Explainable Artificial Intelligence (XAI)

- **Goal:** Explain Alzheimer’s diagnosis made by a deep neural network from anatomical shape and tabular biomarkers. Most existing work on XAI focus on CNN, but
  1. Inputs are heterogeneous,
  2. Shapes are non-Euclidean,
  3. Network differs substantially from a standard CNN.
- Axioms of explanations:
  1. Approximate
  2. Scale
  3. Linearity
  4. Independence
  5. Consistency

**Methods**

- **Shapley Value of feature i of input z:**
  \[ \Delta_i = f(x_{S_i}(z); x_{\overline{S}_i}(z)) - f(x_S; x_{\overline{S}}) \]

  \[ s_i(z \mid f) = \frac{1}{|F|} \sum_{S \subseteq F} |S|! \cdot (|F| - |S| - 1)! \cdot \Delta_i \]

  - **Scales exponentially** in the number of features |F|.
  - \( x_{\overline{S}_i} \): Simulate absence of point by replacing it with projection onto the convex hull.
  - Approximate Shapley value:
    1. Represent output of first layer as normal distribution using sampling theory (Ancona et al., 2019; Cochran, 1977).

**Experiment: Real Data from ADNI**

- Use Wide and Deep Network with left hippocampus shape (\( k = 1024 \) points) and 9 tabular biomarkers.
  - 1308 visits for training, 169 for hyper-parameter tuning.
  - Balanced accuracy: 0.942 on test set (176 patients).
- Explanation of diagnosis for individual patient:

  - **Hippocampus**
    - p-tau = 43.51
    - AV45-PET = 0.999
    - APOE4 = 1
  - **Tabular biomarkers**
    - Education = 20
    - Age = 81.284

  - **Synthetic Data**
    - Use PointNet to classify 100 point clouds of “X” and “T”.
    - Compare against exact Shapley value.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>SRC</th>
<th>NDCG</th>
<th>NE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV Exact</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>65,538</td>
</tr>
<tr>
<td>Occlusion</td>
<td>16.1311</td>
<td>0.3180</td>
<td>0.8659</td>
<td>17</td>
</tr>
<tr>
<td>SV Sampling</td>
<td>0.0008</td>
<td>0.9505</td>
<td>0.9986</td>
<td>32,000</td>
</tr>
<tr>
<td>SV Sampling</td>
<td>0.0340</td>
<td>0.5440</td>
<td>0.9481</td>
<td>512</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.0443</td>
<td>0.6918</td>
<td>0.9641</td>
<td>512</td>
</tr>
</tbody>
</table>

MSE: mean squared error. SRC: Spearman’s rank correlation. NDCG: normalized discounted cumulative gain. NE: forward passes.