Improving the Explainability of Graph Neural Networks for Power Grid Topology Error Identification

Master‘s Thesis Presentation

Cora Hartmann
Albert-Ludwigs-University Freiburg
Chair for Algorithms and Data Structures

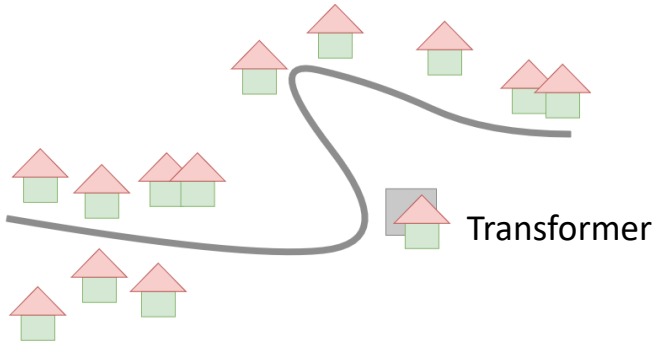
Examiner: Prof. Dr. Hannah Bast, Prof. Dr. Gunther Gust

Advisers: Bodo Rückauer, Sebastian Walter

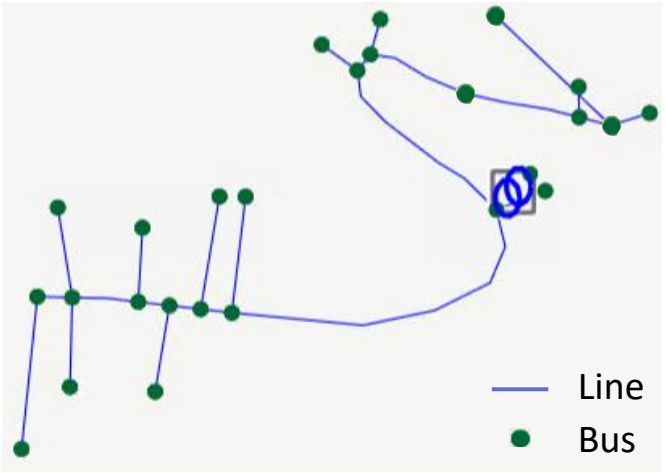
March 21, 2025

Example

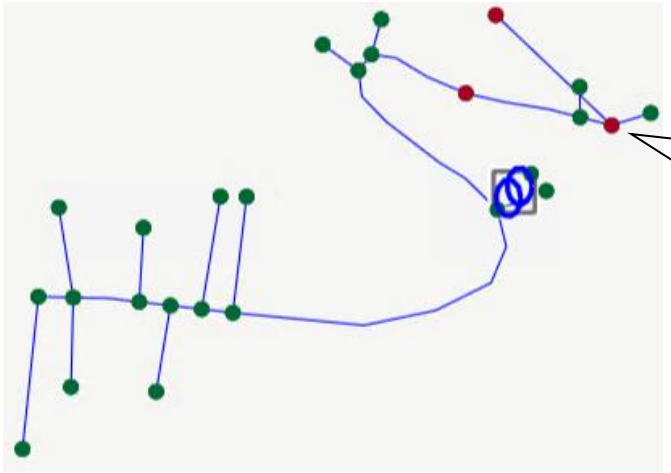
Erroneous Grid Topology



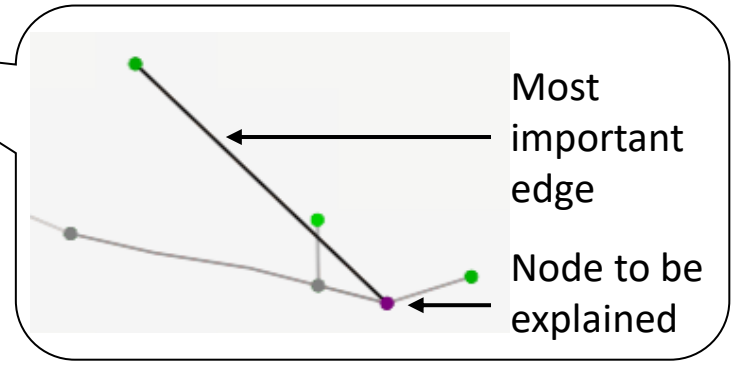
Residential area



Grid topology



Error identification with GNN

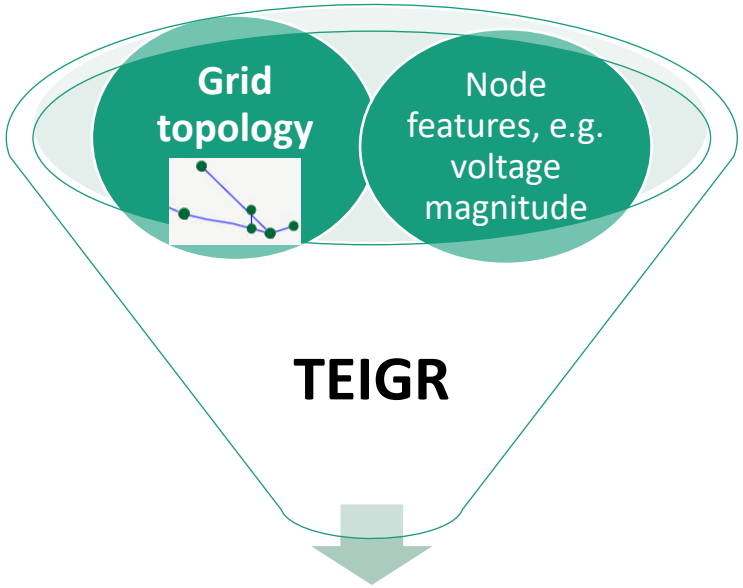


Explanation subgrah

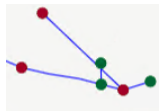
Fraunhofer ISE

Problem

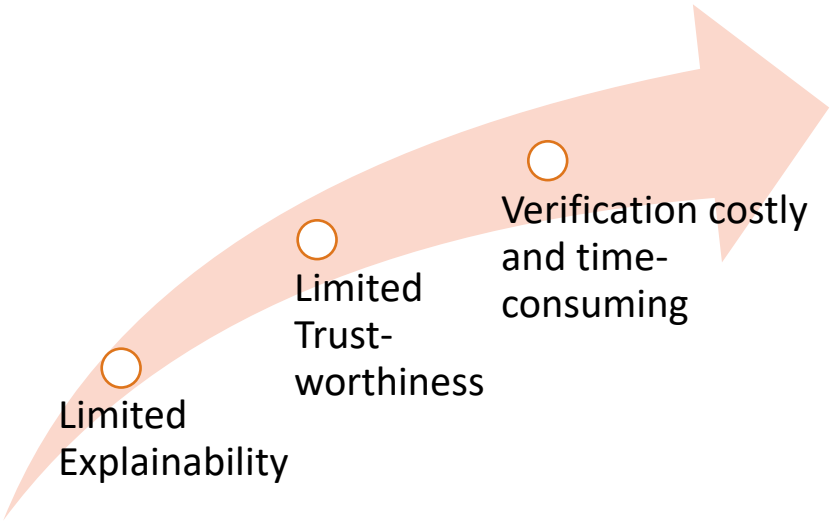
Previous Work



Error Identification:
Classification of nodes as „error“ or „no error“



Problem



Solution Techniques

- Confidence Calibration
- Model Visualization
- Explainable AI

Fraunhofer ISE

Solution

Solution Techniques

- Confidence Calibration
- Model Visualization
- Explainable AI

Main Contributions

We find that TEIGR is well calibrated and only slightly under-confident.

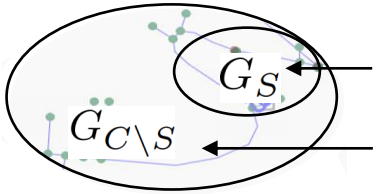
We improve the calibration.

The visualizations reveal clusters based on topographical proximity and the node labels.

We show that model training improves the representation.

Content of this presentation

Evaluation Metrics



Explanation Subgraph

Complement of Subgraph

User Requirement	Metric	Formula
Sufficiency	Fidelity -	$\text{fid}_- = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i^{G_S} = \hat{y}_i)$

