Leveraging Human-Centered Machine Learning to Create More Explainable Machine Learning Models

Bahavathy Kathirgamanathan, Fraunhofer IAIS, Bonn Germany

Partner institutions:

technische universität dortmund

Fraunhofer IAIS

UNIVERSITÄT BONN

Fraunhofer IML

Institutionally funded by:

Bundesministerium für Bildung und Forschung

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

LAMARR
INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Domain knowledge insertion to create more explainable models

Motivation

- Machine Learning models developed in a purely data-driven way may perform well, however they have some limitations
- Data (e.g., numeric values at fine-grained time steps) used for training may not be meaningful to a human

Challenge: The model cannot be sensibly explained

- Standard numeric measures of model quality lack domain-specific meaning

Challenge: They do not help to detect problems and find ways to improve

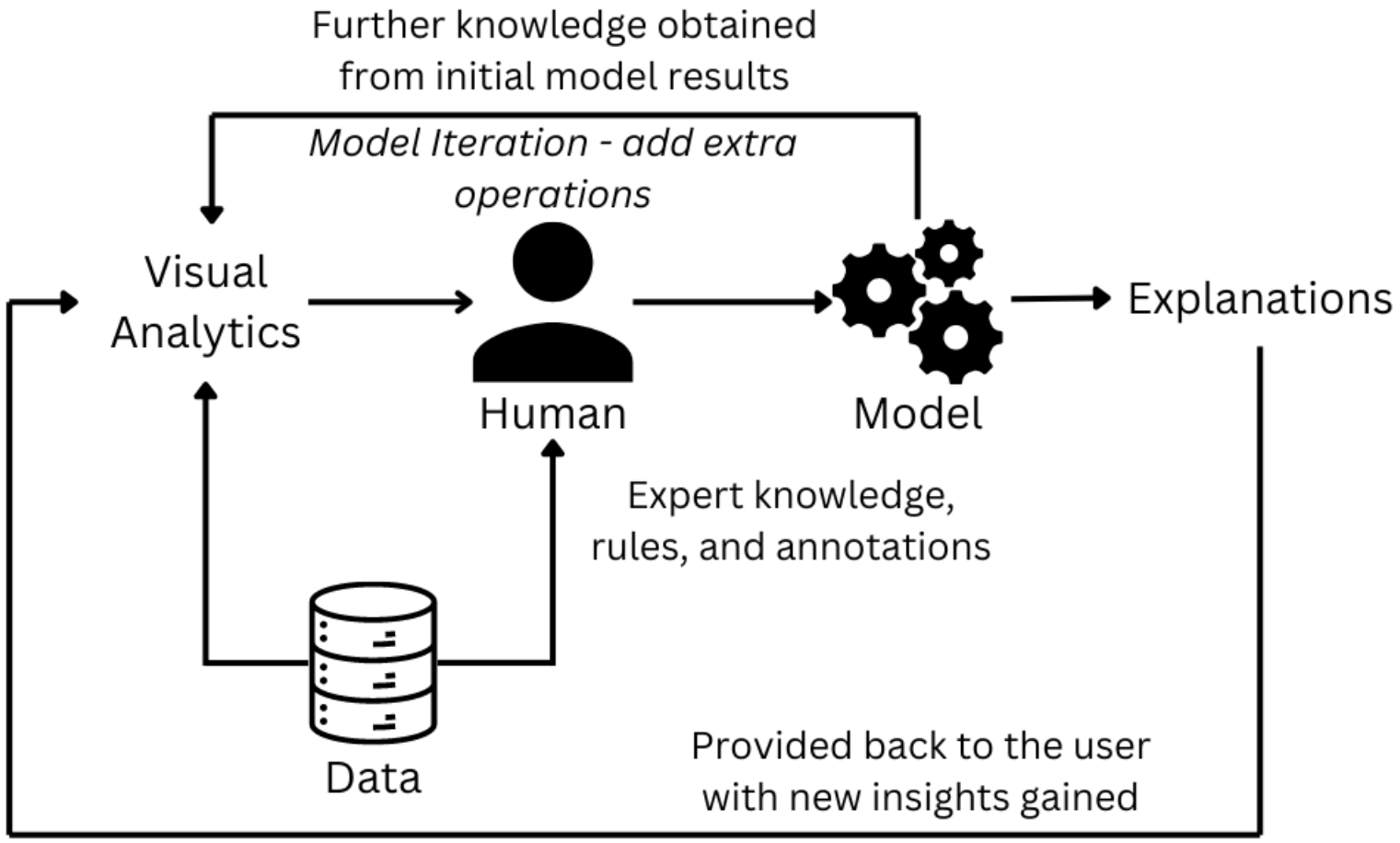
Human-centric machine learning combines computational power with human intelligence and creates explainable and accountable models aligned with human knowledge and logic

Our approach

Employ **interactive visual interfaces** to involve **human expertise** in the model building process

- **Transform raw data** to meaningful structures that match domain concepts
- Investigate model output and provide **expert feedback** for iterative model refinement

The approach is designed to be suitable to be used with a variety of ML algorithms.



Augmenting the model with human knowledge

1. Data abstraction and definition of appropriate units

Time series data –> Time intervals
Geographic data -> Spatial Areas
Social data -> Groups

2. Contextualization

Time series data -> preceding values or trends
Geographic data -> neighboring areas or regions
Social data -> strongly connected nodes

3. Synoptic Feature Generation

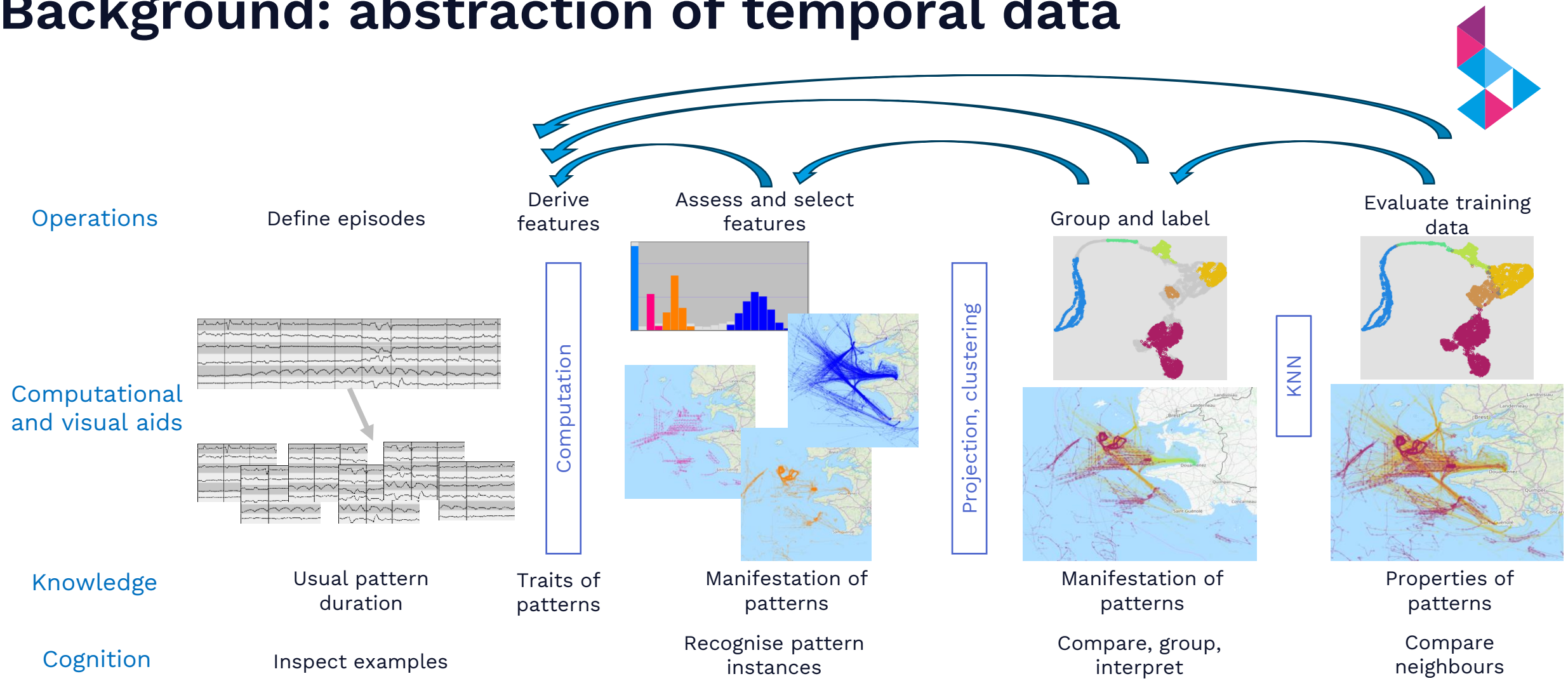
Time series data -> statistical summaries, trend indicators
Geographic data -> internal composition or densities
Social data -> aggregate measures about characteristics of a group

