Validation by Design
Making Machine Learning for Autonomous Driving Interpretable and Validatable

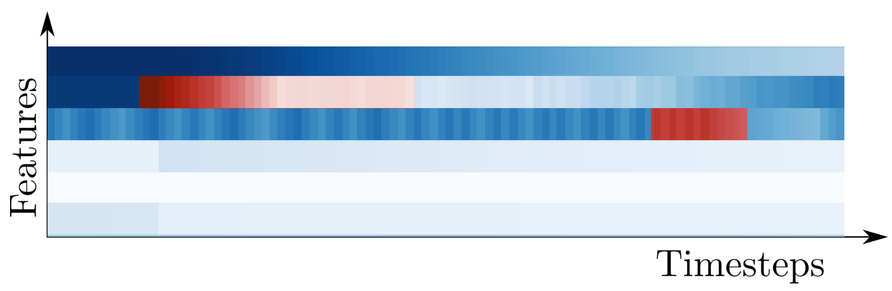
Research Question

- Is safe artificial intelligence in autonomous driving possible?
- How can foreseeable and interpretable behavior be ensured even before delivery?
- Can interpretability and performance of a machine learning (ML) algorithm be complementary?

Feature Generation Method [1]

- Use intrinsic properties of ML structures to establish interpretability.
- Best of both worlds: Fuse deep learning (Convolutional Neural Networks, Recurrent Architectures etc.) and classical methods (Random Forests, Mixture of Experts etc.) for best performance.
- Visualization as valuable byproduct: Applied interpretability methods generate visualizations for more insight.
- Layerwise Relevance Propagation highlights salient regions in input according to

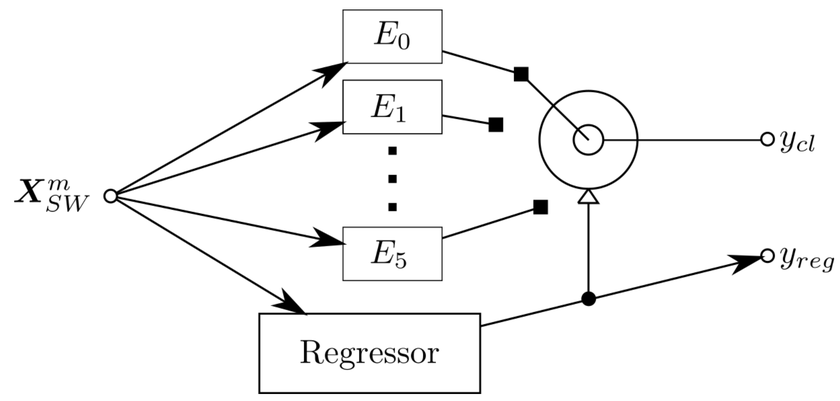
$$R_i^{(l)} = \sum_j \left(\alpha \cdot \frac{z_{ij}^+}{\sum_i z_{ij}^+} + \beta \cdot \frac{z_{ij}^-}{\sum_i z_{ij}^-} \right) R_j^{(l+1)}, \tag{1}$$



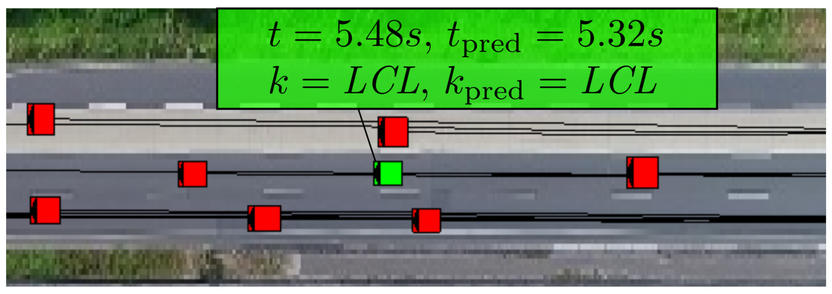
- Interpretable feature generation method facilitates enriched datasets wile remaining fully interpretable.

Mixture of Experts (MoE) Architecture [3]

- Early classification to improve driving comfort & safety
- Experts specialized on different prediction horizons.



