ACML 2023

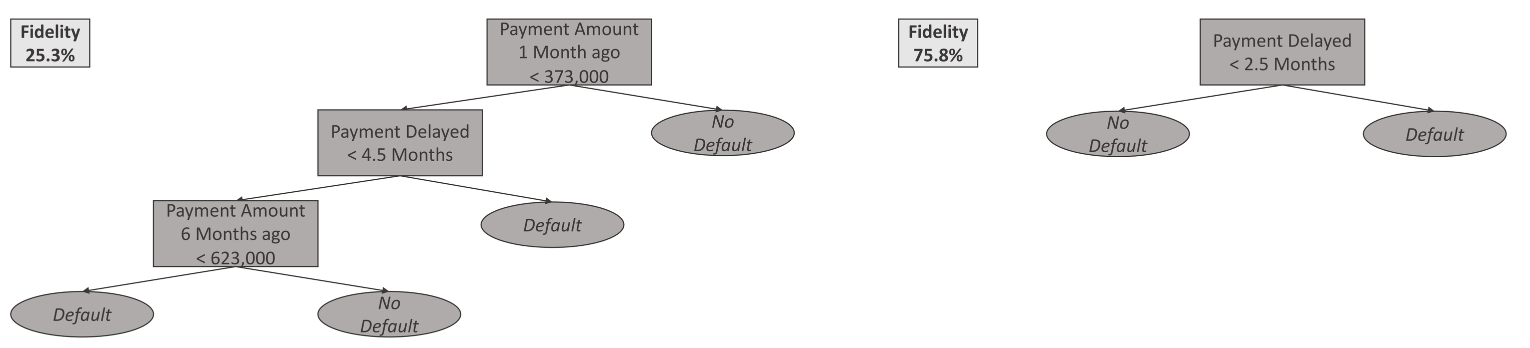
Explaining Neural Networks without Access to Training Data

Motivation

Neural Networks achieve impressive results in a variety of tasks ⟷ Humans can't understand what was actually learned by the model

- Possible solution: Learn a surrogate model that has a high fidelity to the neural network
  – Requires access to the training data to achieve a high fidelity
- Frequently training data is not available e.g., due to privacy concerns

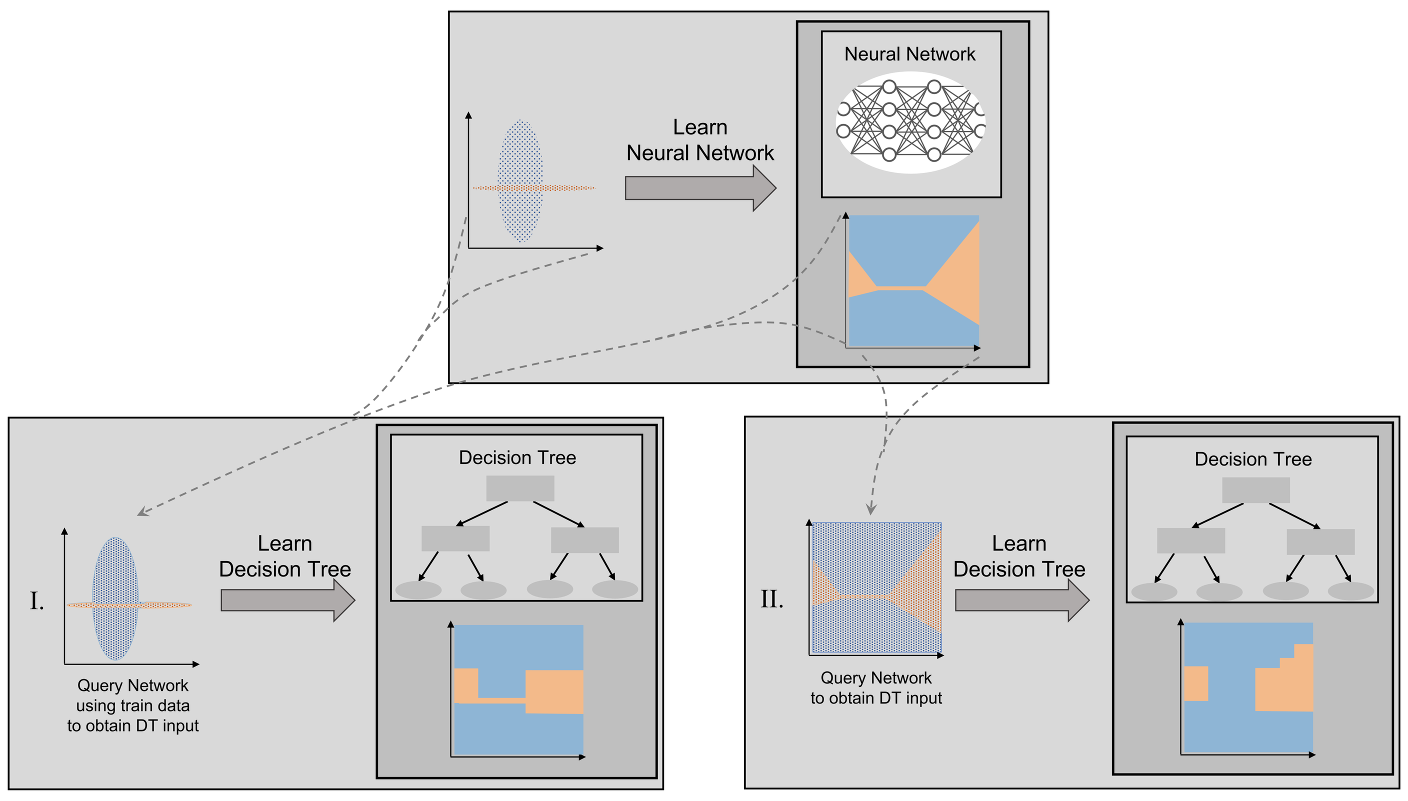
Example: Credit Card Default Prediction with Neural Networks



