CIML-R: Causally Informed Machine Learning Based on Feature Relevance

Martin Surner,Abdelmajid Khelil

Institute for Data and Process Science, Computer Science Department
Landshut University of Applied Sciences, Am Lurzenhof 1, 84036 Landshut, Germany

Swiss Conference on Data Science 2024 (SDS2024)

31 May 2024

Table of Contents

Motivation

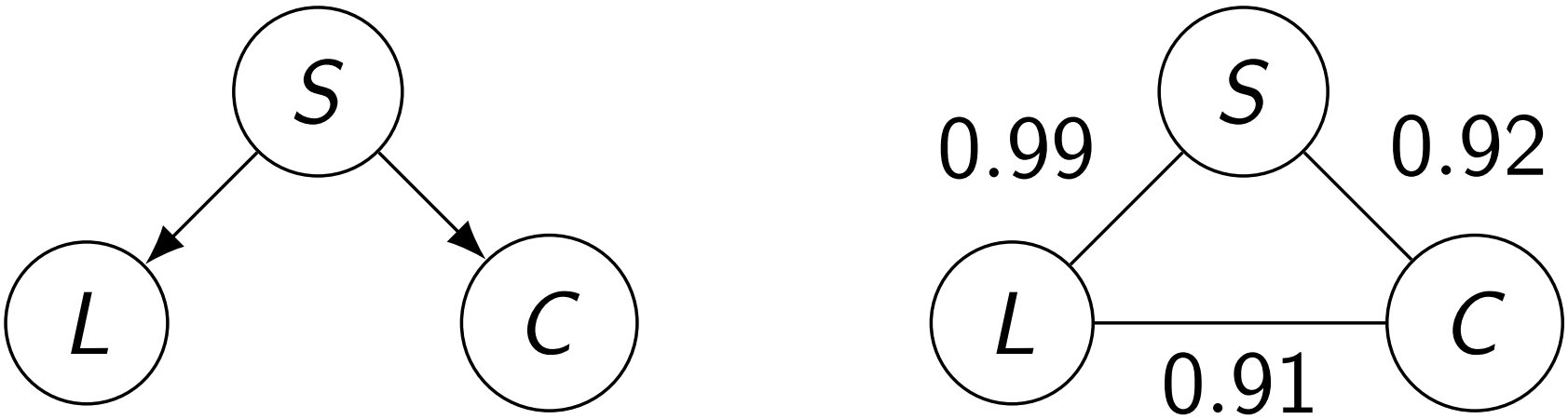


Figure: Causal (left) and correlational (right) connection of variables smoking (S), lighter (L) and cancer (C).

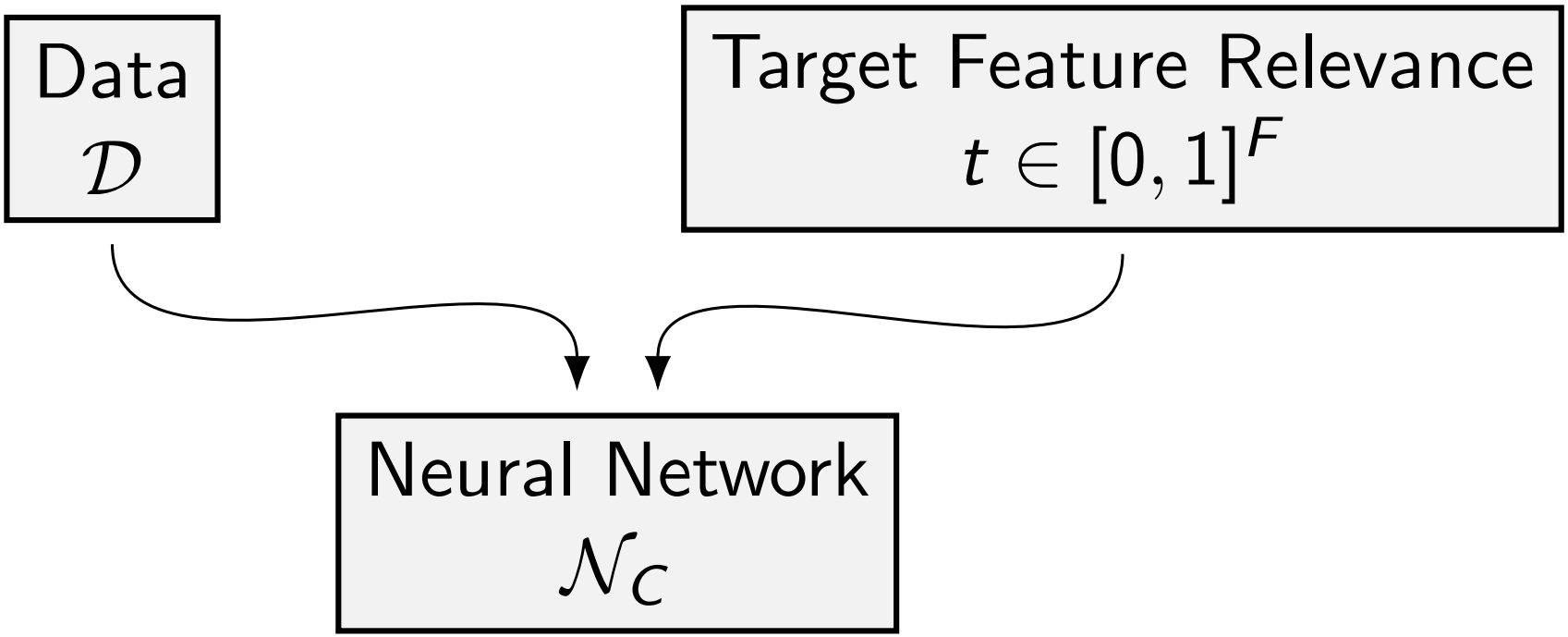
- We should inform the model of the causal structure.