4. Iterative, Human-Controlled Workflow and Insight Injection

For all data types:

- *Correct issues with raw data*
- *introduce new features or adjust based on insights*
- *Validate that the data and model aligns with domain knowledge*

Background: abstraction of temporal data



Natalia Andrienko, Gennady Andrienko, Alexander Artikis, Periklis Mantenoglou, Salvatore Rinzivillo
Human-in-the-Loop: Visual Analytics for Building Models Recognising Behavioural Patterns in Time Series
IEEE Computer Graphics & Applications , 2024, **44**(3), 14-29 https://doi.org/10.1109/MCG.2024.3379851

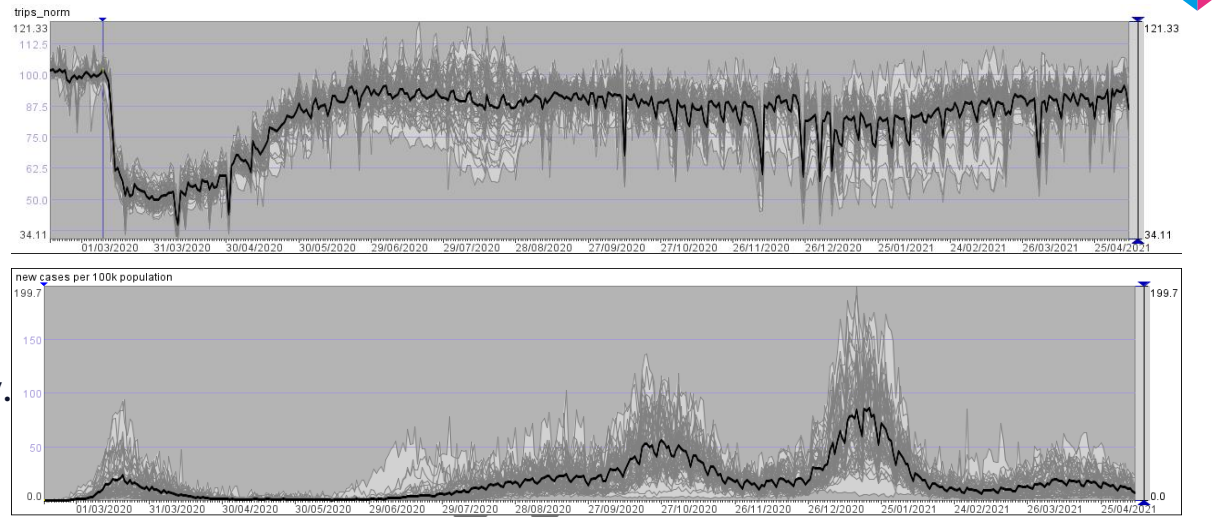
However, this framework focuses on data preparation without addressing model development

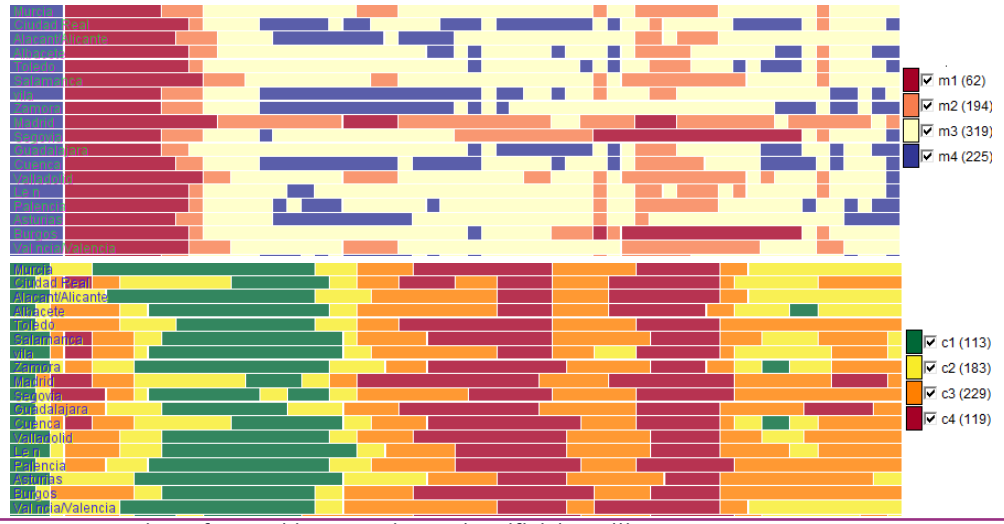
Case Study: Two-way relationships between population mobility behavior and the development of the COVID-19 pandemic

Data: Time series of daily values of Mobility and COVID cases for the provinces of Spain for the period from 17/02/2020 to 09/05/2021 [1,2]

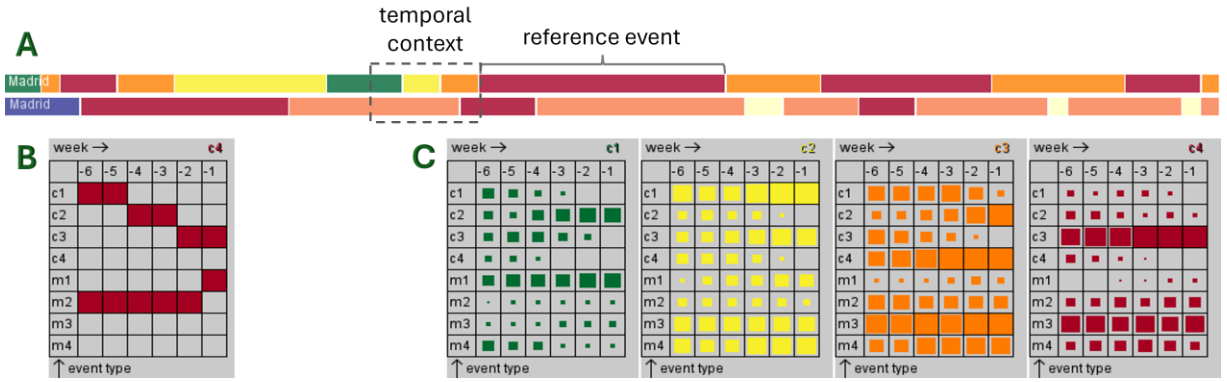
Data Pre-processing: Smoothed and divided into episodes with a length of 7 days.

Temporal abstraction: Each episode is assigned a class based on the (1) level of COVID-19 and (2) level of mobility. Sequences of same class episodes united into *events*.