N Number of nodes

$\hat{y}_i$ Prediction for node i

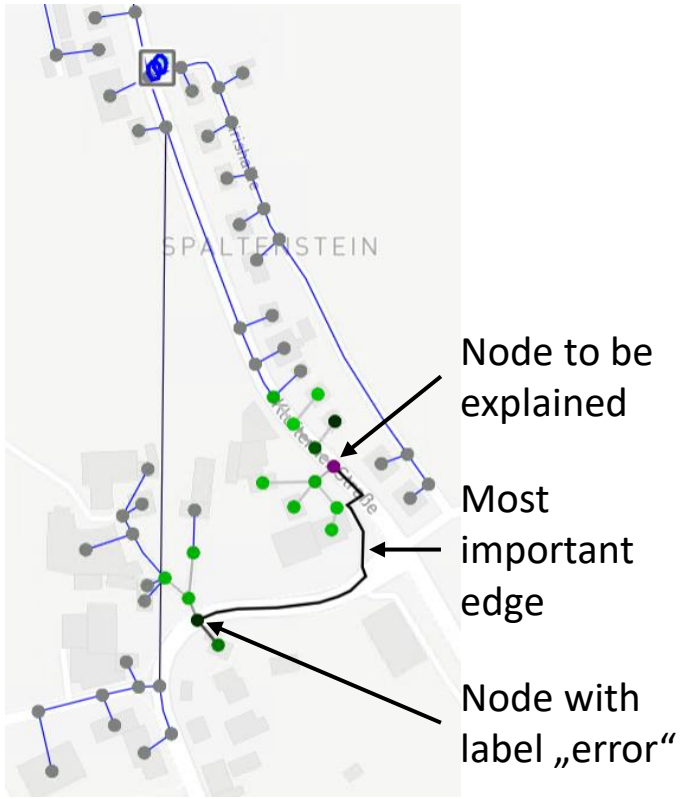
$\mathbb{1}$ Indicator function

G_S Explanation Subgraph

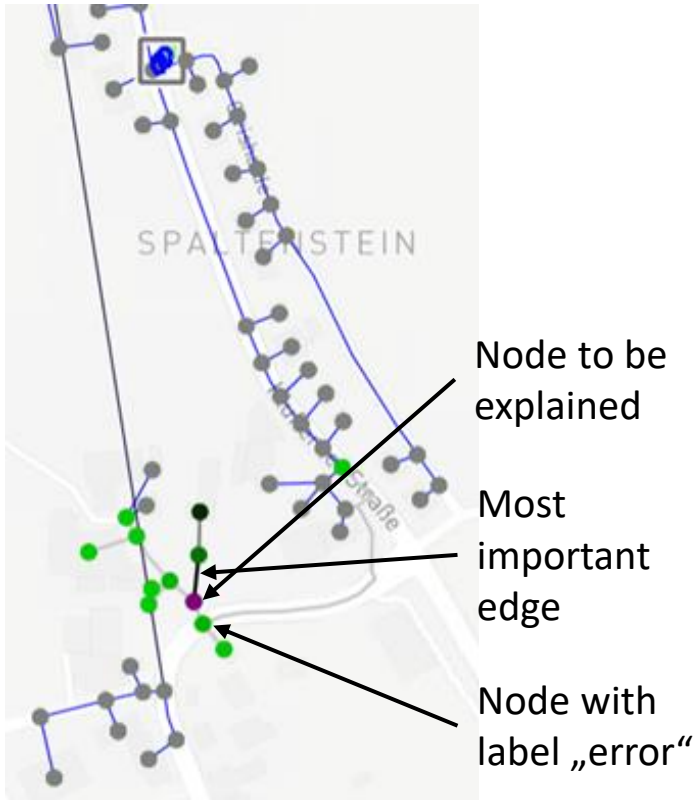
Fraunhofer ISE

Evaluation: Qualitative Analysis

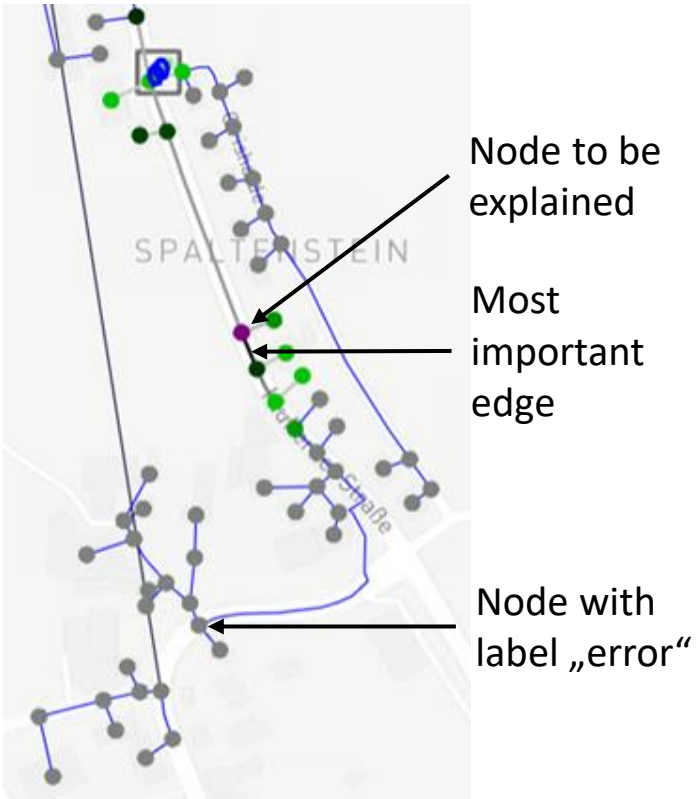
Explanation Categories for False Positives (FPs)



(a) Neighbor node is real error, and explanation reflects this.



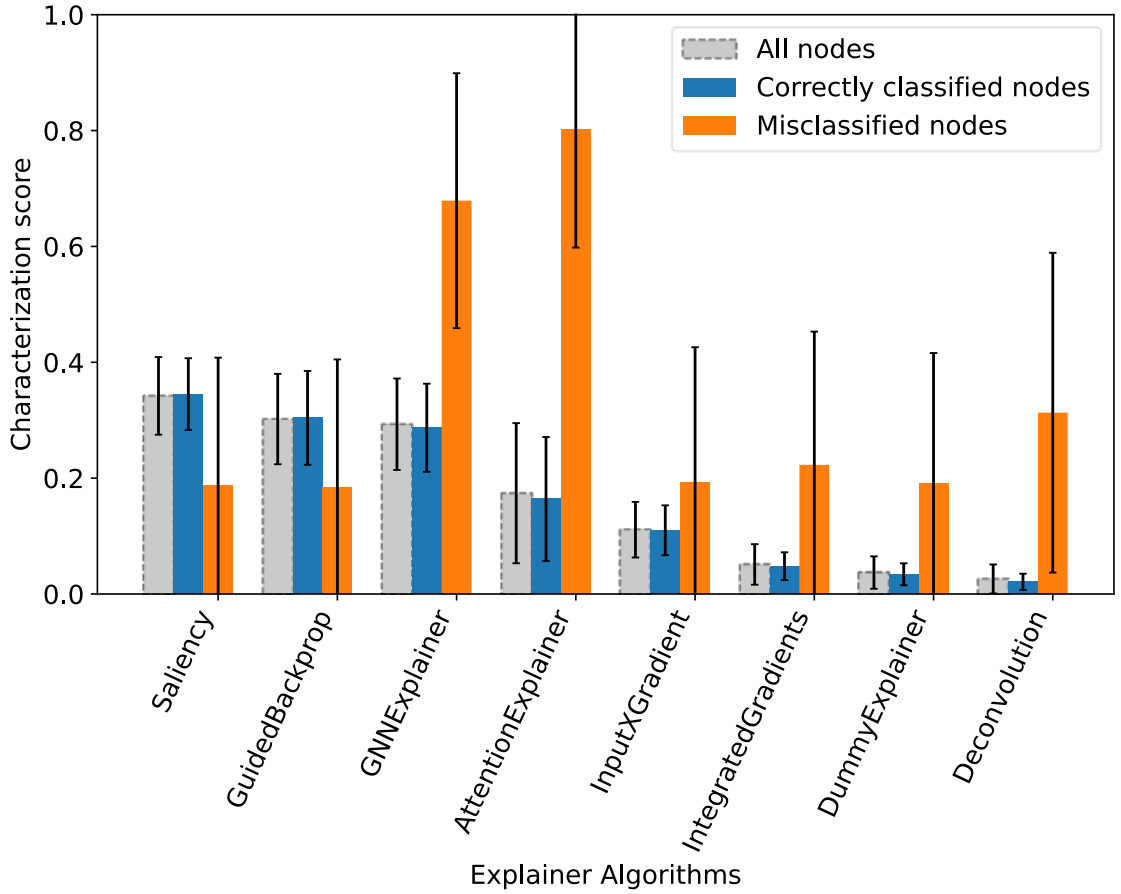
(b) Neighbor is real error, but explanation doesn‘t reflect this.



(c) No real error close by.