- Interpretable expert classifiers, e.g. Decision Trees
- Small trees $\rightarrow$ better interpretability
- Focus of early experts: vehicle constellations
- Focus of late experts: acceleration profiles
- **Exemplary Lane Change Scenario**
  Maneuver: Lane Change Left (LCL)



- Regressor correctly assigns early expert
- Vehicle constellation: Slow leading vehicle, left lane fast
- Decision: Lane Change Left after current vehicle on lane passed

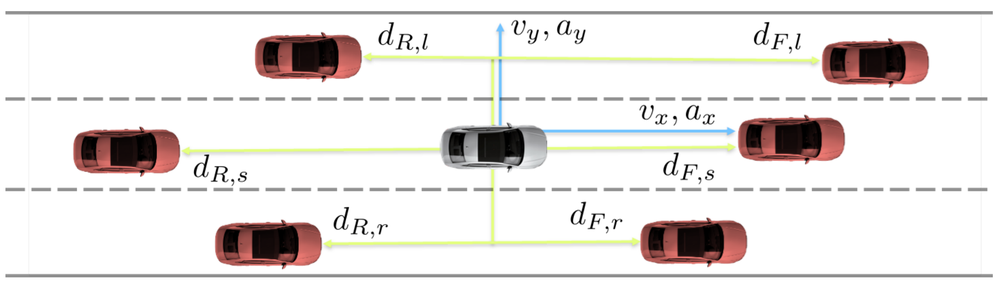
Dataset

- Public dataset highD [2] for reproducability
- Multivariate time series of lane changes
- Classification labels:
  Lane change direction left *LCL*, right *LCR* and no lane change *NLC*.
- Regression labels: Time to lane change in seconds.
- Dataset split 70/20/10 into training, validation and test set, containing samples according to

	LCR	NLC	LCR	Total
Training	1548	1548	1548	4644
Validation	449	449	449	1347
Test	209	209	209	627

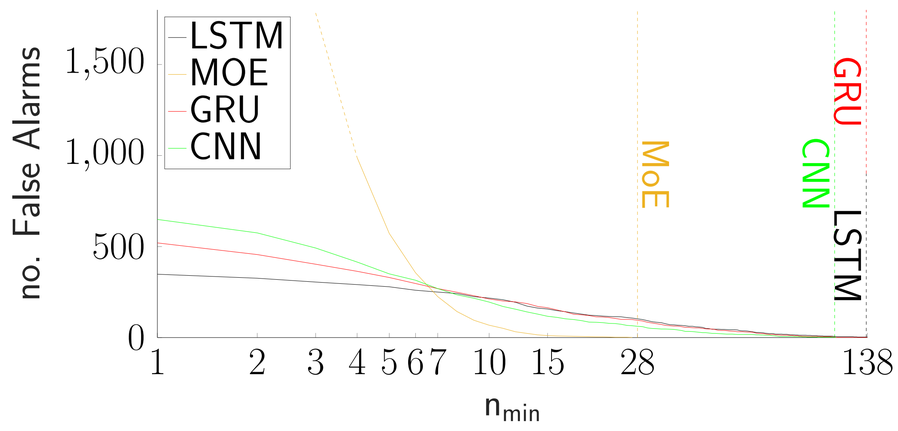
- Feature vector is describing the vehicle constellation and dynamic properties.
- A single sample with F features and T discrete timesteps is defined as

$$\boldsymbol{X}^m \in \mathbb{R}^{F \times T}, \tag{2}$$



Results

- End-to-end interpretable approach for early detection
- Smoothing by n_{min} subsequent identical decisions



- MoE false alarm rate outperforms reference methods
- Mean reliable prediction time μ_{trel} competitive with reference methods

Ref. Method	CNN	GRU	LSTM	MoE
False Alarms	62	95	103	0
μ_{trel}	3.89s	3.91s	3.84s	3.44s

References

[1] Oliver Gallitz et al. Interpretable feature generation using deep neural networks and its application to lane change detection. In *IEEE ITSC*, pages 3405–3411, 2019.

[2] Robert Krajewski et al. The highd dataset. In *IEEE ITSC*, pages 2118–2125, 2018.

[3] Oliver Gallitz et al. Interpretable machine learning structure for an early prediction of lane changes. Lecture Notes in Computer Science, pages 337–349. Springer, 2020.

Oliver Gallitz, Technische Hochschule Ingolstadt
Michael Botsch, Technische Hochschule Ingolstadt
Wolfgang Utschick, Technical University of Munich

CARISSMA Automotive Safety Research
Technische Hochschule Ingolstadt
BayWISS-Kolleg Mobilität & Verkehr
BayWISS Bayerisches Wissenschaftsforum