Created features: Temporal contexts of COVID-19 events



Case Study: Initial Model Development

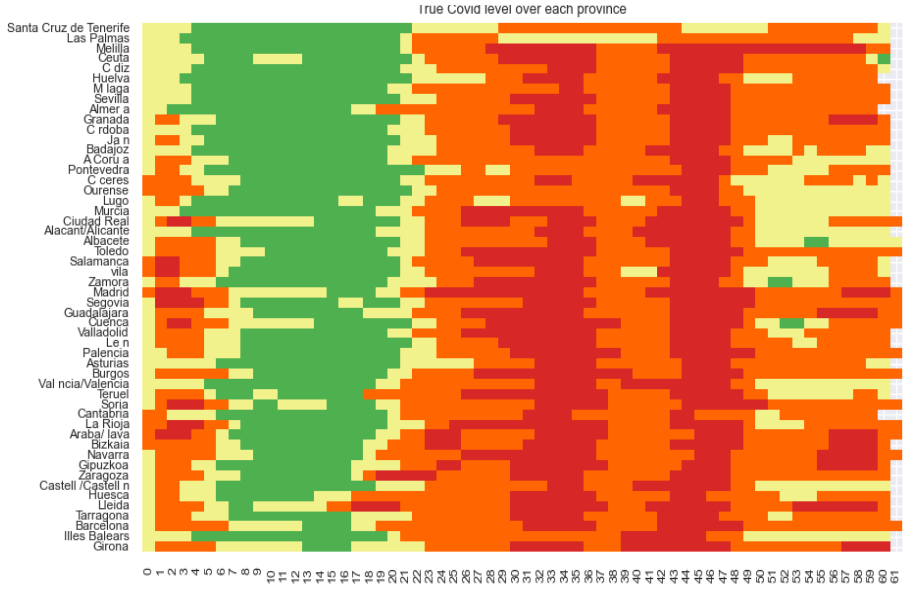
Modelling Task: Use historical data from the prior six weeks of event contexts to predict the current level of COVID-19.

Machine Learning Model used: Random Forest model using a 5-fold cross validation strategy

Model Output: Base model accuracy of 0.71

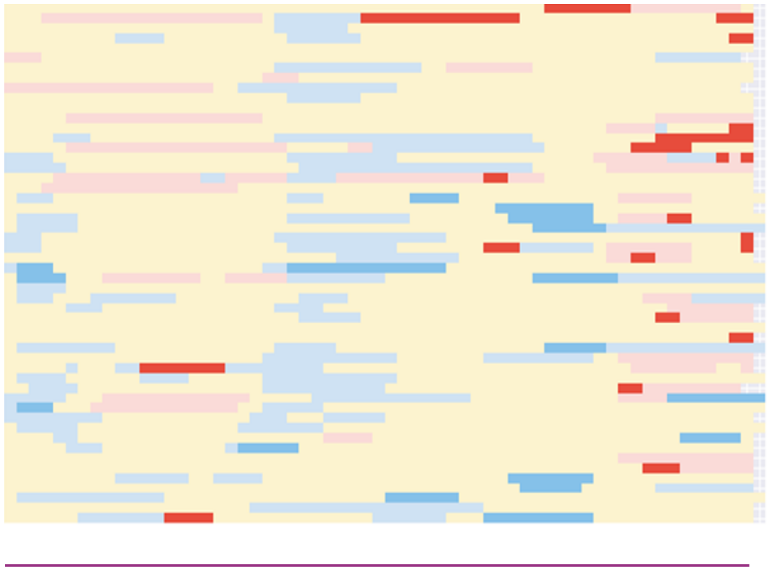
Visual analytics techniques are used to explore the spatio-temporal distribution of the model errors. Detected patterns provide clues to help a human understand what and how to improve.

True classes



Predicted classes

Differences



Iterative Model Development

Summary of the models tested

- Human insights are incorporated into the model through data operations, aiming for compatibility with common ML techniques.

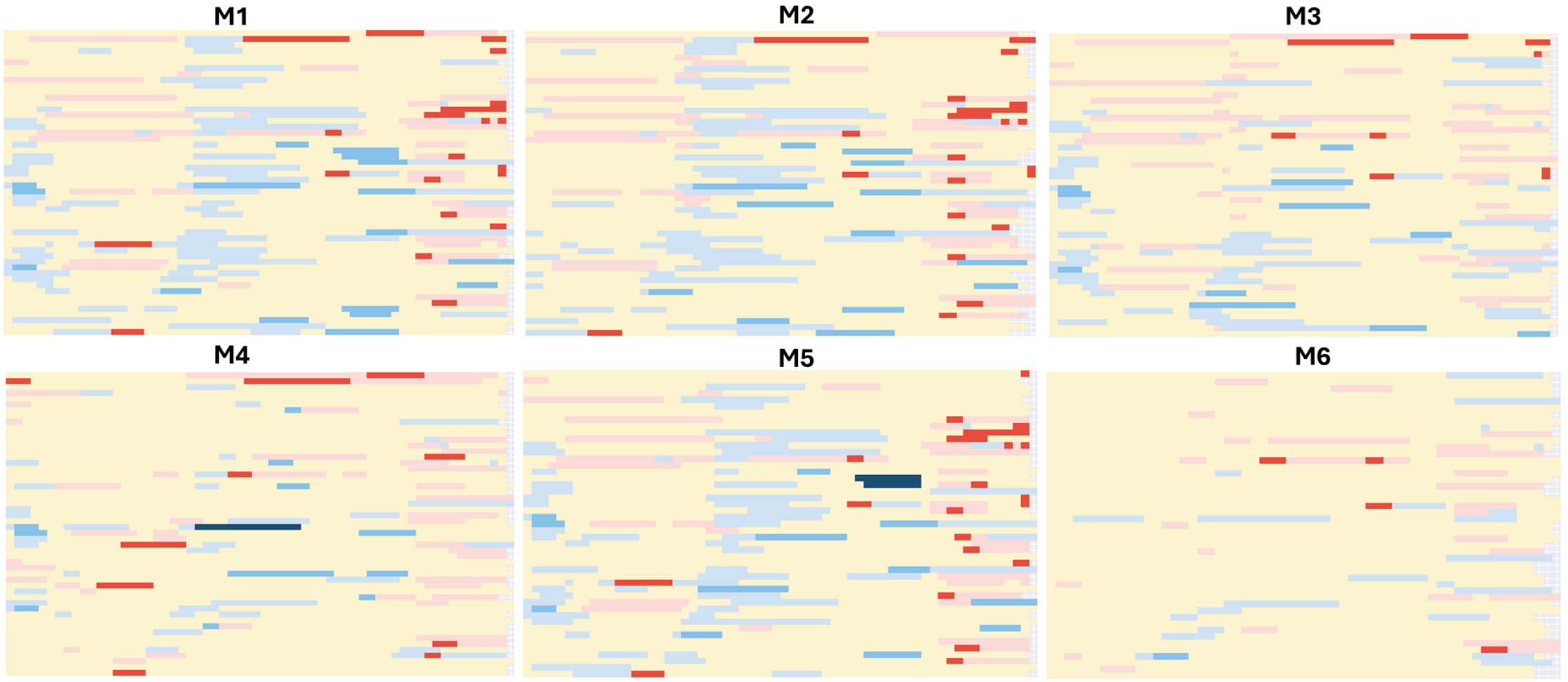
- The operations on the data guided by the iterative visual exploration and human feedback finally led to an improvement of model accuracy from 0.71 to 0.86

Model	Operations Undertaken	5-Fold CV Mean Accuracy
M1	Basic Model – No extra operations	0.71244 ± 0.041
M2	First 30 days removed	0.71827 ± 0.016
M3	Number of weeks since the start added as a feature	0.74798 ± 0.047
M4	Duration of the event added as a feature	0.78847 ± 0.019
M5	All island provinces are removed	0.71636 ± 0.041
M6	All of the above operations	0.86334 ± 0.036

Table 1. A summary of the operations applied during the iterative Model development