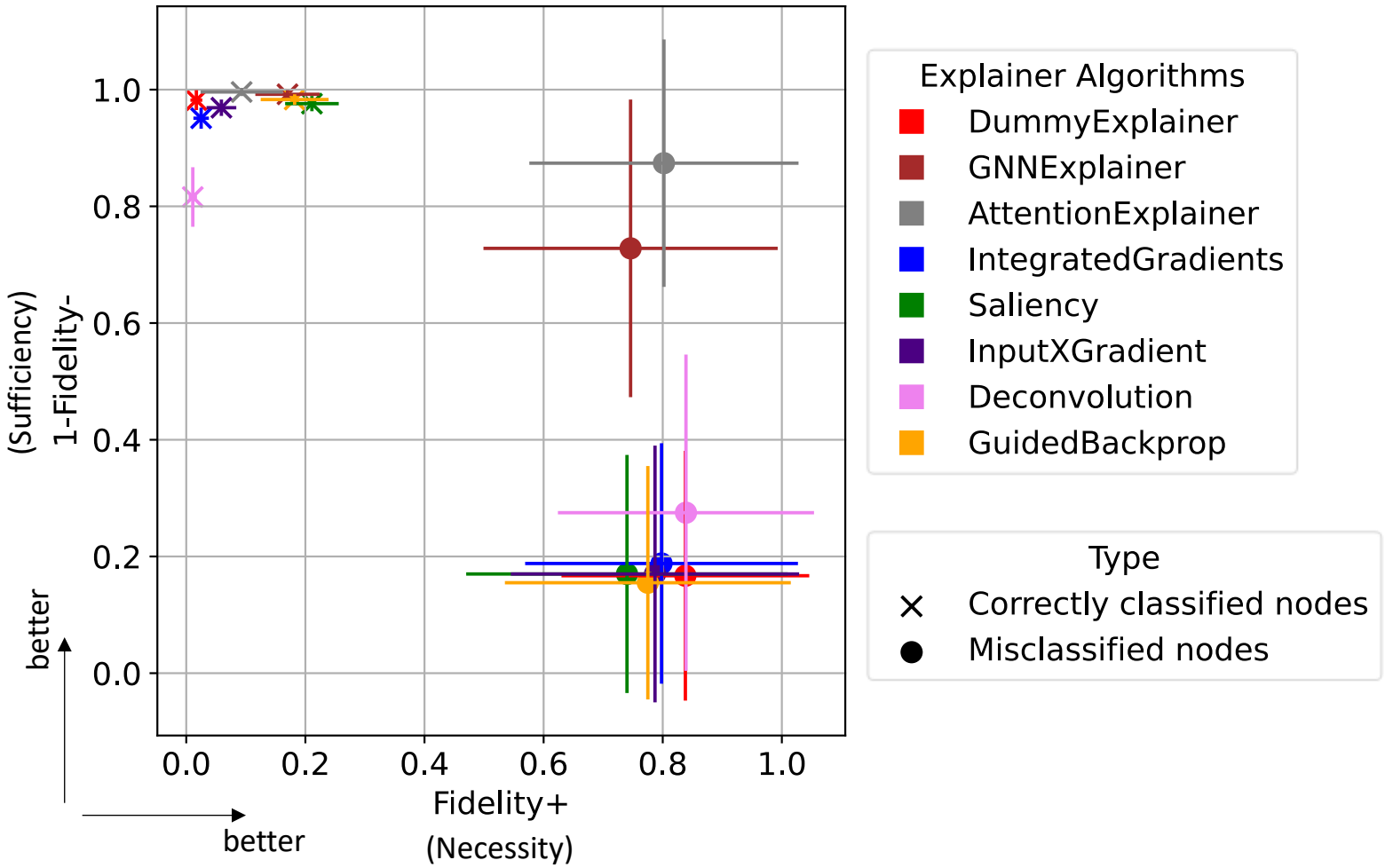
Fraunhofer ISE

Evaluation: Characterization Score Analysis



Fraunhofer ISE

Evaluation: Fidelity Analysis



Fraunhofer ISE

Main Contributions

Recap and Conclusion

Solution Techniques	Main Contributions
Confidence Calibration	We find that TEIGR is well calibrated and only slightly under-confident. We improve the calibration.
Model Visualization	We visualize the internal representations of TEIGR with dimension reduction methods. We show that model training improves the representation.
Explainable AI	We categorize the explanations for incorrectly classified nodes. We analyze the explainability of correctly classified vs. misclassified nodes. We enhance the loss function which improves the explainability.

Thank you for your attention!

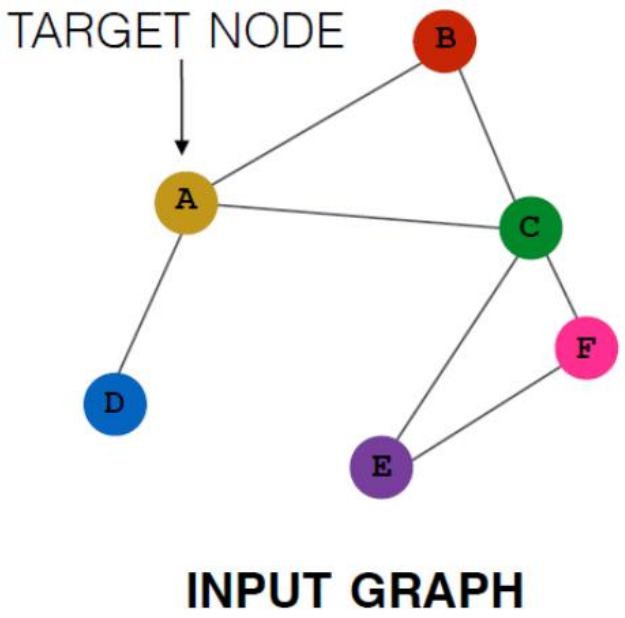
Fraunhofer ISE

Solution Techniques

Solution Techniques	Research Questions
Confidence Calibration	R1.1 How well is TEIGR calibrated? R1.2 Over or under-confidence? R1.3 Improvement with calibration methods?
Model Visualization	R2.1 Do the visualizations show clusters? R2.2 Influence of model features on visualization? R2.3 Do the clusters become more differentiated through model training?
Explainable AI	R3.1 Division of explanations into categories for FPs and FNs? R3.2 Difference of correctly vs. wrongly classified node explanations? R3.3 Improvement by adding an explainability term to the loss function?