(a) Sample-Based Distillation (b) I-Net Distillation

- Sample-based approach to learn a surrogate model can't generate reasonable explanations without training data

Related Work: Sample-Based Distillation

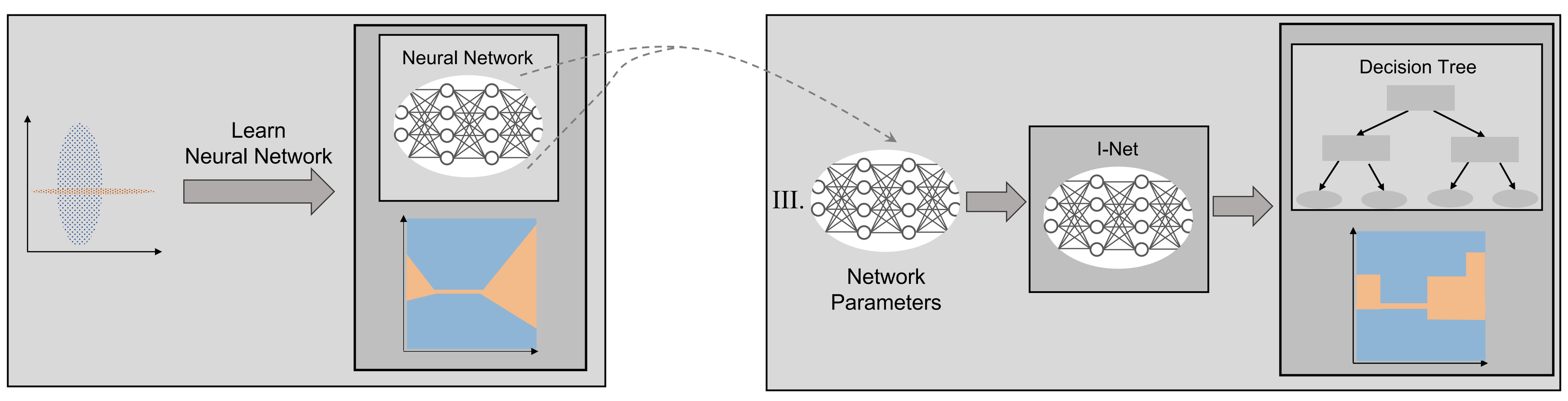


General Procedure:

1. Select an input data set $X = \{\boldsymbol{x}^{(j)}\}_j^M$
   - Usually training data used (I)
   - Alternative: Randomly sample data points (II)
2. Query neural network using X to generate labels $\boldsymbol{y} = \{y^{(j)}\}_j^M$
3. Train surrogate model (e.g. decision tree) on $\{\boldsymbol{x}^{(j)}, y^{(j)}\}_j^M$

Data used for querying the model is very important
➔Information that is not explicitly queried cannot be contained in the explanation!

Interpretation-Networks as Sample-Free Approach



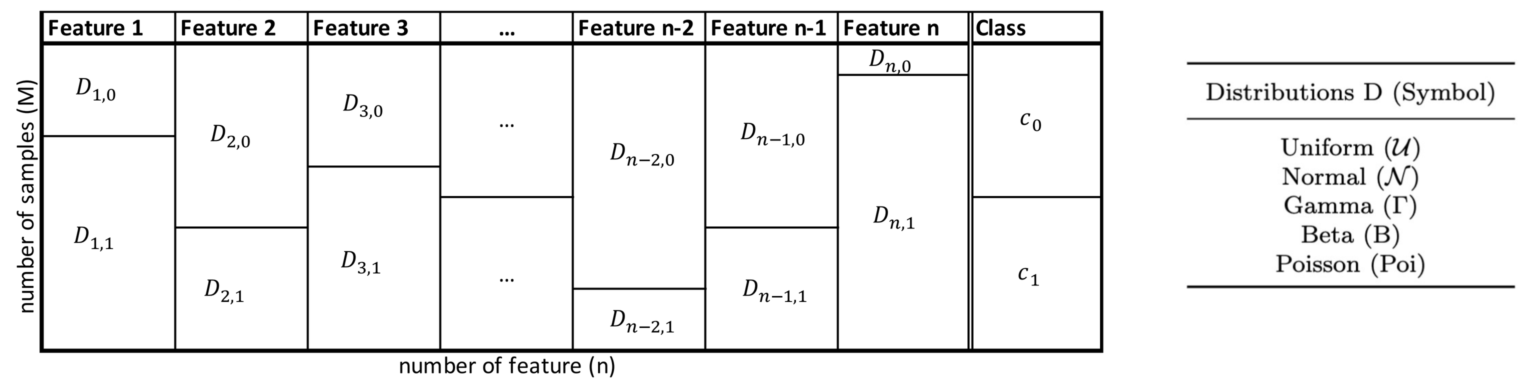
- I-Nets as sample-free approach to generate global surrogate models
- General Procedure:
  1. Train a set of neural networks on synthetic data and extract their learned parameters
  2. Train a second neural network using the extracted parameters as input

No samples are required when generating explanations using the I-Net
➔I-Nets utilizes the network parameters that implicitly contain all relevant information

Methodology

Synthetic Data Generation

- Generating realistic data for training the set of neural networks is crucial
  → Allows generalization to real-world application
- Consider different, diverse distributions that are reasonable for numerous real-world phenomena

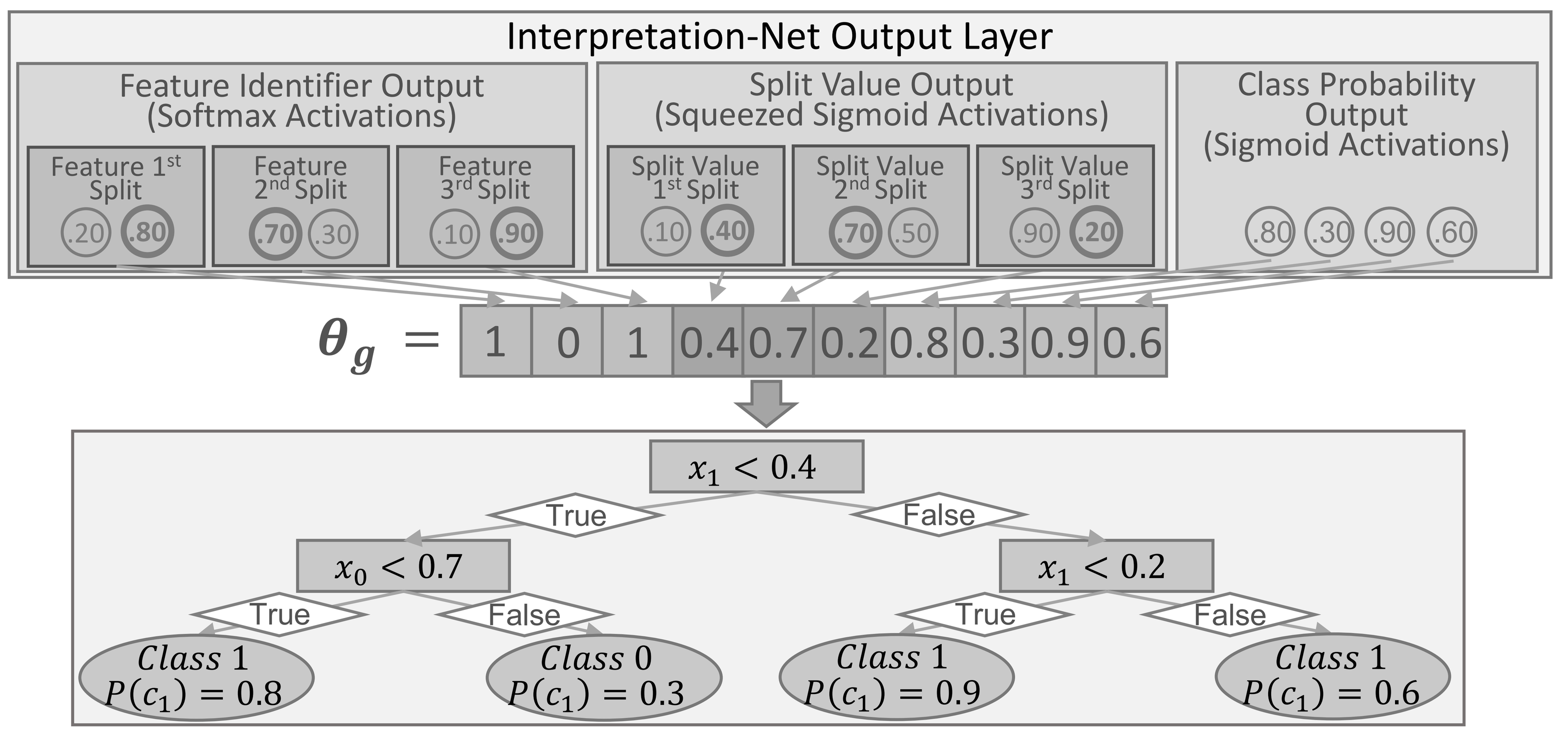


What is the advantage of I-Nets?
➔Training can be performed on synthetic data
➔During the training, we can access the training data of the model