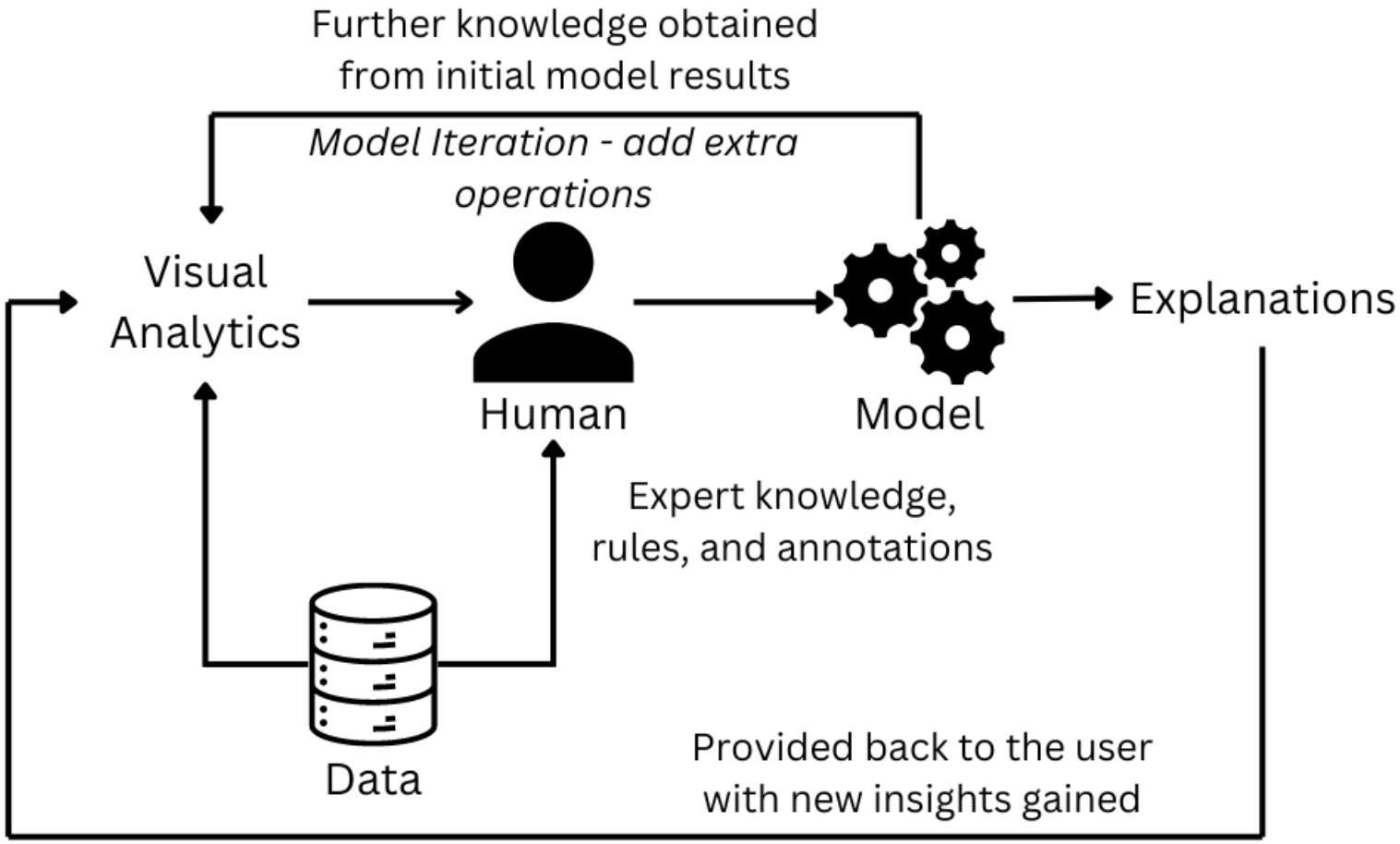
Iterative Model Development

Reducing errors and removing patterns in the error distribution



Proposed General Framework

- Model will be refined based on insights obtained from visualizing the data and/or the model results
- Human knowledge gained can also be used to provide explanations
- Key idea: Visual Analytics enables humans to understand and organize data and define relevant concepts to be used in the ML algorithm and for explanation



LAMARR

INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

A Hybrid Human-Centric approach to combining Rule-based and Attribute based explanations

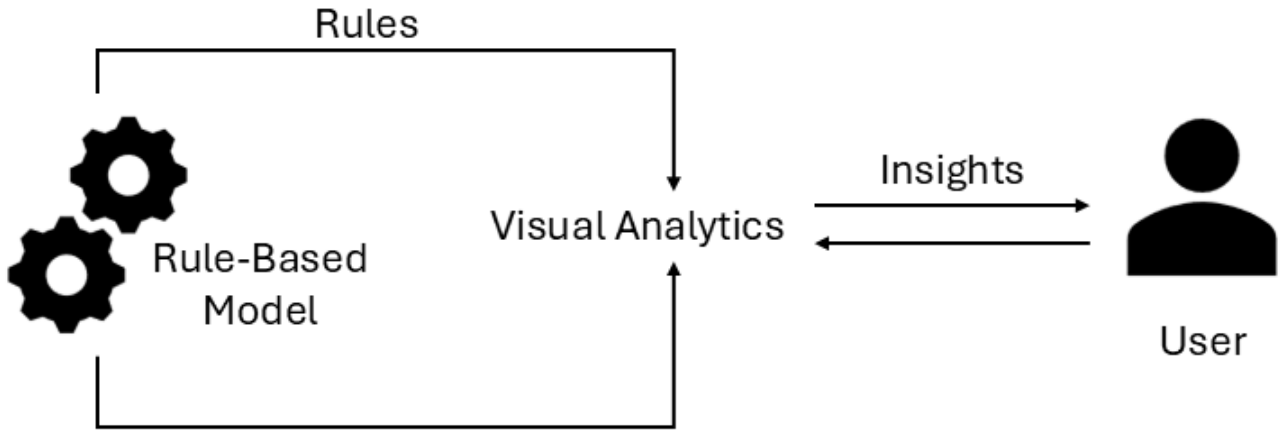
Motivation

- Rule based and Feature attribution methods have distinct strengths and weaknesses.

- Hypothesis: By using them both together, the overall user trust and model interpretability can be improved further

- Motivation:
  - Enhance interpretability
  - Support interactive exploration and alignment

+ Intuitive and align with human reasoning
+ Stable results
- Does not scale well



+ Quantitative values are available
+ Can be summarized simply for any amount of data
- Do not align with human reasoning
- Results may be instable