Graph Neural Networks (GNNs)

Message Passing



Features

Description of GNN Features used for Topology Error Identification

Feature	Unit	Description
P	W (Watt)	Measured active power
Q	Var (Volt-ampere reactive)	Measured reactive power
V_{mag}	V (Volt)	Measured voltage magnitude
V_{ang}	° (Degrees)	Measured voltage angle
$\hat{V}_{\text{mag}}$	V (Volt)	Expected voltage magnitude
$\hat{V}_{\text{ang}}$	° (Degrees)	Expected voltage angle
$V_{\text{mag\_diff}}$	V (Volt)	$= \hat{V}_{\text{mag}} - V_{\text{mag}}$: Voltage magnitude difference
$V_{\text{ang\_diff}}$	° (Degrees)	$= \hat{V}_{\text{ang}} - V_{\text{ang}}$: Voltage angle difference

Grids

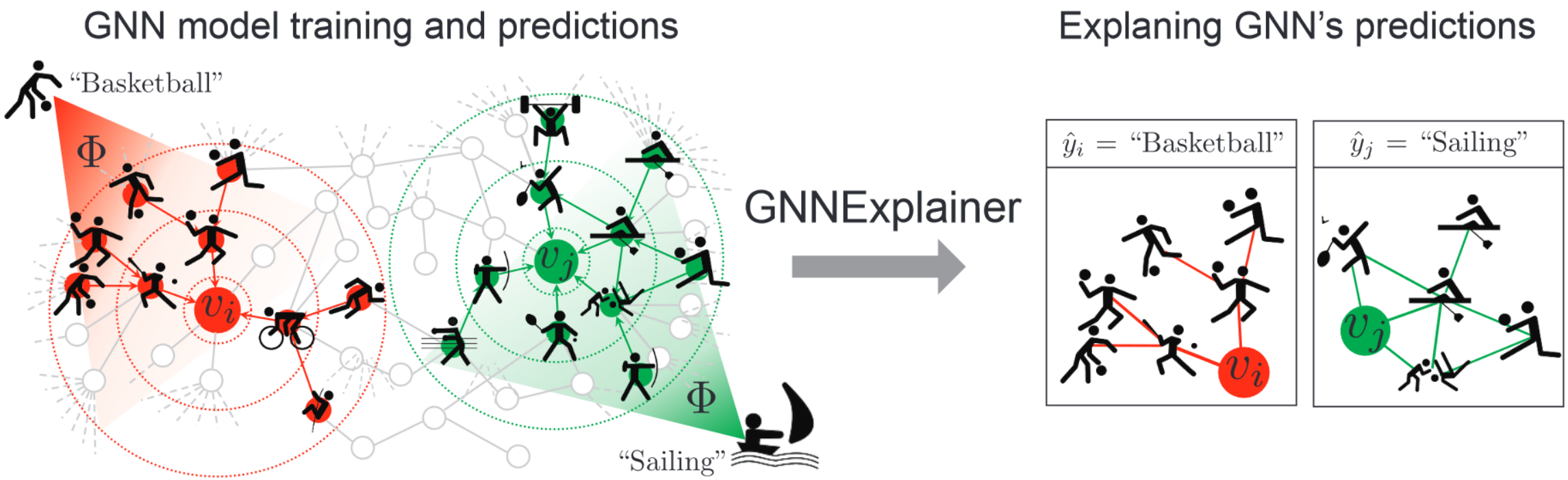
Grid Topologies and Their Properties

Grid	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	μ_{deg}	dia	μ_{sp}
Minimal (syn)	26	48	1.92	13	5.23
Spaltenstein (syn)	80	156	1.97	24	10.23
Eggenweiler (syn)	115	226	1.98	37	13.98
Oberraderach (syn)	282	560	1.99	43	17.65
Manzell Nord (syn)	349	694	1.99	45	18.30
E301 (real)	105	210	2.02	23	8.92
E212 (real)	266	538	2.03	36	15.44
E208 (real)	188	385	2.06	28	12.44

Hyperparameter Configuration

Hyperparameter	Short Description	Value
K	Number of update layers	3
H_{number}	Number of attention heads	8
H_{width}	Width of the attention heads	16
δ_{dropout}	Dropout rate	0.0118
$\sigma(\cdot)$	Non-linear activation function	ReLU
B	Batch size	200
η	Learning rate	0.0026
$opt(\cdot)$	Optimizer	RMSProp

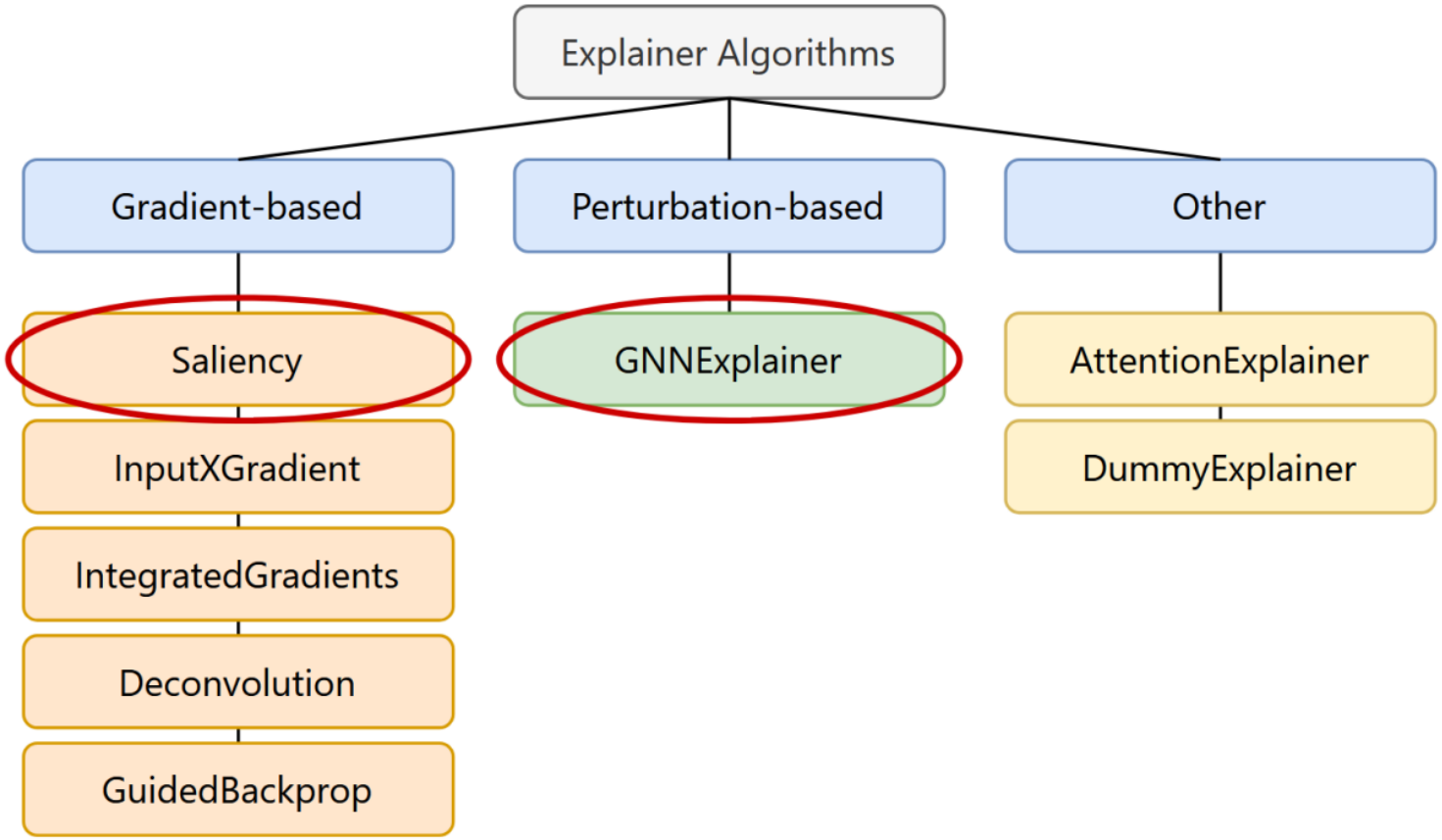
GNN Explainability



Fraunhofer ISE

Explainer Algorithms

Overview of the used explainer algorithms, categorized into gradient- and perturbation-based



Confidence Calibration

Definition Calibration:

$$\mathbb{P}(\hat{y}_u = y_u | \hat{p}_u = p) = p, \forall p \in [0, 1]$$

Reliability Curve:

$$\mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{u \in B_m} \mathbb{1}[\hat{y}_u = y_u]$$

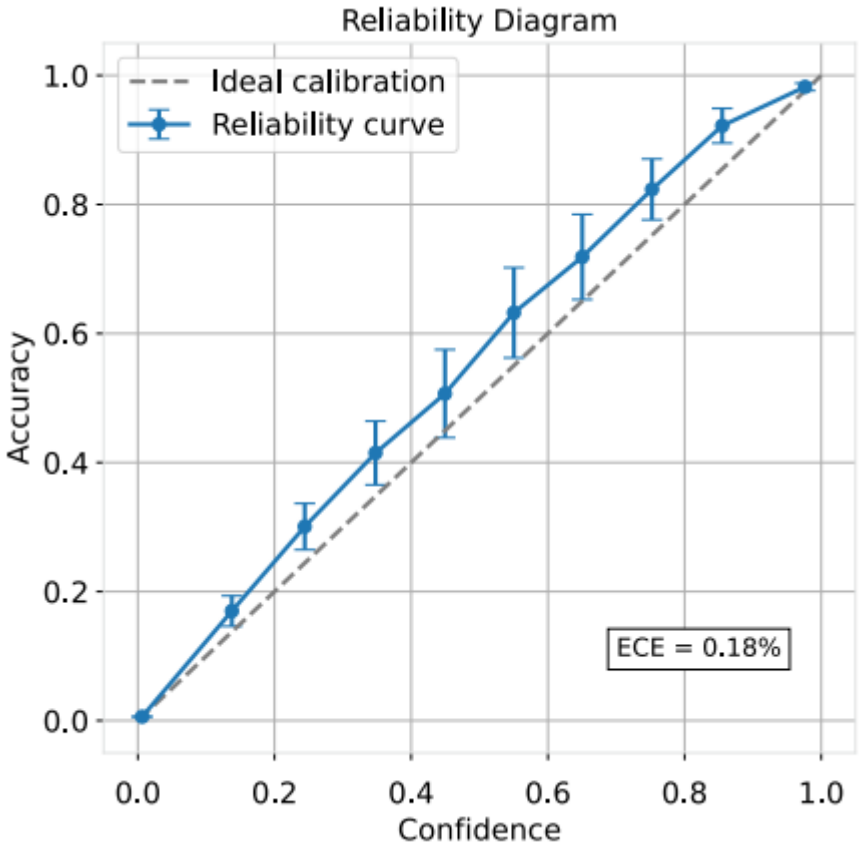
$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{u \in B_m} \hat{p}_u$$

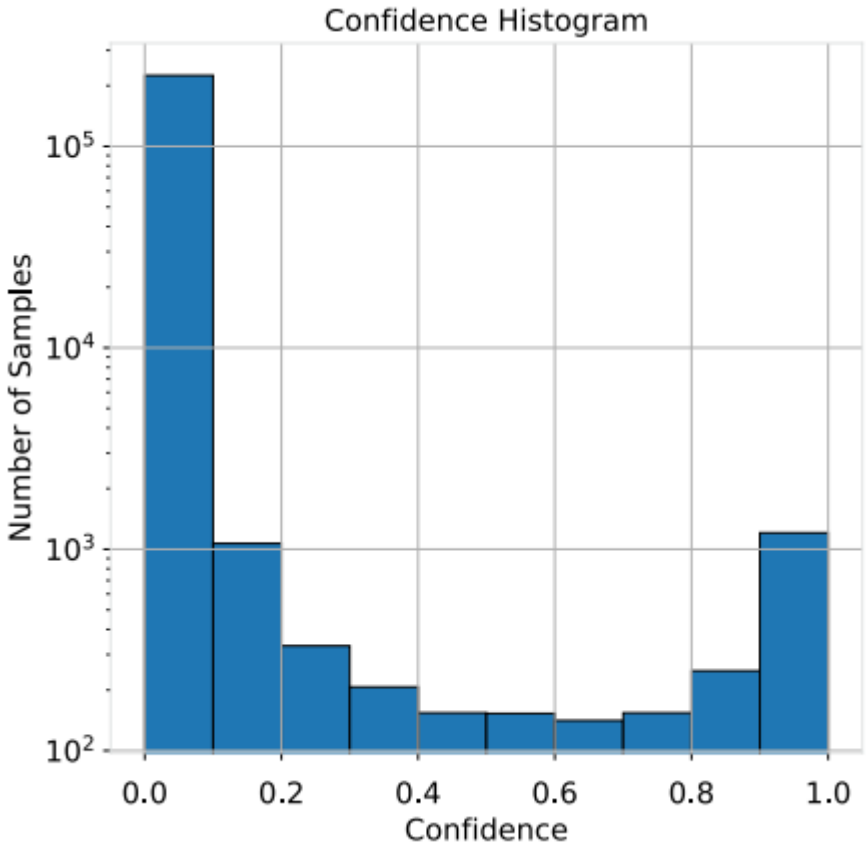
Expected Calibration Error (ECE):

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|$$

Confidence Calibration

Reliability diagram (left) and corresponding confidence histogram (right) for uncalibrated GNN



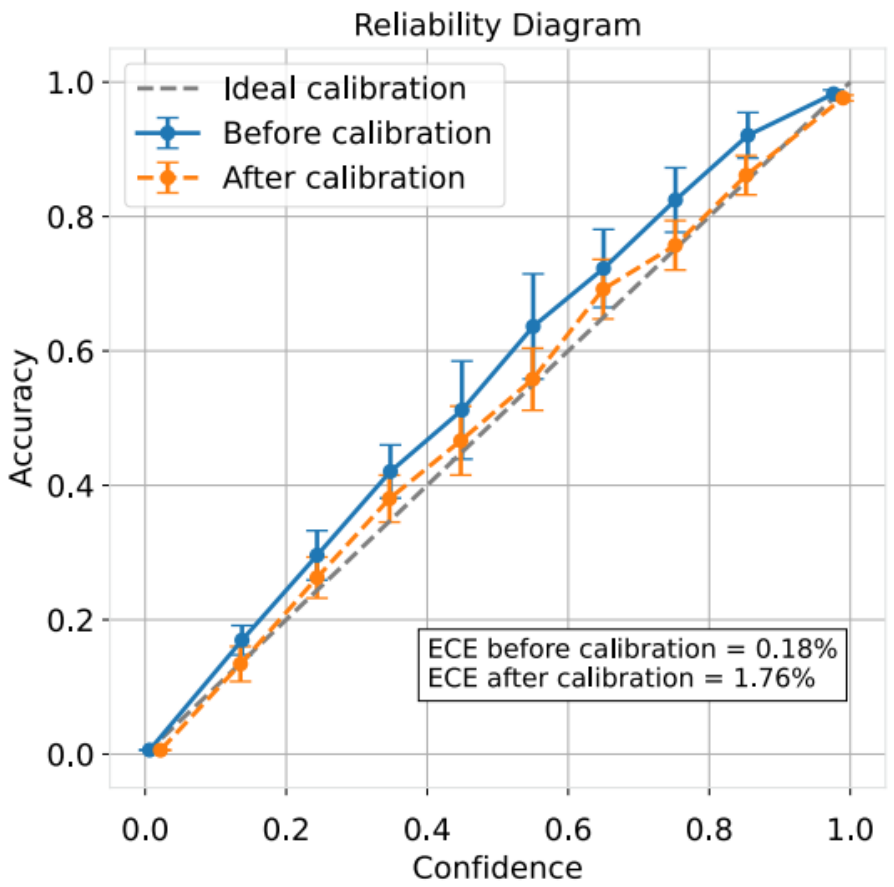


Fraunhofer ISE

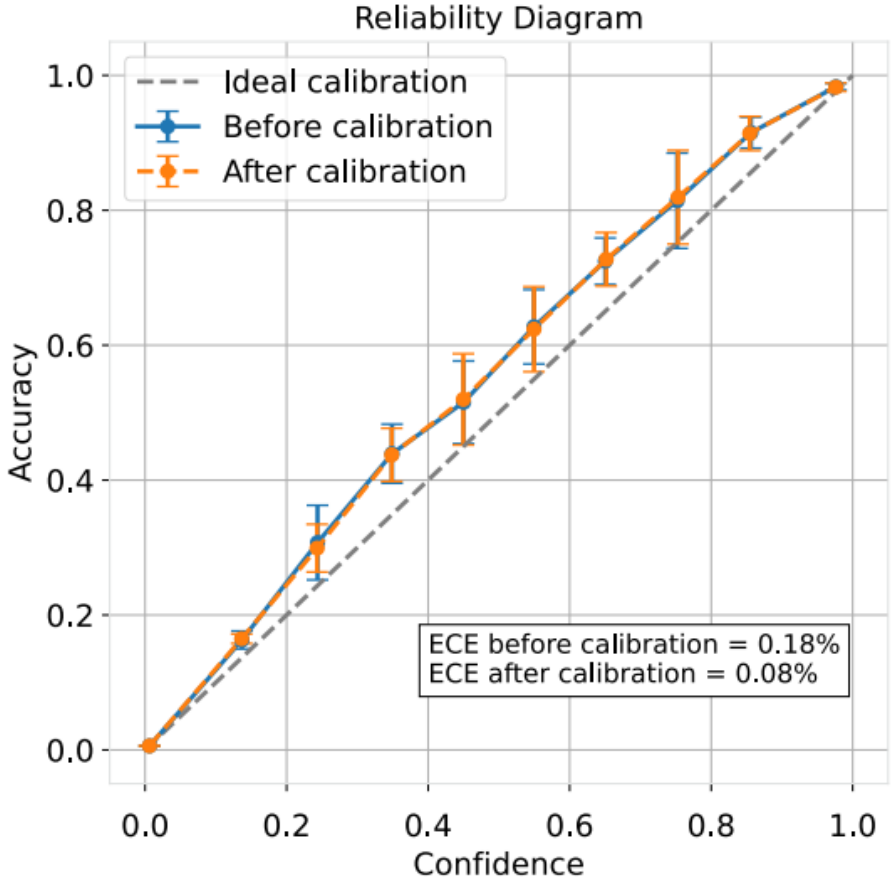
Confidence Calibration

Calibration Methods: Histogram Binning and Temperature Scaling

Histogram Binning



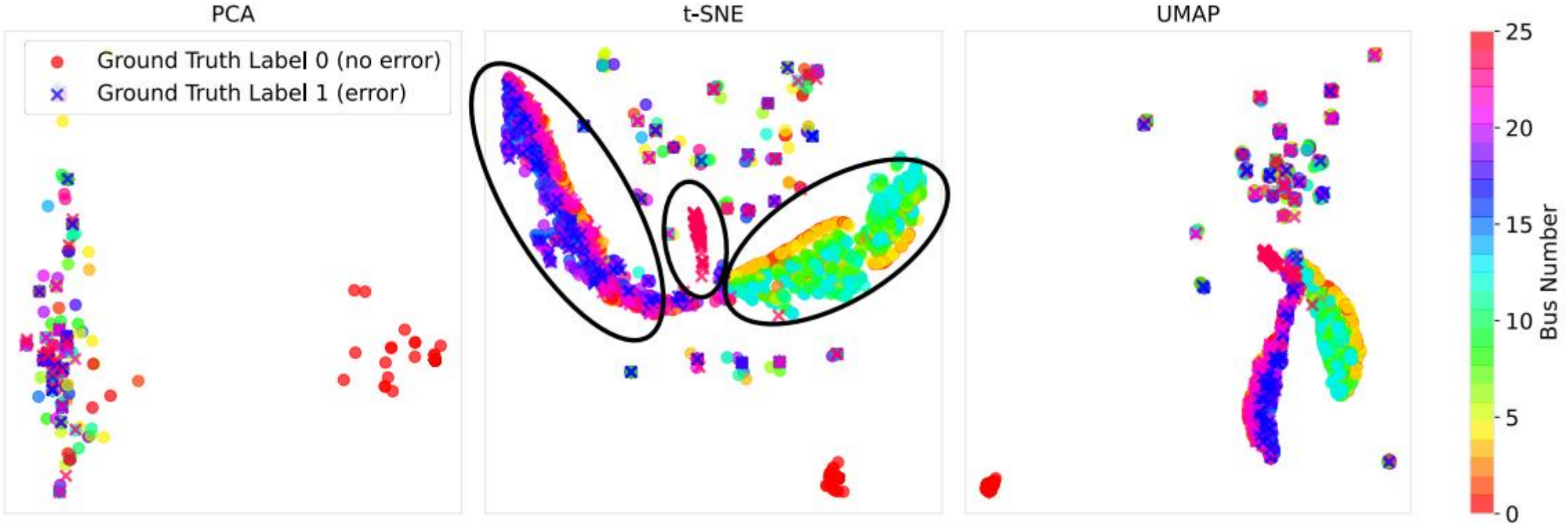
Temperature Scaling



Fraunhofer ISE

Model Visualization

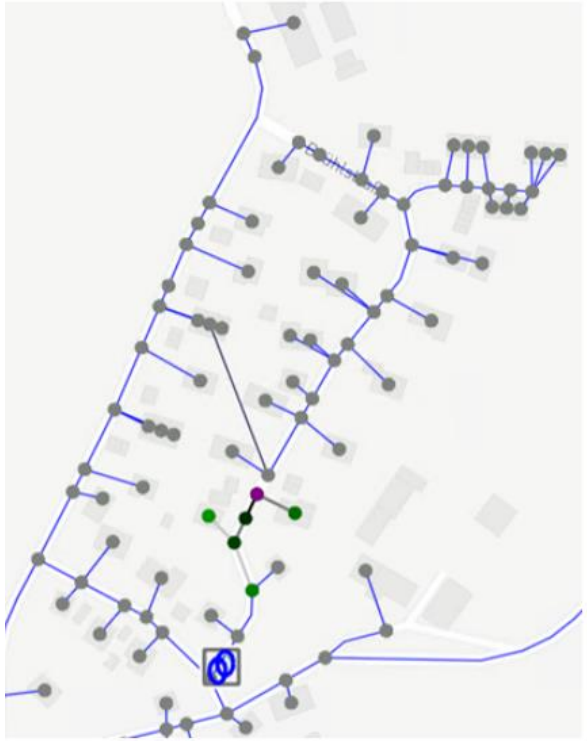
PCA, t-SNE and UMAP visualizations of the trained TEIGR



Fraunhofer ISE

Future Work

- Addressing “Missing Connection” Errors
- Sensitivity Analysis to Enhance Explainability
- Enhance Loss Function by Fine-tuning Pre-trained Model
- Alternative Approaches to Enhancing the Loss Function
- Adapting TEIGRs Usage Based on High Characterization Scores
- Advanced Confidence Calibration Methods



FN-Category 2: Explanation not helpful, because the error type is “missing connection”.