Pearl and Mackenzie 2018

Related Work

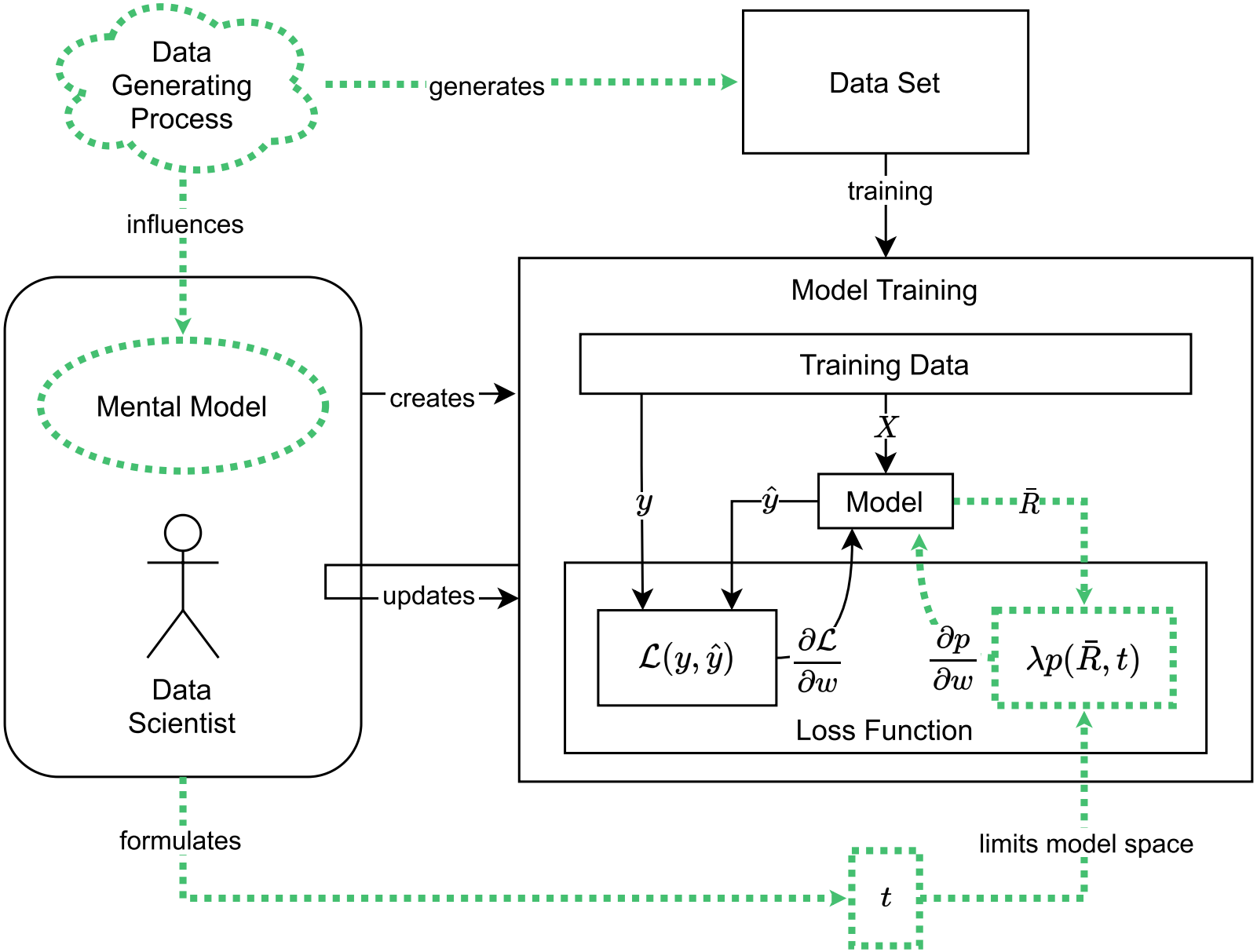
- **Causality:** Shows limits of statistical learning and motivates for informed machine learning.
- **Domain Generalization:** Used to test interventional setting.
- **Explanation Guided Learning:** Methodology for informing the neural network.

Causally Informed Machine Learning



- Target feature relevance t informs $\mathcal{N}_C$.

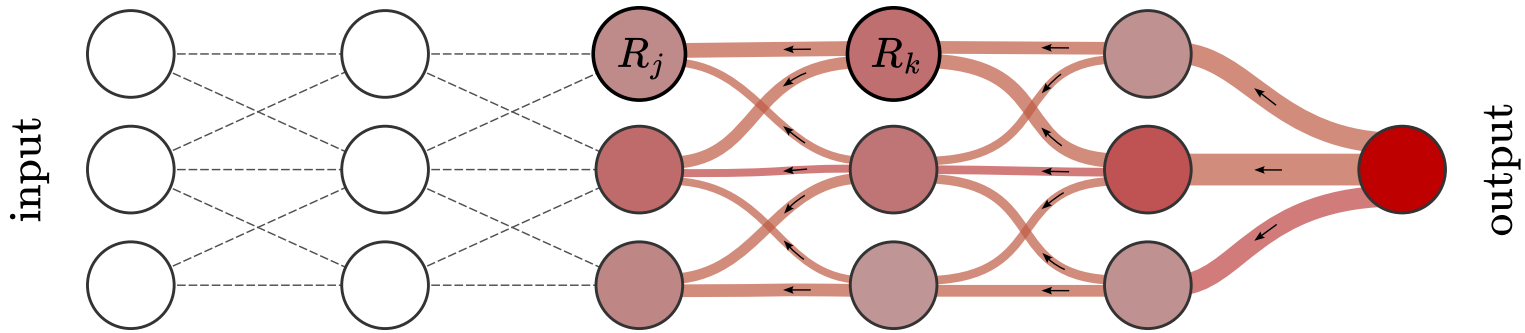
System Model



Target Feature Relevance t

- Motiviated by the Average Causal Effect (ACE).
- Formulated based on the prior knowledge / (human) mental model.
- E.g. $t := (1, 0)$ for (S, L).

Layer-Wise Relevance Progagation (LRP)



LRP-0 Rule

$$R_j = \sum_k \frac{a_j \cdot w_{jk}}{\sum_i a_i \cdot w_{ik}} R_k$$

- We use LRP to compare the actual feature relevance R to t.

Montavon et al. 2019

Loss Formulation

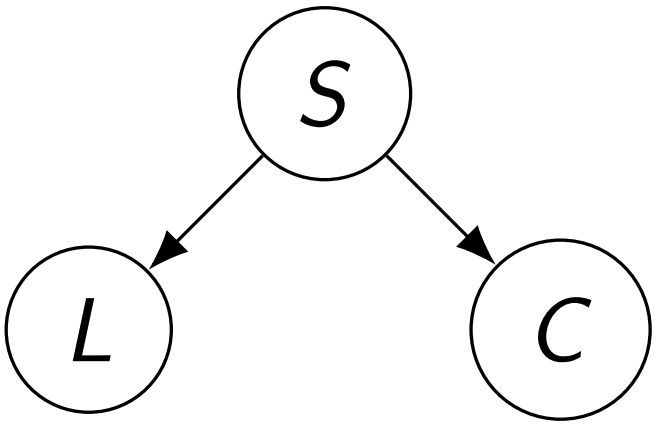
Mean Absolute Error Penalty

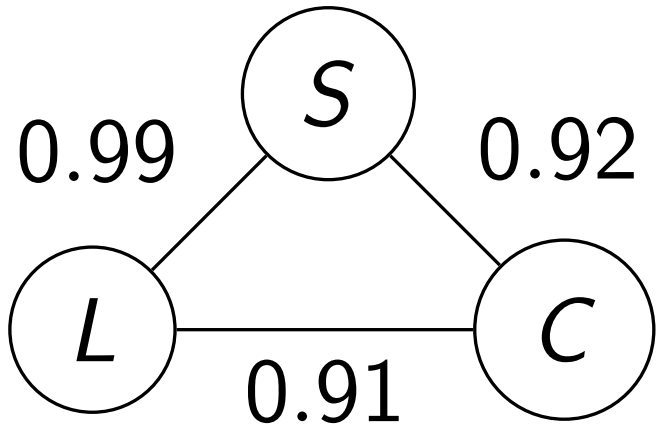
$$p(\bar{R}, t) = \frac{1}{n} \sum_{i=1}^{n} \left| \bar{R}_i - t_i \right|$$

Custom Loss Function

$$\mathcal{L}_C(\hat{y}, y, \bar{R}, t) = \mathcal{L}(\hat{y}, y) + \lambda p\left(\bar{R}, t\right)$$

Experimental Setup (Data Set)





- Based on smoking example (with S and L as variables).
- Two validation sets (identically and non-identically distributed).
- Shows whether mechanism $S \rightarrow C$ was learned.

Experimental Setup (Data Generation)

Observational Validation Set $\mathbf{V}_o$ and Training Set $\mathbf{T}_o$

$$S, N_L, N_C \sim U(-0.5, 0.5)$$
$$C := \mathbf{1}_{0.7 \cdot S + 0.3 \cdot N_C \geq 0}$$
$$L := 0.9 \cdot S + 0.1 \cdot N_L$$

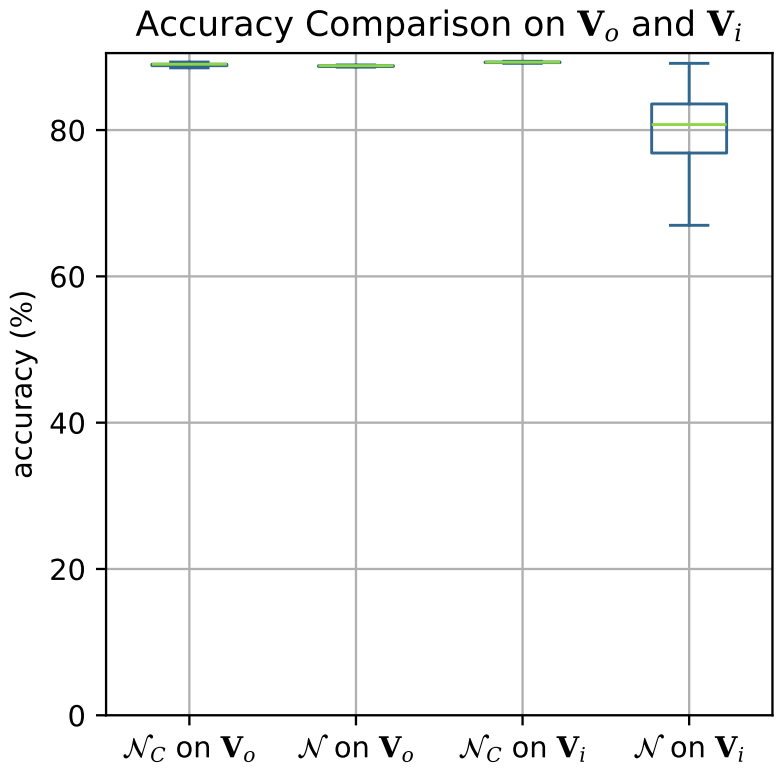
Interventional Validation Set $\mathbf{V}_i$ (Non-IID)

$$L := 0.1 \cdot S + 0.9 \cdot N_L$$

Experimental Setup (Models)

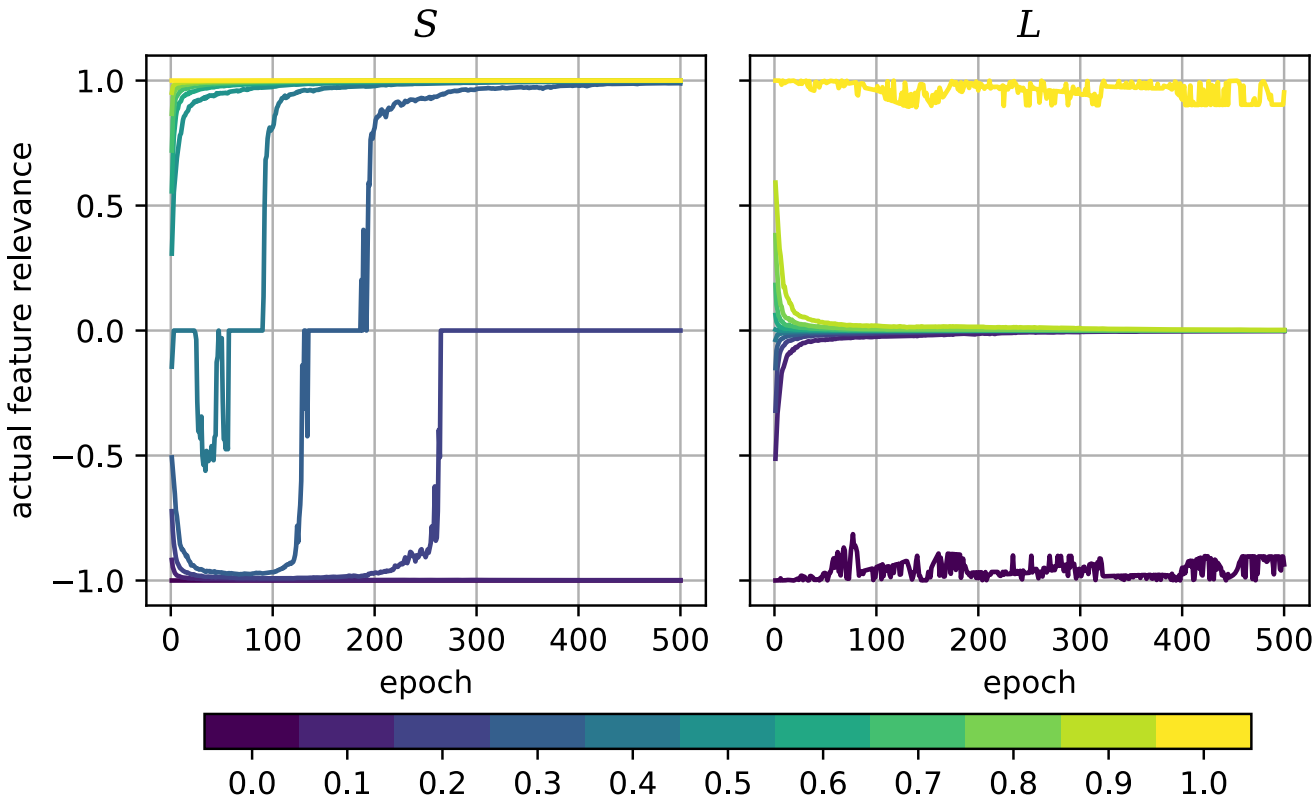
- Basic model with 2, 7, 7 and 2 neurons using linear layers and ReLU.
- Two versions:
  - $\mathcal{N}_C$: Trained with CIML-R extension.
  - $\mathcal{N}$: Trained without CIML-R extension.
- 300 training trials, both models start with the same parameters.
- Hyperparameters:
  - $\lambda := 1.0$.
  - $t := (1, 0)$ for the features (S, L).

Model Accuracy



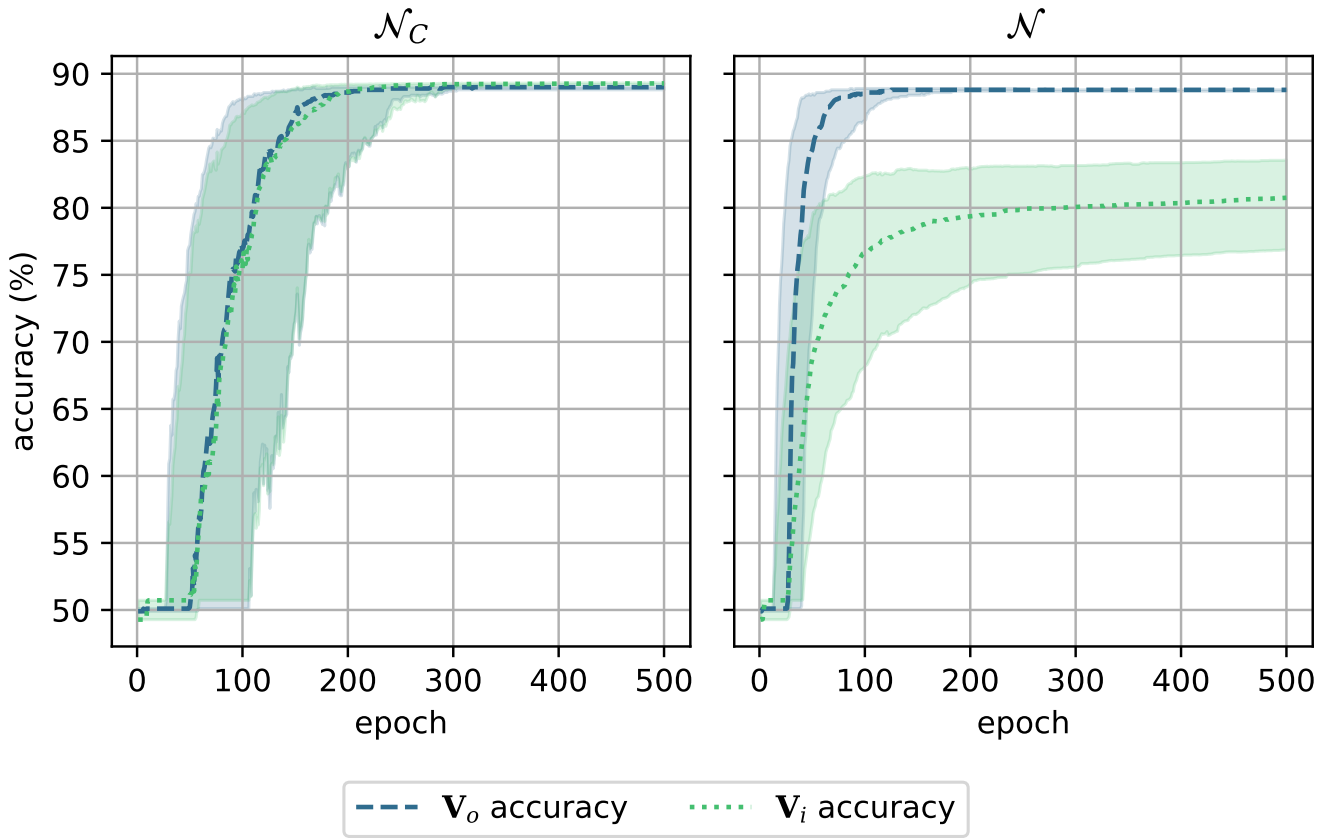
- Comparable performance on $\mathbf{V}_o$ and for $\mathcal{N}_C$ on $\mathbf{V}_i$.
- Performance of $\mathcal{N}$ on $\mathbf{V}_i$ is lower and shows higher variance.