Proposed Workflow

1. Visualise Rules

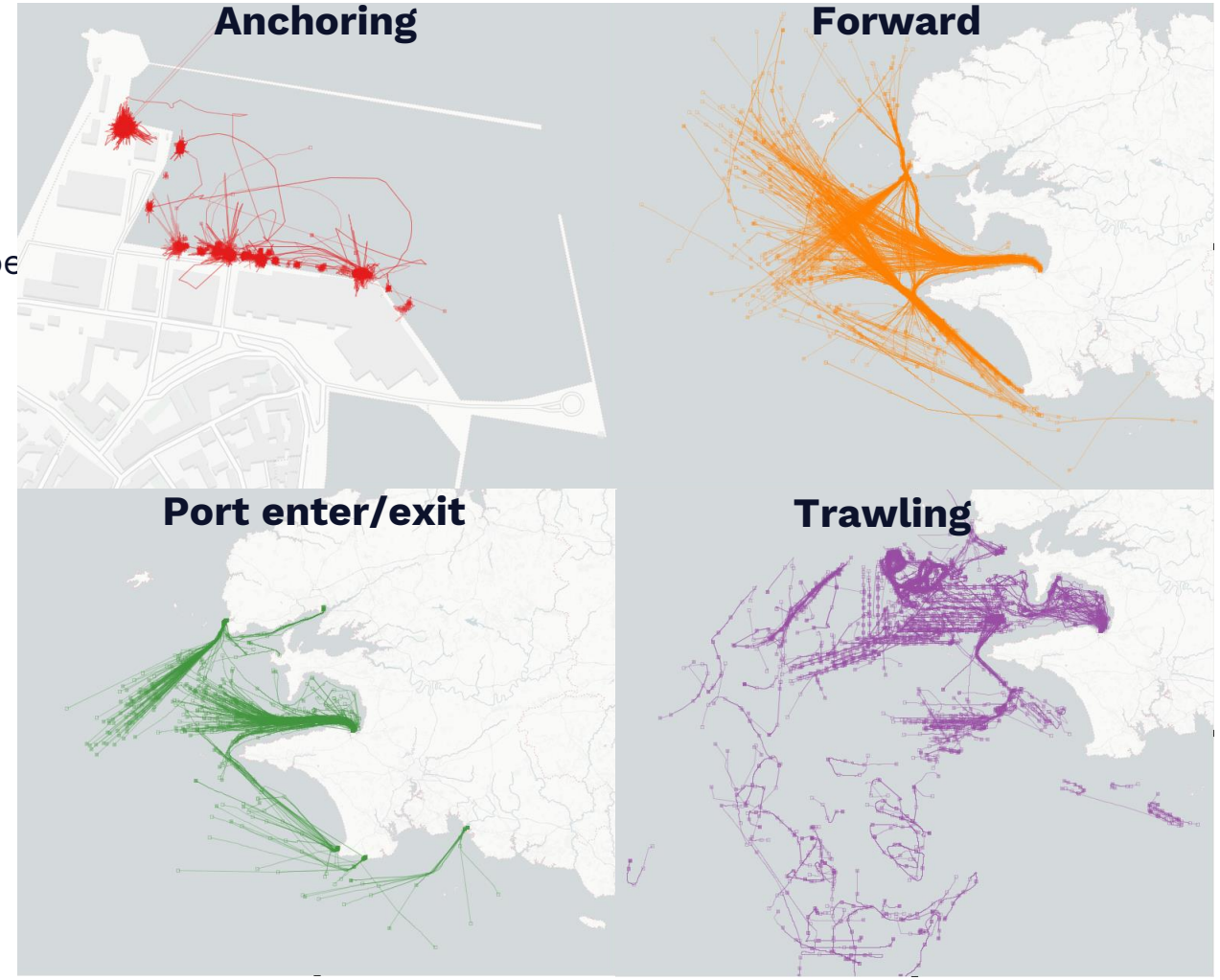
2. Compute and Visualise Feature Attribution Scores

3. Interactively Explore the Ruleset

4. Iterate

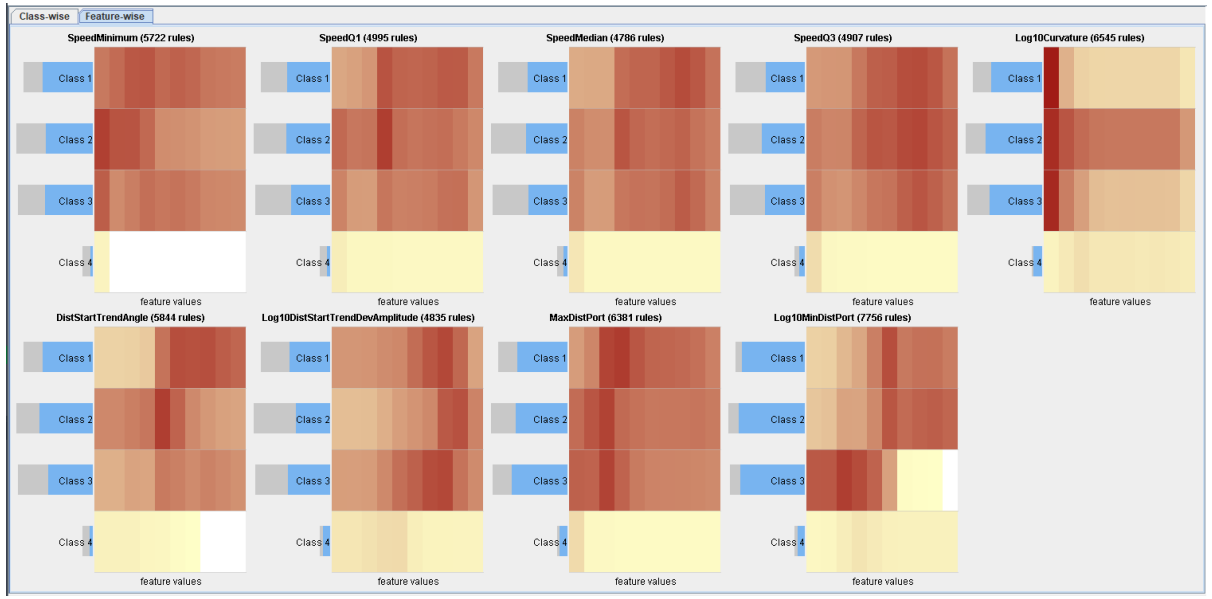
Case Study – Movement Vessel Dataset

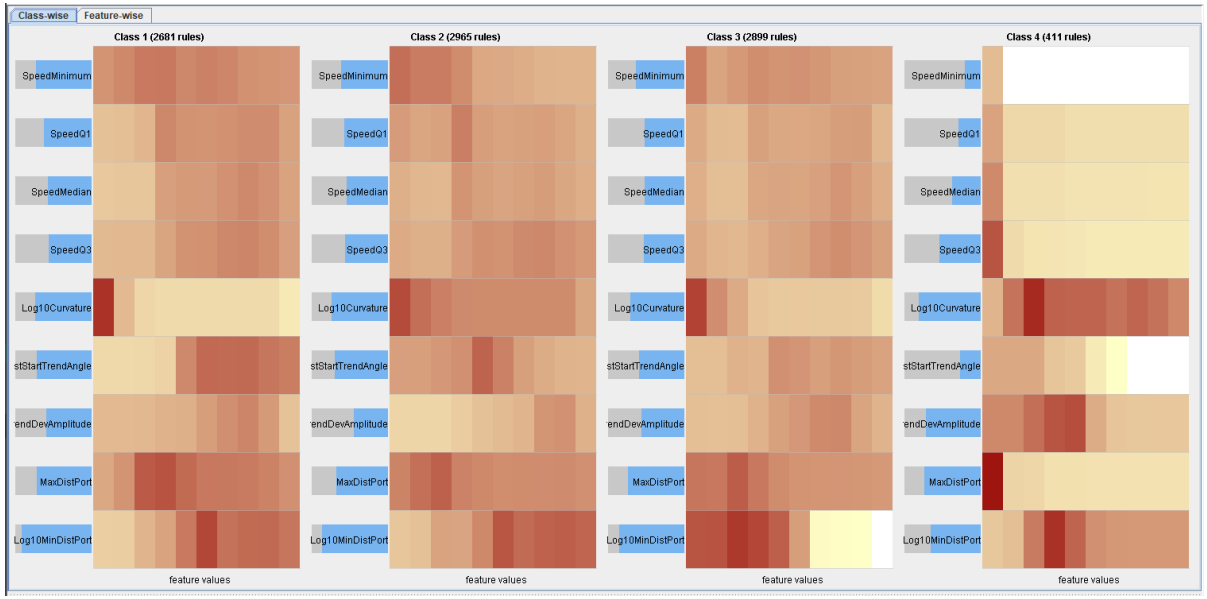
- Dataset with trajectories from fishing vessels in France
- Each episode was represented by 9 features which characterize the speed, trajectory shape and the proximity to the port
- Episodes were classified into four activity classes.
- A random forest classifier was trained achieving an accuracy of **0.97**



