymbol{\varepsilon_i}(\ldots) \approx \boldsymbol{\varepsilon_j}(\ldots);\ \forall\ i, j$$

- However, we observe that we get rather different explanations from different explainers.
- A recent experimental study (Bodria et al., 2023) showed that XAI methods for computer vision may yield highly variant results. A fact that corresponds with our experience using XAI models.

source: (Bodria et al., 2023)

Let‘s call this:

The variance problem

XAI: What‘s left for us to do… some thoughts (Problem 4)

One strong motivation for XAI is:

We want to explain black box models in order to trust them.

However:

Can we trust our XAI models?

- Might it be that our XAI models behave like black boxes* aswell ?
- Are we explaining a black box with a black box ?
- Do we need **trustworthy XAI** ?

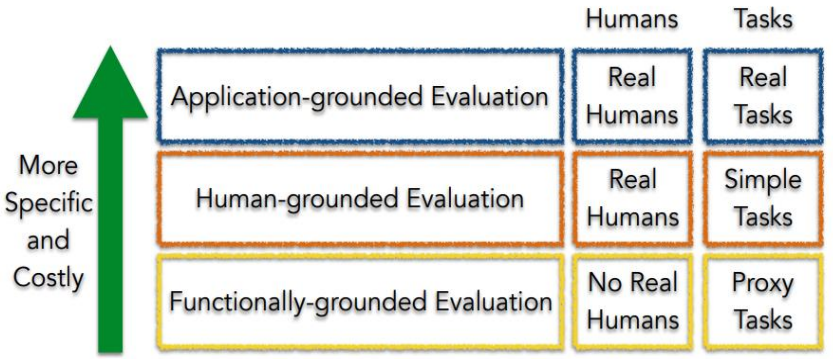
* hyperparameters, interpretation of explanations, SW libraries

Let‘s call this:

The nested-black-box problem

XAI: What‘s left for us to do… some thoughts (Problem 5)

- Application-/Human-grounded evaluation is often avoided in papers
  - e.g. from > 60 time series-specific XAI approaches reviewed in (Theissler et al., 2022) only four were evaluated with a user study
- Functionally-grounded evaluation offers a variety of metrics



(Doshi-Velez and Kim, 2017)

1. **quantitative metrics presumably needs to be developed further (see e.g. (Schlegel and Keim, 2023)), but $understand(\varepsilon_i(\dots))$ is hard to quantify**
2. **users should be involved in testing**
3. **Should we really automatically view ante-hoc methods as interpretable?**
   - e.g. deep decision trees, shapelets for time series, prototypes retrieved from latent space, …
   - Note that “transparency” has been discussed at three levels in (Lipton , 2018). Adding explainability to a black-box does not necessarily make the entire model understandable, but rather sheds light on specific parts of the model or the model’s decisions.
   - i.e. testing these methods regarding loss of accuracy w.r.t. non-interpretable model might not be sufficient

could be a whole talk (workshop day) on its own

Let‘s call this:

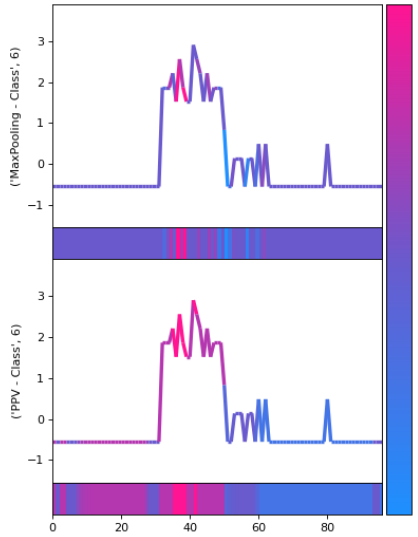
The evaluation problem

Some more thoughts…

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

Aalen University

Higher-order explanations



Can we build “higher-order explanations”?

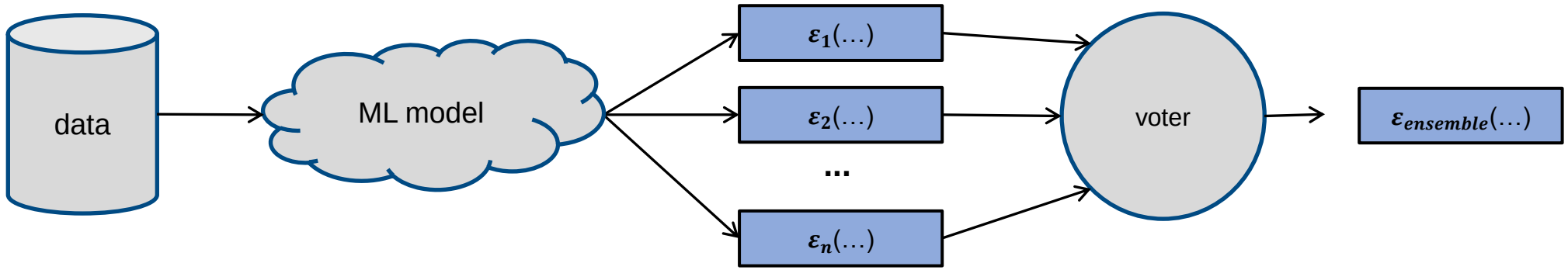
- Using higher-level ways of explanations in addition to saliency maps

- Textual explanations, possibly in domain-specific terminology, seem to be a promising direction.
- Concepts seem promising
- Factuals and counterfactuals seem promising

Ensemble XAI

If we think of machine learning, one option to tackle the variance problem are ensembles.

Can we exploit the idea of ensembles for XAI?



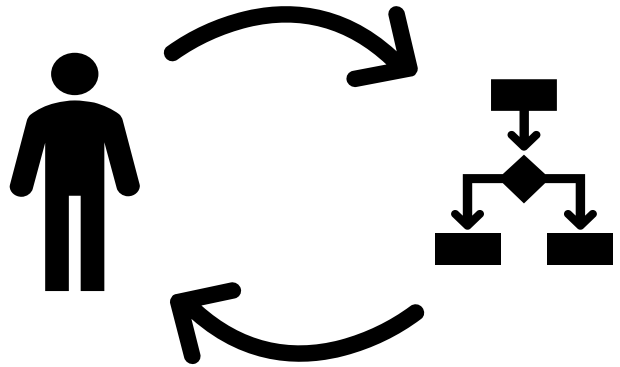
(We meant to call our hybrid SHAP-based explanation for EEG data an “ensemble” –

but we chickened out because we only had two explainers…)

Guide model training with XAI

Can we exploit XAI for more efficient ML model training?

- to guide the model
- to overrule the model
- to evaluate and adapt the training data



(small-scale research project started)

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

Observation:

- large part of work in XAI done on tabular data and computer vision

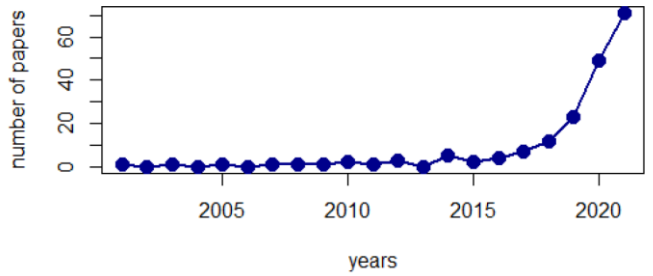
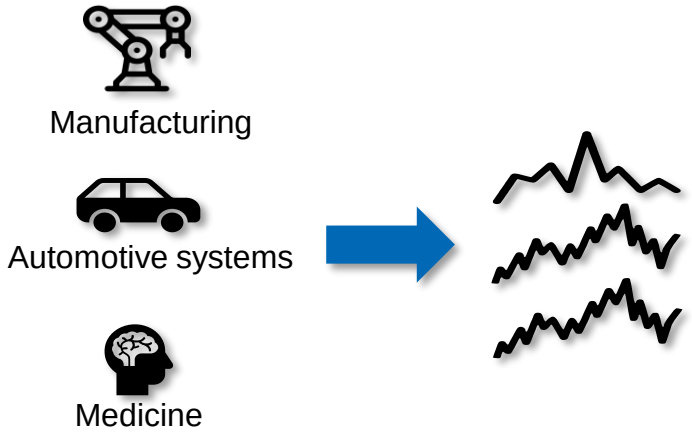


FIGURE 1. The number of papers published per year on XAI for time series classification started to increase significantly in 2019, suggesting an increase in the topic's relevance. The search was performed on Scopus

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. for computer vision)

But: time series are omnipresent



and many more…

Conclusion

- XAI may be one cornerstone to make AI trustworthy, as demanded e.g. by EU legislation and by many domain experts.
- And we have come a long way
  - see e.g. „Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence" for a current state-of-the-art
- We need to close the gap between $\boldsymbol{\varepsilon_i(\dots)}$ and $\boldsymbol{understand(\varepsilon_i(\dots))}$

So we need to solve plenty of research problems on the way, some of them named here:

- non-inherent semantics problem
- nested-black box problem
- dynamic-target-user problem
- variance problem
- evaluation problem (that's a tough one)

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