Relevance Evolution



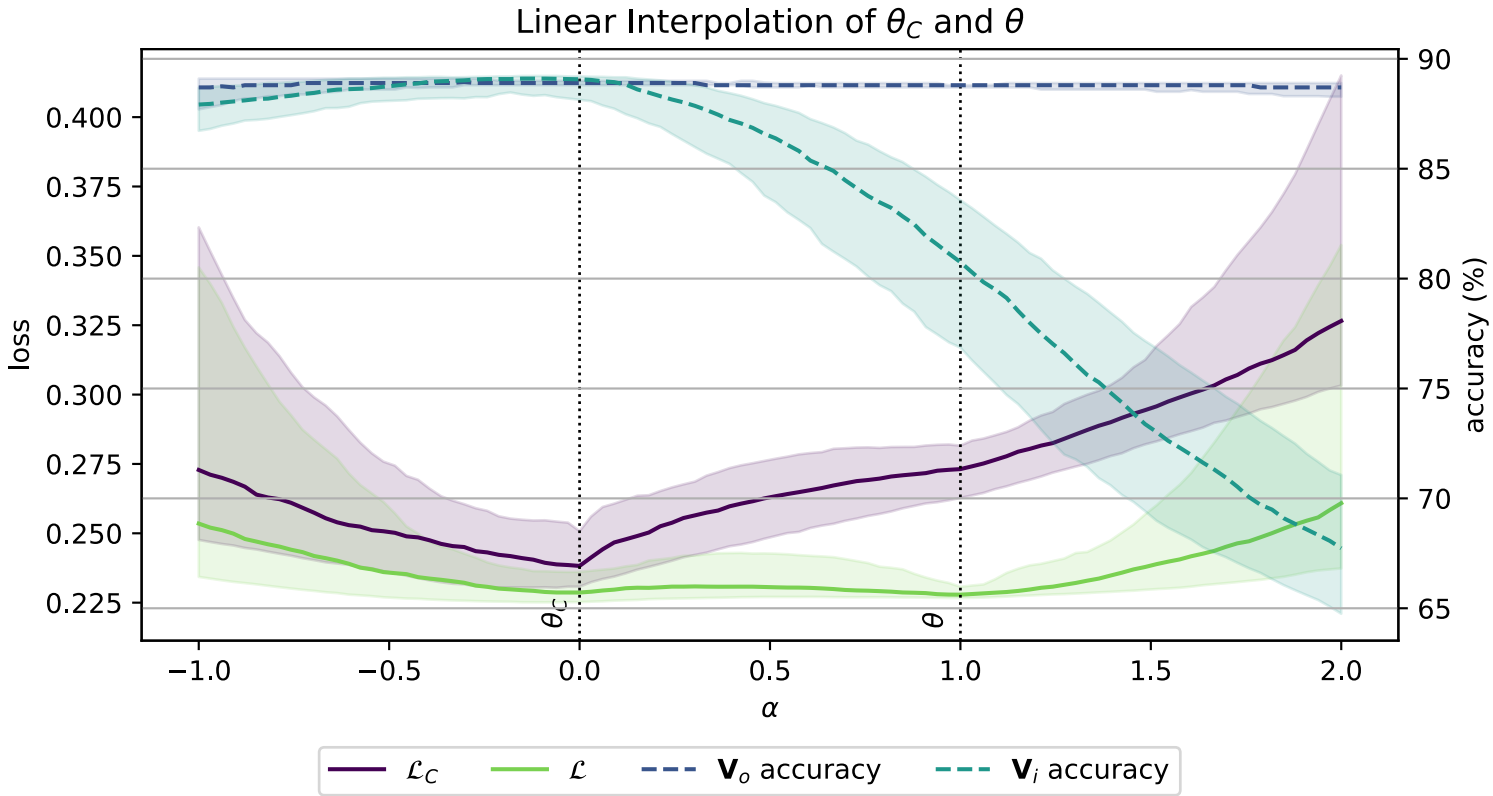
- Plots show distribution of the (actual) feature relevance values R for each epoch.
- During training, distribution shifts toward values of t.

Training Performance (Accumulated)



- $\mathcal{N}_C$: Accuracies converge.
- $\mathcal{N}$: Accuracy for $\mathbf{V}_i$ stagnates at around 80%.

Model Interpolation (Accumulated)



- Interpolation for $\theta_\alpha := \theta_C + \alpha(\theta - \theta_C)$.
- Ascend between θ_C and θ guides the gradient descent.

Conclusion & Future Work

- CIML-R informs the neural network about the causal structure resulting in a successful guidance in our experiments.
- To extend the method to non-tabular datasets, such as image data, semantic segmentation is worth investigating.
- We plan to test CIML-R on other datasets or scenarios involving more complex causal structures.

Thank you.

Contact Details

Martin Surner
Research Associate
Landshut University of Applied Sciences
Martin.Surner@haw-landshut.de

References


Montavon, Grégoire et al. (2019). "Layer-Wise Relevance Propagation: An Overview". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Lecture Notes in Computer Science, pp. 193–209. DOI: 10.1007/978-3-030-28954-6_10.


Pearl, Judea and Dana Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. 1st. USA: Basic Books, Inc.