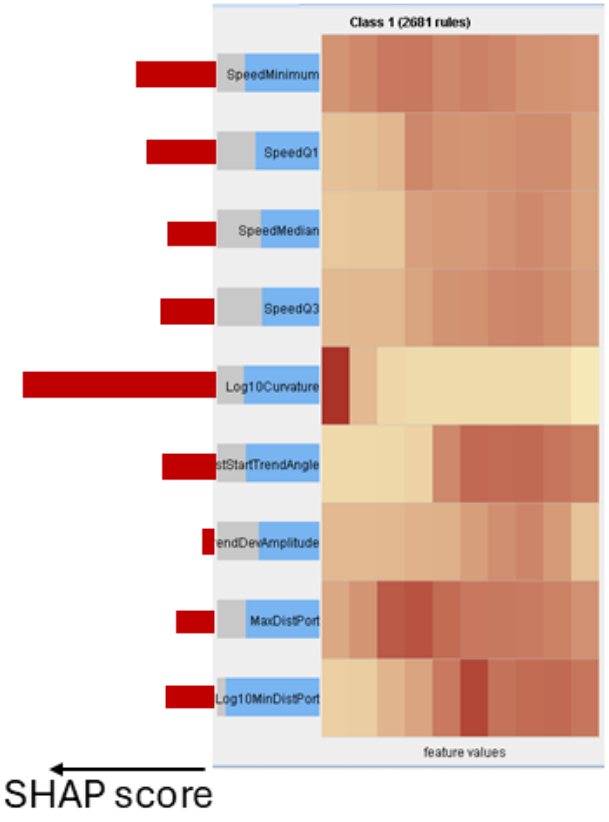
Rule-based explorer

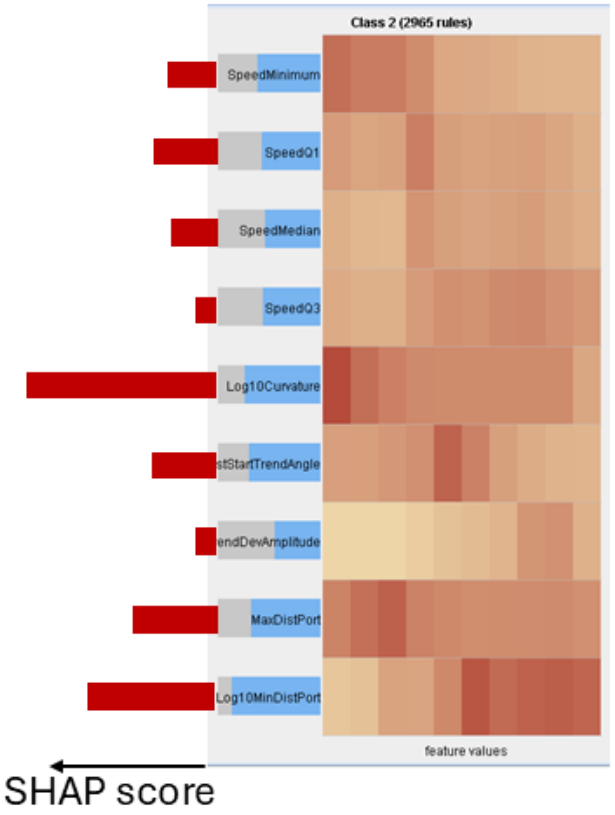
- RuleExplorer – a software prototype to visualize and interact with rules