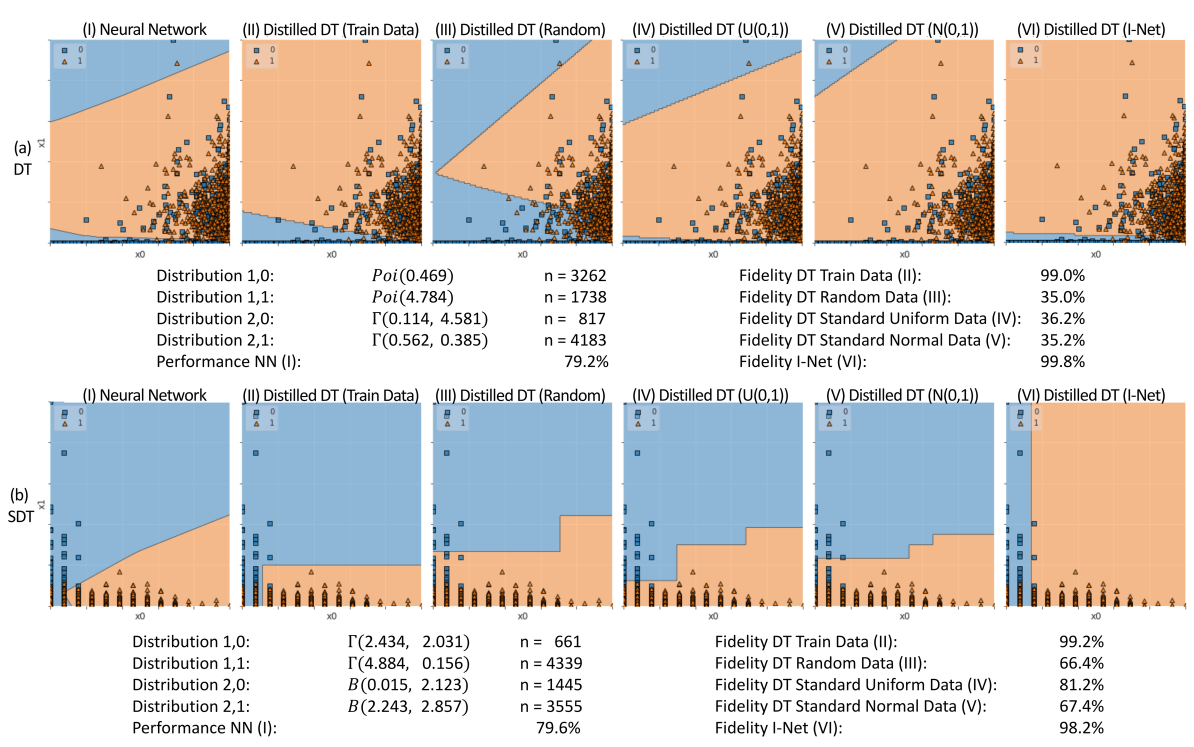
I-Net Output Representation

Three separate types of output layers:

1. Feature Identifier Output
   - One softmax layer for each internal node
   → "Classification task" at each layer
2. Split Value Output
   - One neuron with sigmoid for each internal node
   → sigmoid activation can be used as variable values are in [0, 1]
3. Class Probability Output
   - One neuron with sigmoid for each leaf node (for binary case)
   - One softmax layer for each leaf node (for multi-class case)



Results: Visual Comparison of Decision Boundaries



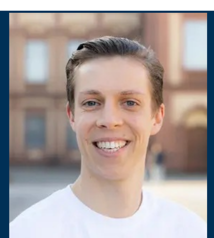
Without access to training data, surrogate models learned using sample-based approaches neglect relevant parts and focus on explaining irrelevant aspects

Results: Performance Comparison

Dataset	$\mathcal{I}$-Net	Multi-Distribution	Standard Uniform	Standard Normal
Titanic (n=9)	**95.51 ± 0.00**	71.12 ± 17.16	86.07 ± 3.30	86.29 ± 7.75
Medical Insurance (n=9)	82.71 ± 0.00	88.12 ± 6.71	89.47 ± 4.19	**90.75 ± 8.83**
Breast Cancer Wisconsin Original (n=9)	**97.10 ± 0.00**	83.62 ± 13.09	39.42 ± 13.90	31.88 ± 0.00
Wisconsin Diagnostic Breast Cancer (n=10)	**80.36 ± 0.00**	56.43 ± 17.65	37.86 ± 15.56	33.39 ± 5.42
Heart Disease (n=13)	73.33 ± 0.00	74.67 ± 9.45	**85.67 ± 5.97**	80.33 ± 7.67
Cervical Cancer (n=15)	**84.71 ± 0.00**	65.41 ± 27.77	71.88 ± 9.64	60.82 ± 30.29
Loan House (n=16)	**100.00 ± 0.00**	77.05 ± 24.41	**96.89 ± 7.42**	59.84 ± 33.84
Credit Card Default (n=23)	**75.80 ± 0.00**	69.16 ± 17.58	74.76 ± 0.05	34.33 ± 20.31
Mean Fidelity	**86.19**	73.20	72.75	59.70

The I-Net consistently outperforms a sample-based distillation if the training data is not accessible.

Sascha Marton
sascha.marton@uni-mannheim.de
University of Mannheim



Jun.-Prof. Dr. Stefan Lüdtke
stefan.luedtke@uni-rostock.de
University of Rostock

Dr. Christian Bartelt
christian.bartelt@uni-mannheim.de
University of Mannheim

Andrej Tschalzev
andrej.tschalzev@uni-mannheim.de
University of Mannheim

Prof. Dr. Heiner Stuckenschmidt
heiner.stuckenschmidt@uni-mannheim.de
University of Mannheim