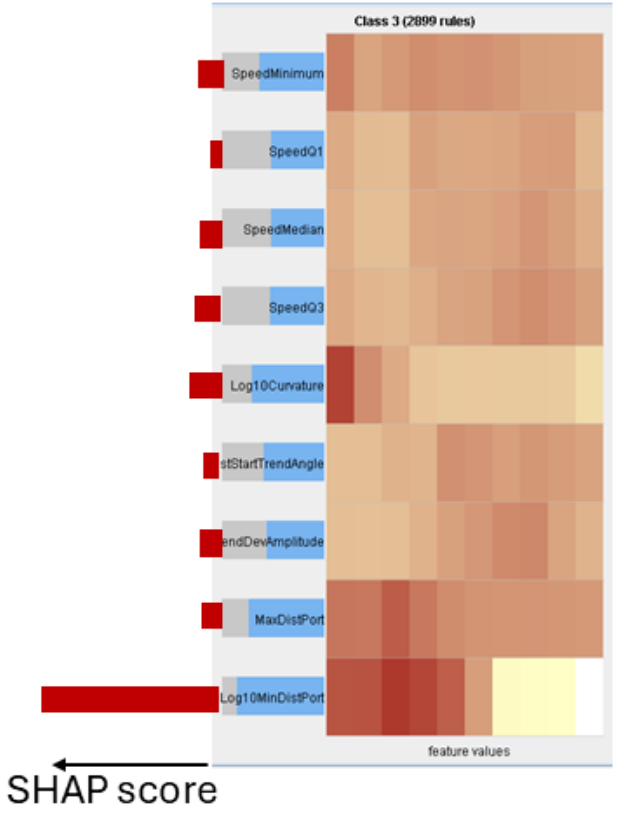


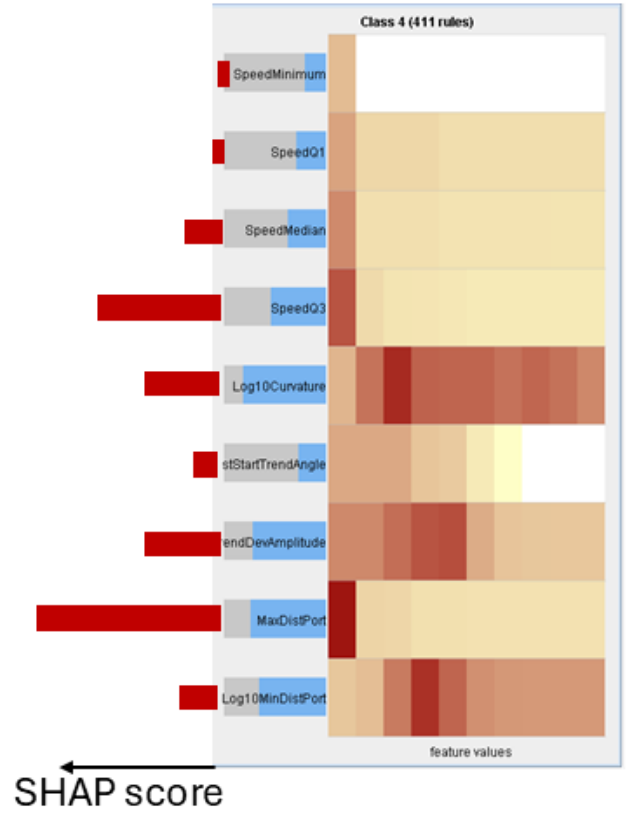


Hybrid View – feature attribution + Rules









Key Objectives

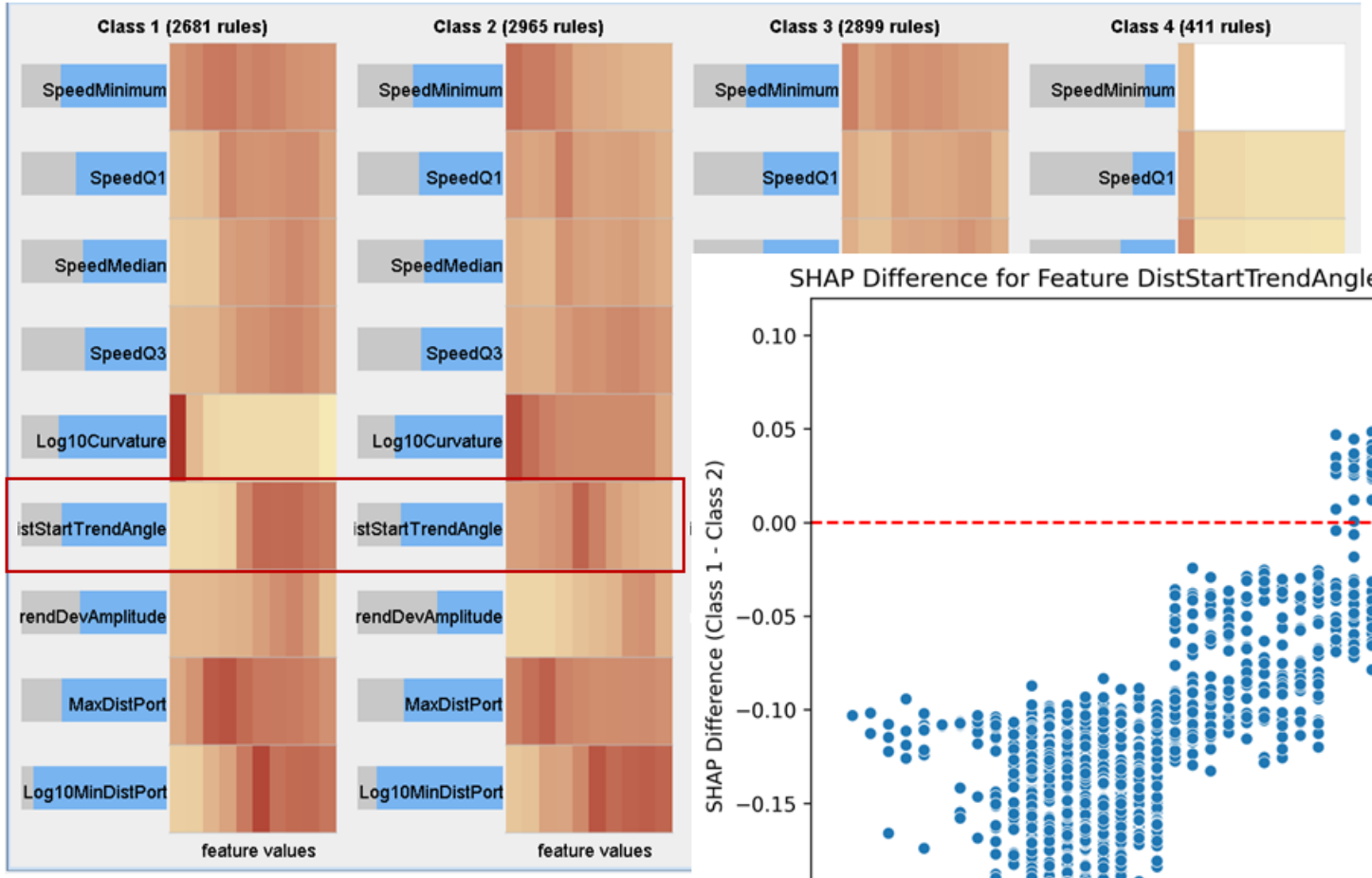
Enhancing interpretability of model behavior:

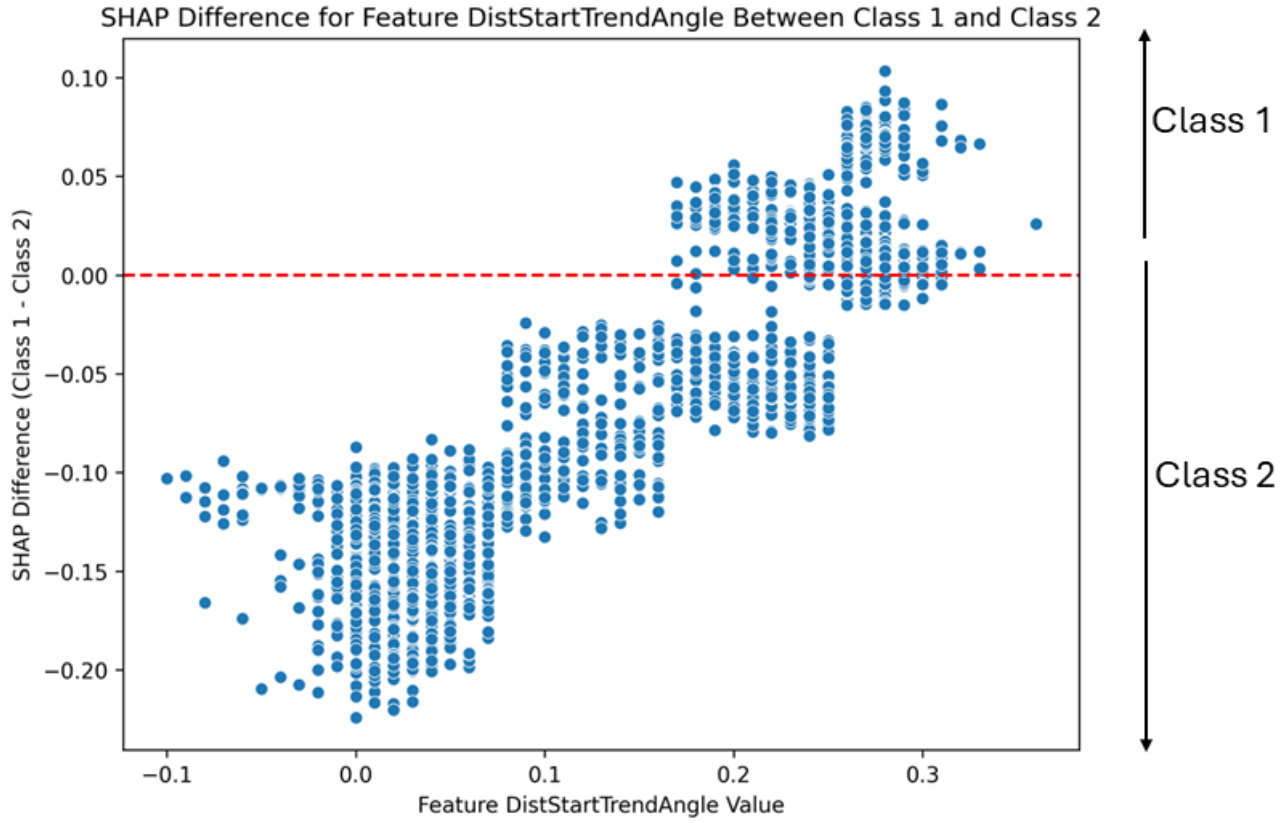
- Feature impact scores provide complementary information that is not directly accessible from rule-based explanations alone.
- By quantifying the contribution of individual features to model predictions, they support a more detailed understanding of the model's reasoning and foster greater transparency.

Supporting interactive exploration:

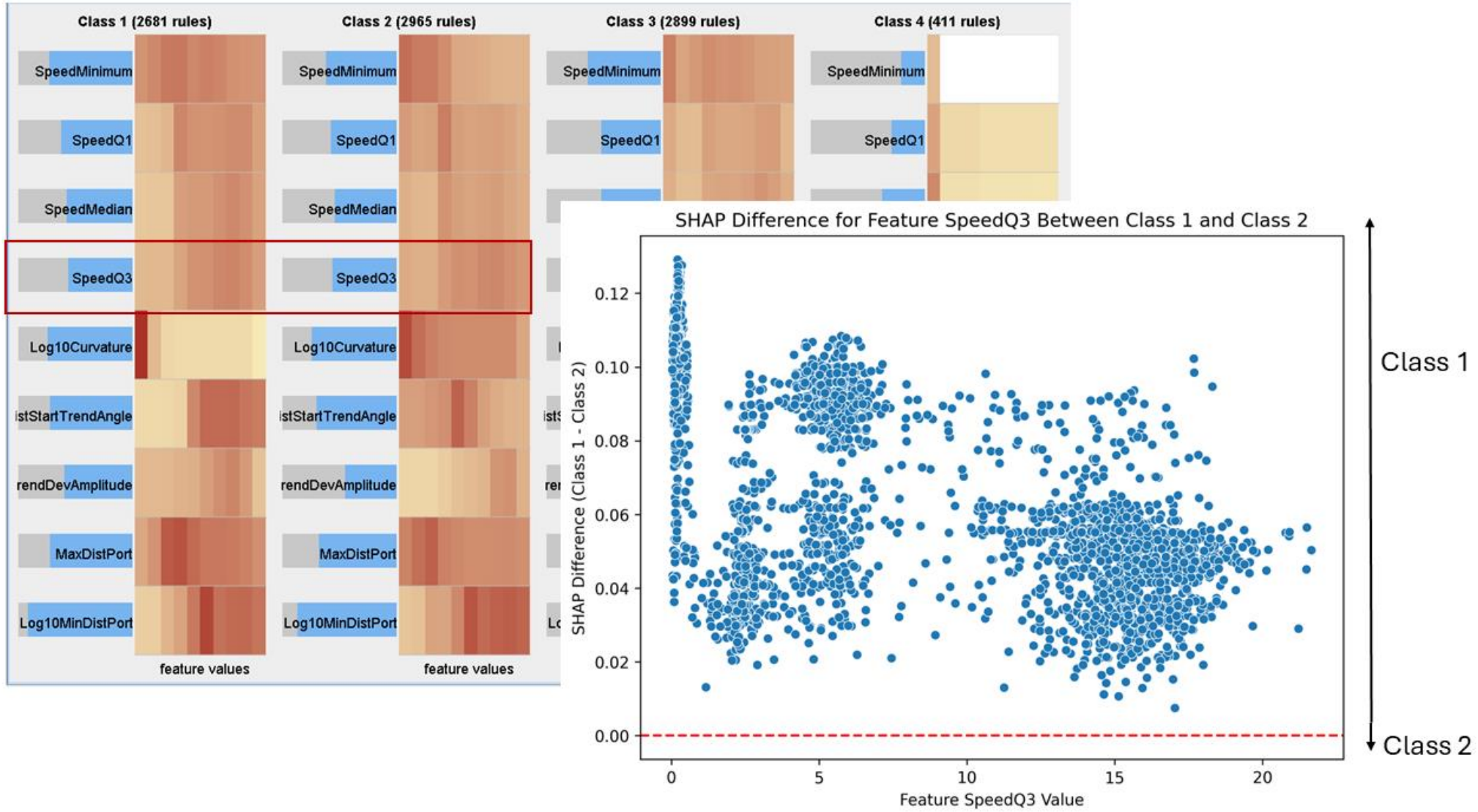
- Awareness of feature importance enables users to more effectively navigate and interrogate the rule space, prioritizing rules that include or omit key features.
- The comparison of rule-based and attribution-based explanations allows users to identify discrepancies and improve mutual trust in the system.

Enhancing Interpretability

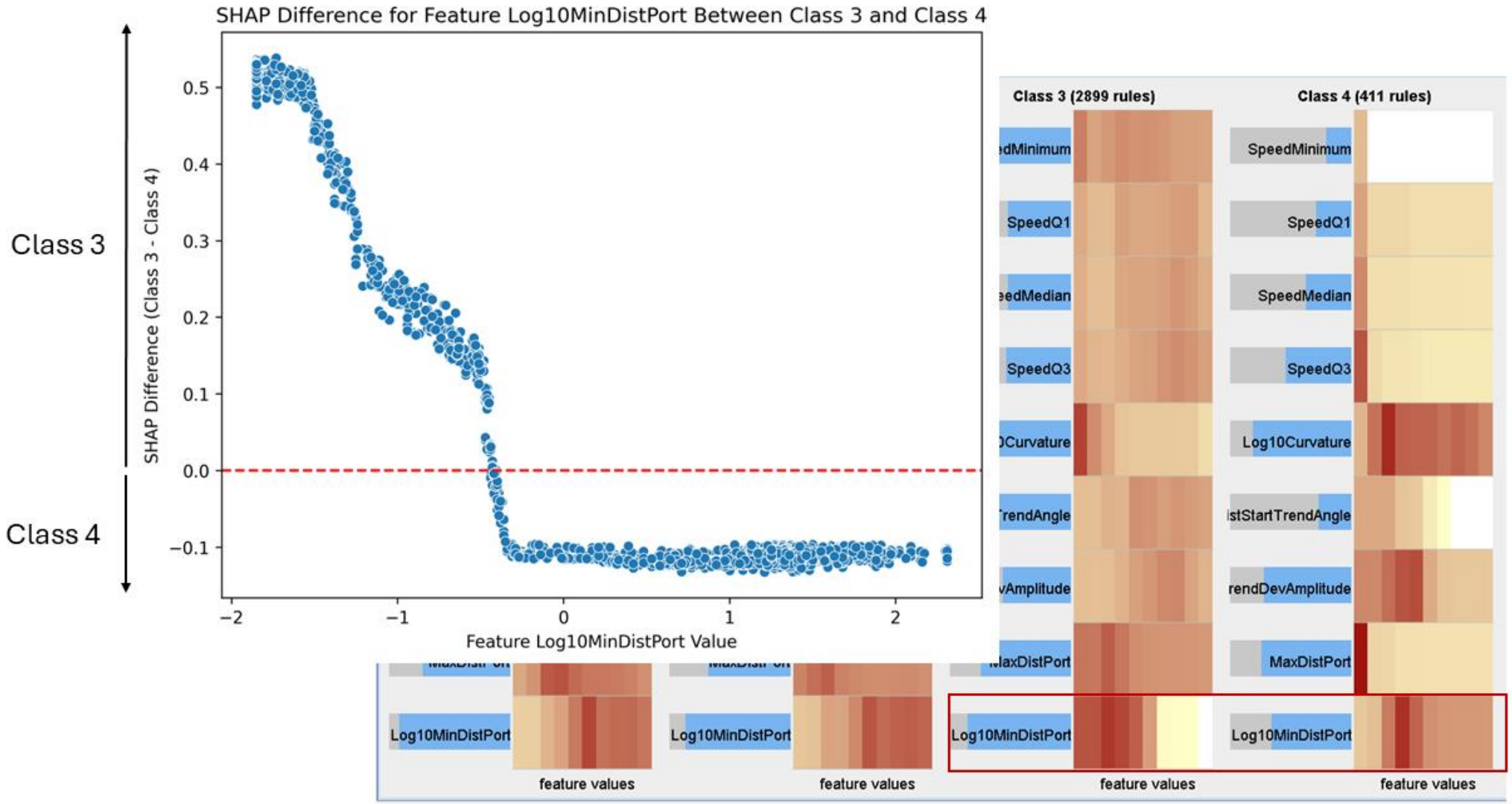




Enhancing Interpretability



Enhancing Interpretability



Limitations and Future work

- Shap stability – Explore stabilization techniques
- Refine visualization of feature attributions
- Generalisation and Validation – look at further datasets
- Provide Local Explanations

Ideas for Collaboration

- Interesting datasets to apply our techniques ?
  - Particularly temporal or spatio-temporal datasets
- Ideas for improving the framework?
- Techniques to compliment the framework?

LAMARR
INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

The Team:

Prof. Dr. Natalia Andrienko

Prof. Dr. Gennady Andrienko

Dr. Bahavathy Kathirgamanathan

Institute: Fraunhofer IAIS, Sankt Augustin, Germany



References

[1] Ponce-de Leon et al., **COVID-19 Flow-Maps an open geographic information system on COVID-19 and human mobility for Spain**. Scientific Data 8(1), 310 (Nov 2021).

[2] Ponce-de Leon et al., **COVID19 Flow-Maps daily cases reports** (2021). https://doi.org/10.5281/zenodo.5217386

LAMARR
INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Contact

Dr. Bahavathy Kathirgamanathan
Scientific Coordinator, Lamarr Human Centered AI Systems

Email:
bahavathy.kathirgamanathan@iais.fraunhofer.de

www.lamarr-institute.org

Partner institutions:

technische universität dortmund

Fraunhofer IAIS

UNIVERSITÄT BONN

Fraunhofer IML

Institutionally funded by:

Bundesministerium für Bildung und Forschung

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen