LAB MLDS

Project Presentation

Presenters:

Sona Mehdizade

Josephina Thiele

Jakub Eryk Tarka

Free University of Berlin

Summer Semester 2024

Instructor: Prof. Dr. Grégoire Montavon

II. Getting Insights into Quantum-Chemical Relations

Presented by: **Josephina Thiele**

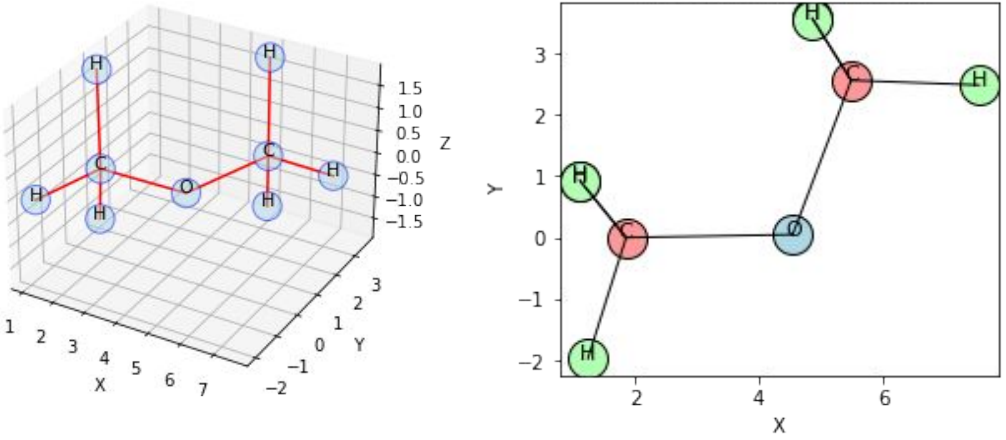
Dataset Overview

Dataset with atom types and coordinates of 7165 molecules consisting of maximum 23 atoms.

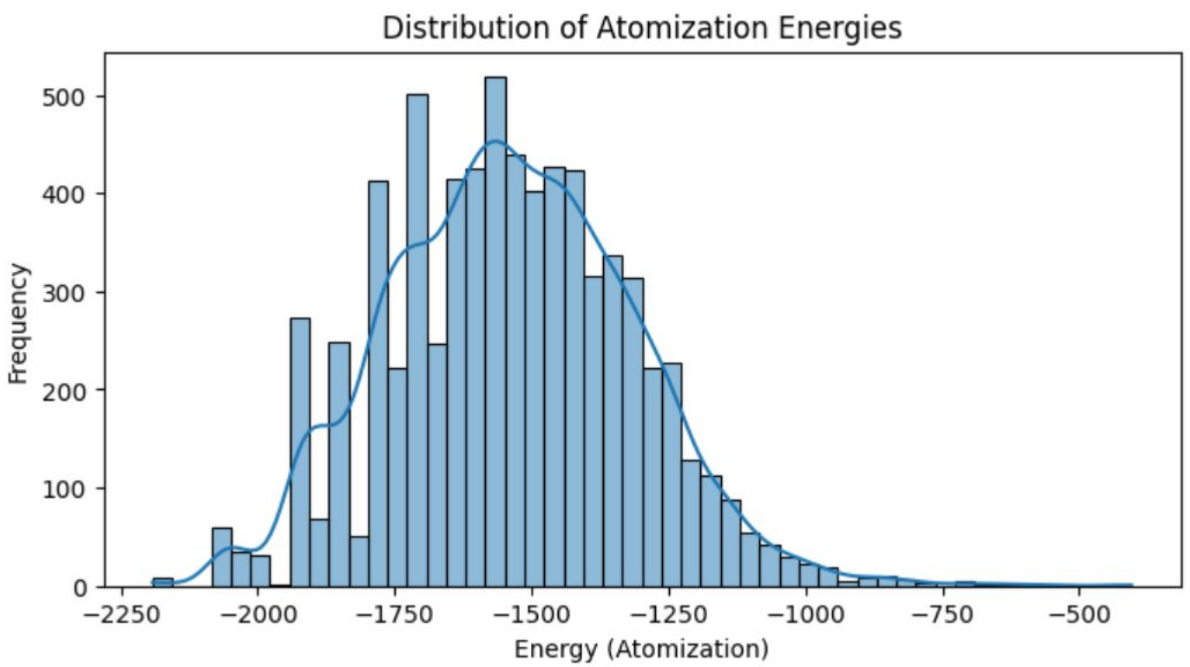
Atom types: Hydrogen (H), Carbon (C), Nitrogen (N), Oxygen (O), Sulfur (S)

Target variable is the atomization energy (AE) of the molecules.

Dimethyl ether

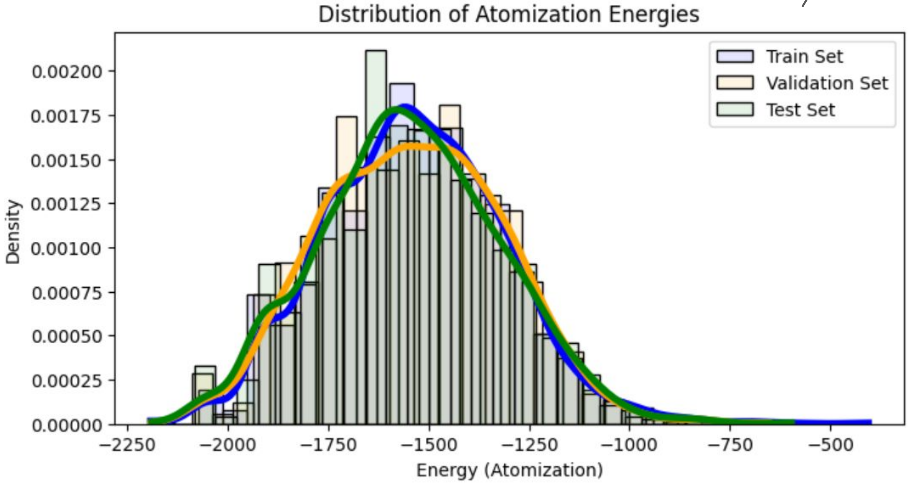


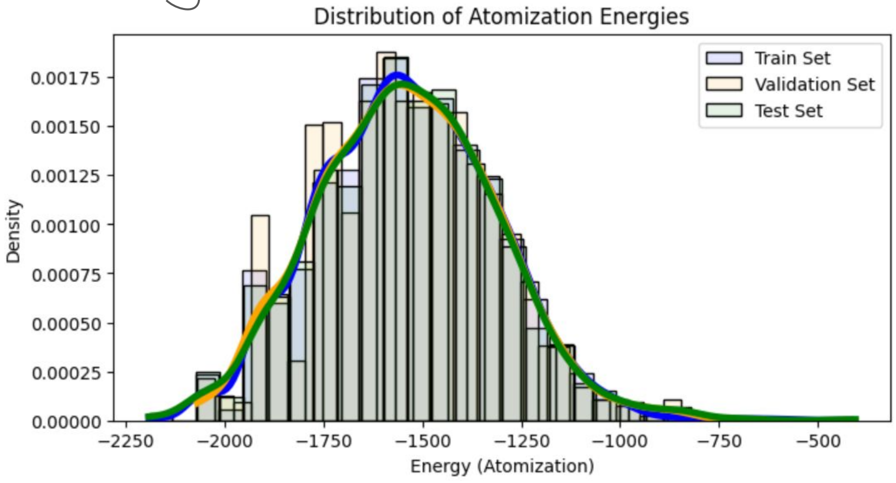
The Target Variable: Atomization Energy



Train-Validation-Test Split

Stratified Splitting





Model 1: Simple Atom based representation of molecules

Feature Map 01

One hot representation

$$\phi(\mathcal{E}_i) = \begin{pmatrix} I(\mathcal{E}_i = \mathrm{H}) \\ I(\mathcal{E}_i = \mathrm{C}) \\ I(\mathcal{E}_i = \mathrm{N}) \\ I(\mathcal{E}_i = \mathrm{O}) \\ I(\mathcal{E}_i = \mathrm{S}) \end{pmatrix} \in \mathbb{R}^5$$

$$\boldsymbol{x} = \sum_{i=1}^{|\mathcal{M}|} \phi(\mathcal{E}_i)$$

Model Training 02

Ridge Regression

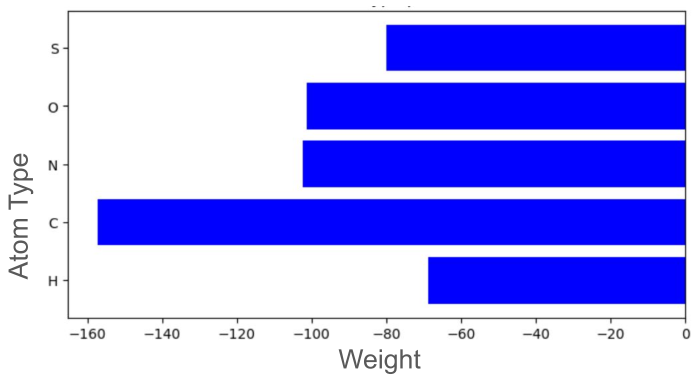
Linear Model with objective:

$$J(\boldsymbol{w}) = \mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{x} - t)^2 + \lambda \|\boldsymbol{w}\|^2]$$

Optimal parameter $\lambda = 10^{-5}$
Mean Absolute Test Error:
15.32 kJ/mol

Model Explanation 03

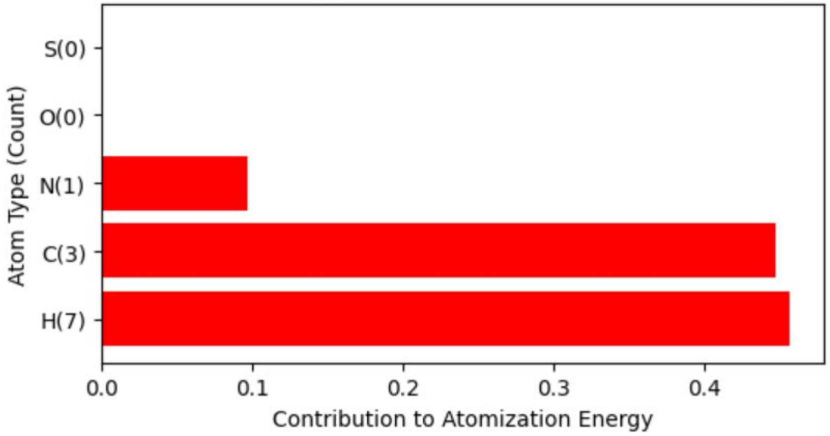
Weights



Model 1 Explanations

Local Explanation 04

Contribution of Atom Types to AE for n-Alkyl N-C3H7



Chemical Explanation 05

Typical AE for Atom Pairs

	C	H	N	O	S
C	347	413	305	358	259
H	413	432	391	467	347
N	305	391	160	201	*Na*
O	358	467	201	146	*Na*
S	259	347	*Na*	*Na*	266

C=C	C ≡ C	O=O	O=C	C ≡ O	N=O	N=N	N ≡ N	C ≡ N	C=N
614	839	495	745	1072	607	418	941	891	615

Model 2: Representation with pairs of atoms and distance

Feature Map 01

Softmax for distance of atom pairs

$$\phi^A(\mathcal{E}_i) = \begin{pmatrix} \exp -\frac{(\mathrm{dist}(\mathcal{E}_i)-\mu_1)^2}{2\sigma^2} \\ \exp -\frac{(\mathrm{dist}(\mathcal{E}_i)-\mu_2)^2}{2\sigma^2} \\ \vdots \\ \exp -\frac{(\mathrm{dist}(\mathcal{E}_i)-\mu_n)^2}{2\sigma^2} \end{pmatrix}$$

One hot representation atom pairs

$$\phi^B(\mathcal{E}_i) = \begin{pmatrix} I(\mathrm{type}(\mathcal{E}_i) = \mathrm{HH}) \\ I(\mathrm{type}(\mathcal{E}_i) = \mathrm{HC}) \\ I(\mathrm{type}(\mathcal{E}_i) = \mathrm{HN}) \\ \vdots \\ I(\mathrm{type}(\mathcal{E}_i) = \mathrm{SS}) \end{pmatrix}$$

$$\Rightarrow \quad \boldsymbol{x} = \sum_{i=1}^{|\mathcal{M}|} \phi(\mathcal{E}_i)$$

where $\phi(\mathcal{E}_i)$ is the flattened matrix of shape (300,) defined by:

$$[\phi(\mathcal{E}_i)]_{jk} = \phi_j^A(\mathcal{E}_i) \cdot \phi_k^B(\mathcal{E}_i)$$

Model Training 02

Ridge Regression

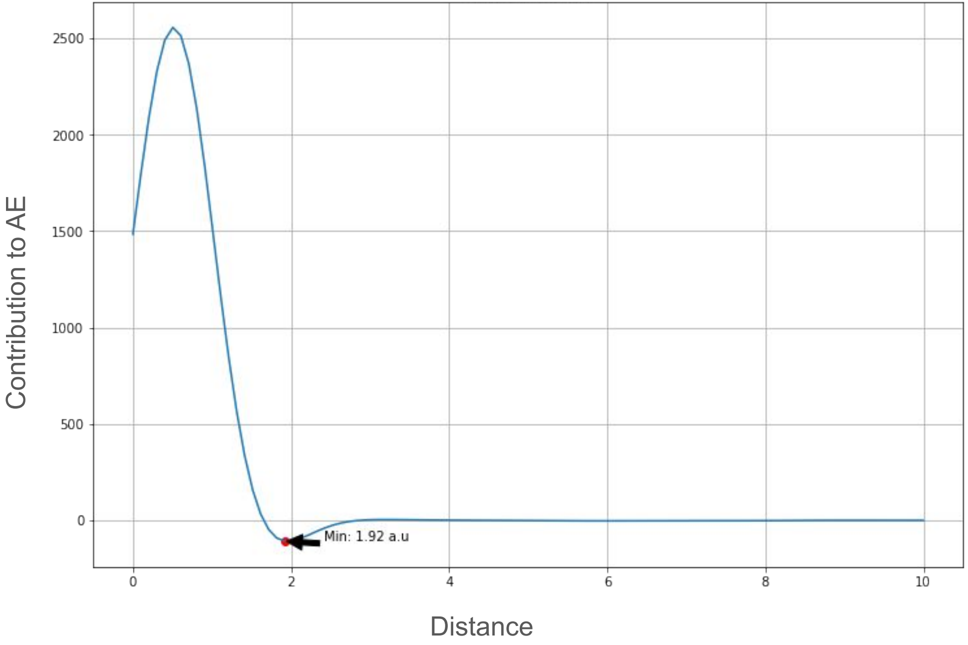
Optimal Parameters:

Ridge regression:	$\lambda = 10^{-5}$
Number of Bins:	n= 20
Variance for Softmax:	σ= 0.5

Mean Absolute Test error: 6.33 kJ/mol

Model 2: Explanations

Potential of atom pair O-H



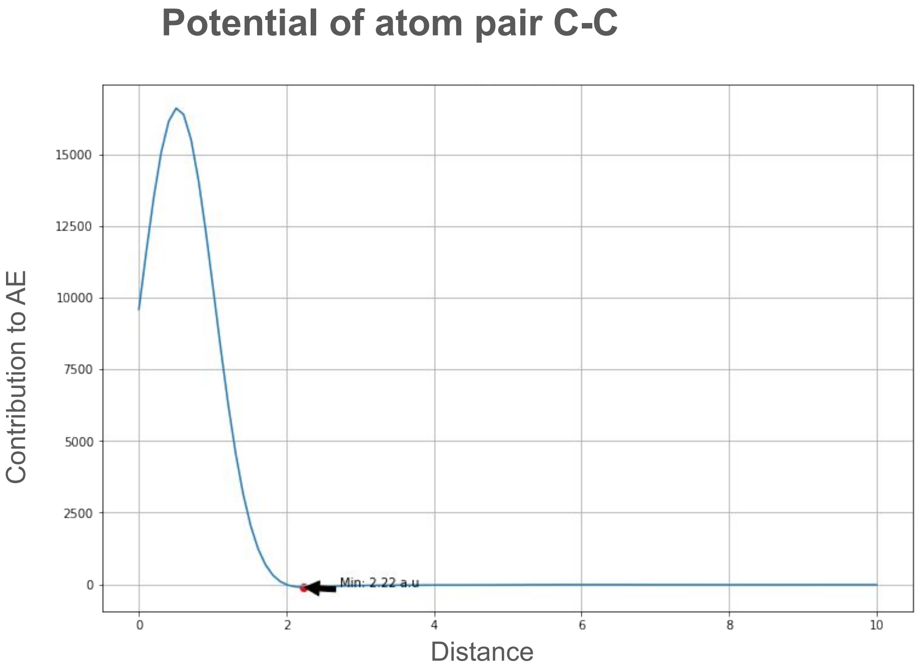
Typical bond lengths for single bonds in AU

	C	H	N	O	S
C	2.87	2.04	2.78	2.70	3.44
H	2.04	1.40	1.91	1.83	2.53
N	2.78	1.91	2.76	2.57	3.31
O	2.70	1.83	2.57	2.80	3.04
S	3.44	2.53	3.31	3.04	3.87

The model's predicted potential for pair H-O is **minimal** at the typical bond length of the atom pair.

Two atoms are in a **stable state** (= lowest energy) at their bond lengths. → At their typical bond length the energy to break their bond, which corresponds to their **AE is lowest**.

Model 2: Explanations



Typical bond lengths single bonds in AU

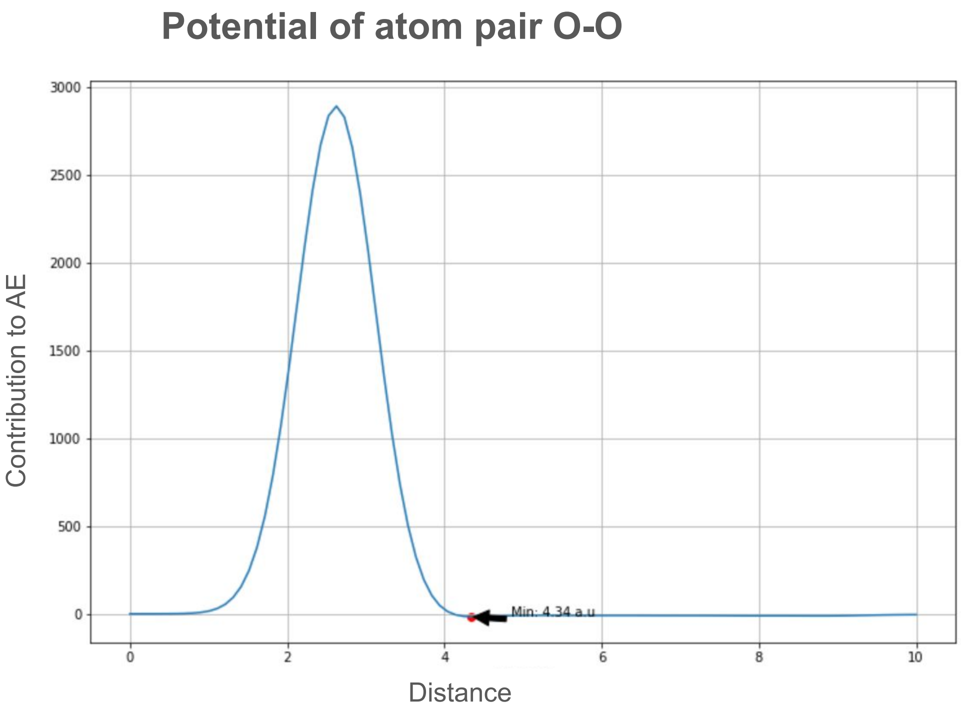
	C	H	N	O	S
C	2.87	2.04	2.78	2.70	3.44
H	2.04	1.40	1.91	1.83	2.53
N	2.78	1.91	2.76	2.57	3.31
O	2.70	1.83	2.57	2.80	3.04
S	3.44	2.53	3.31	3.04	3.87

C=C	C ≡ C	O=O	O=C	C ≡ O	N=N	N ≡ N	C ≡ N	C=N
2.53	2.27	2.29	2.29	2.14	2.31	2.14	2.17	2.40

Typical single bond length of C-C pair is **significantly higher** than the distance with minimal energy predicted by the model.

Typical lengths for **double/ triple bonds** are lower!

Model 2: Explanations



Typical bond lengths single bonds in AU

	C	H	N	O	S
C	2.87	2.04	2.78	2.70	3.44
H	2.04	1.40	1.91	1.83	2.53
N	2.78	1.91	2.76	2.57	3.31
O	2.70	1.83	2.57	2.80	3.04
S	3.44	2.53	3.31	3.04	3.87

C=C	C ≡ C	O=O	O=C	C ≡ O	N=N	N ≡ N	C ≡ N	C=N
2.53	2.27	2.29	2.29	2.14	2.31	2.14	2.17	2.40

Minimal Potential at distance much higher than typical bond length.

Possible Limitations:
- not enough examples of this pair in the dataset
- groups of atoms contribute to AE

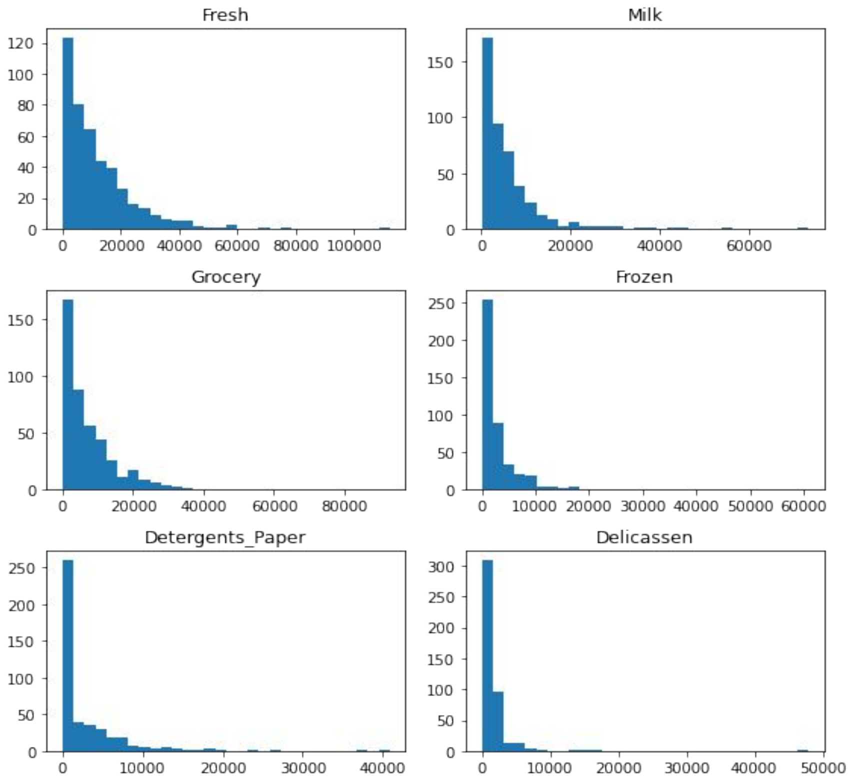
I. Getting Insights into an Unsupervised Dataset

Presented by: **Sona Mehdizade**

Dataset Overview

- **Dataset:** UCI Wholesale customers
  The dataset consists of annual spending on 6 different product categories by 440 wholesale customers.

- **Objective:** Identify patterns and anomalies in spending behavior of customers and extract explanations for these anomalies.



Data Preprocessing - Log Transformation

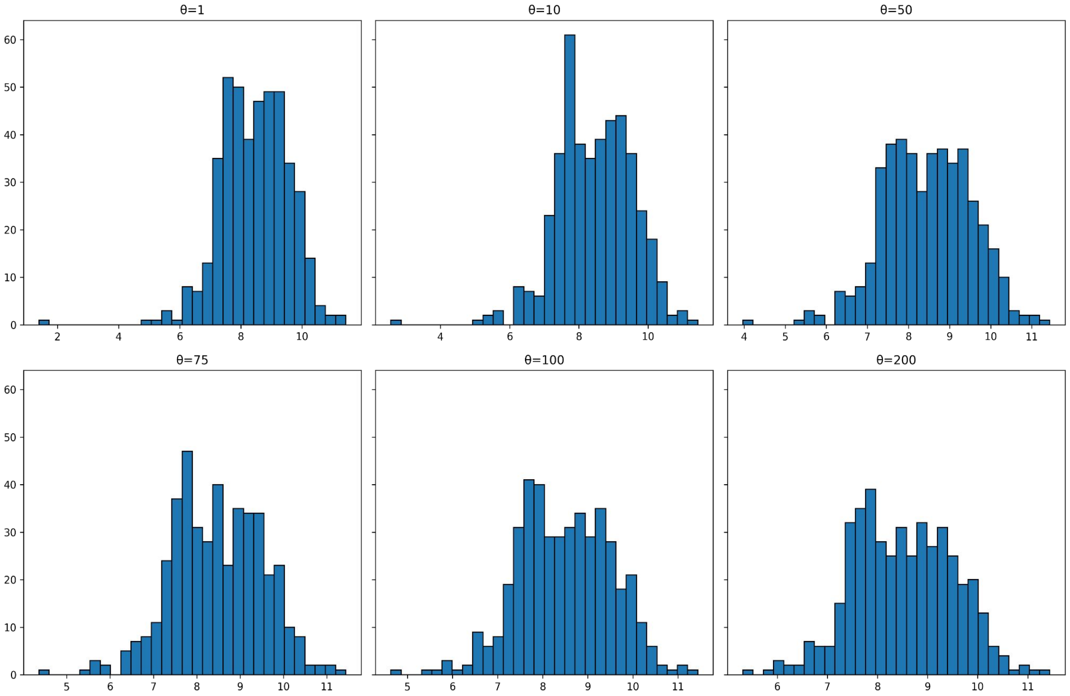
Selection of θ:

- Visual evaluation with histograms
- Jarque-Bera normality test

$$x \mapsto \log(x + \theta)$$

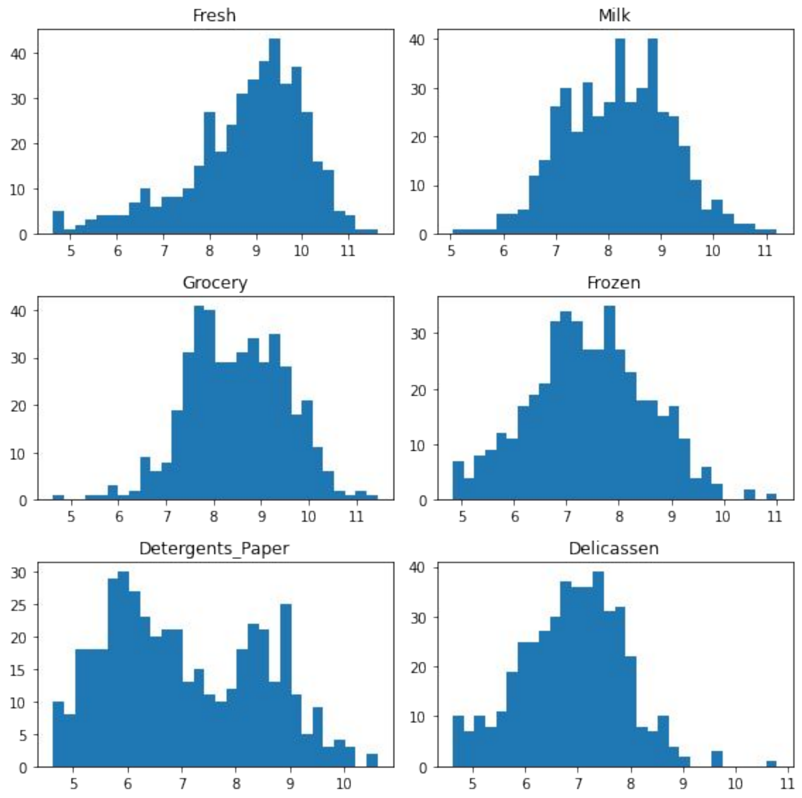
For Grocery category

θ	JB Statistics	P-value
1	210.779	0.000
10	52.777	0.000
50	5.540	0.063
75	2.309	0.315
100	**1.170**	**0.557**
200	1.287	0.526



Distributions after Transformation

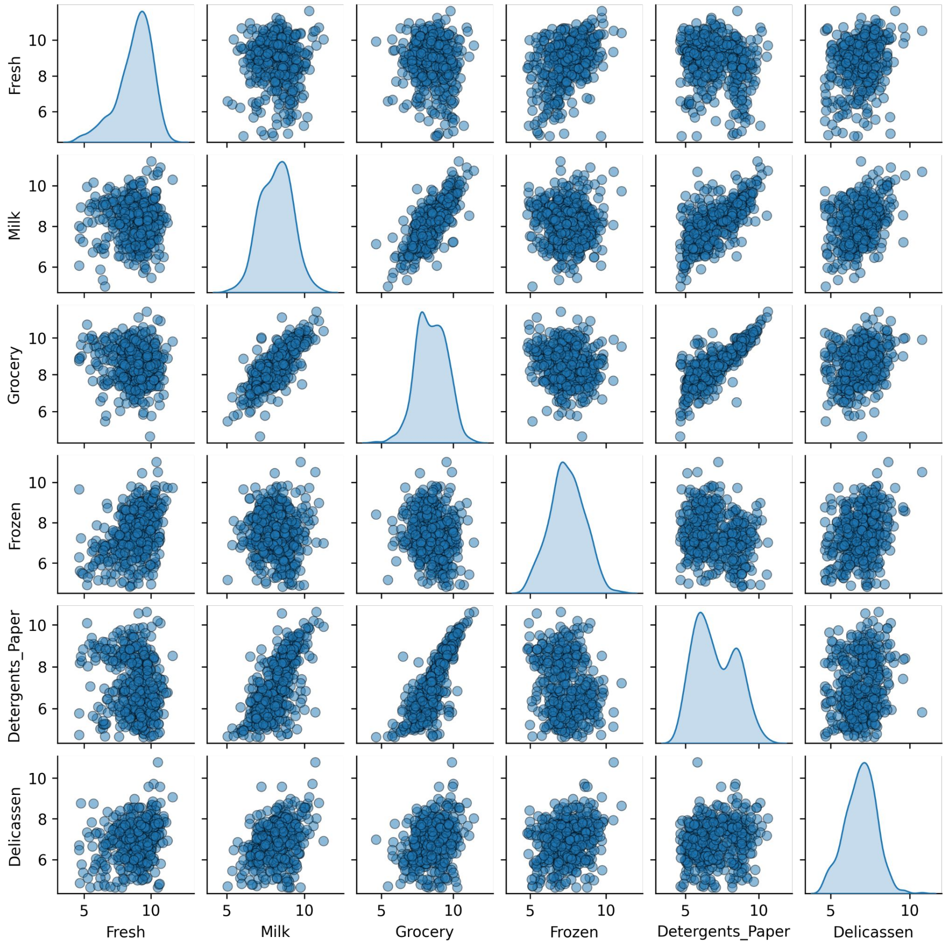
- After transformation (θ=100) , the distributions appears more Gaussian.
- It ensures that high spenders do not overshadow typical spenders, making it possible to analyze patterns among the majority of the customers.



Correlation between spending categories

High Correlation Between Categories:

- Grocery and Detergents/Paper
- Milk and Grocery
- Milk and Detergents/Paper



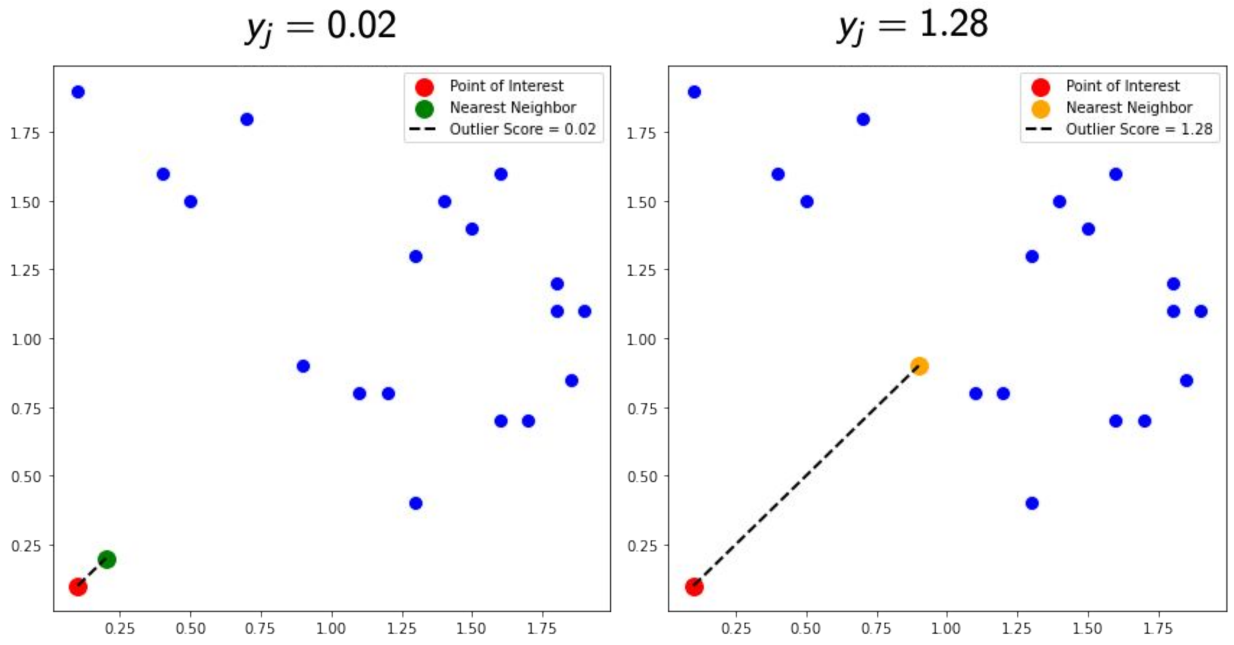
Outlier Detection

1. **Nearest Neighbor-based Anomaly Detection**.

$$z_{jk} = \|\boldsymbol{x}_j - \boldsymbol{x}_k\|^2$$

$$y_j = \min_{k \neq j} z_{jk}$$

If the nearest neighbor is very close but the next one is far, the outlier score can change drastically if the nearest point is missing.

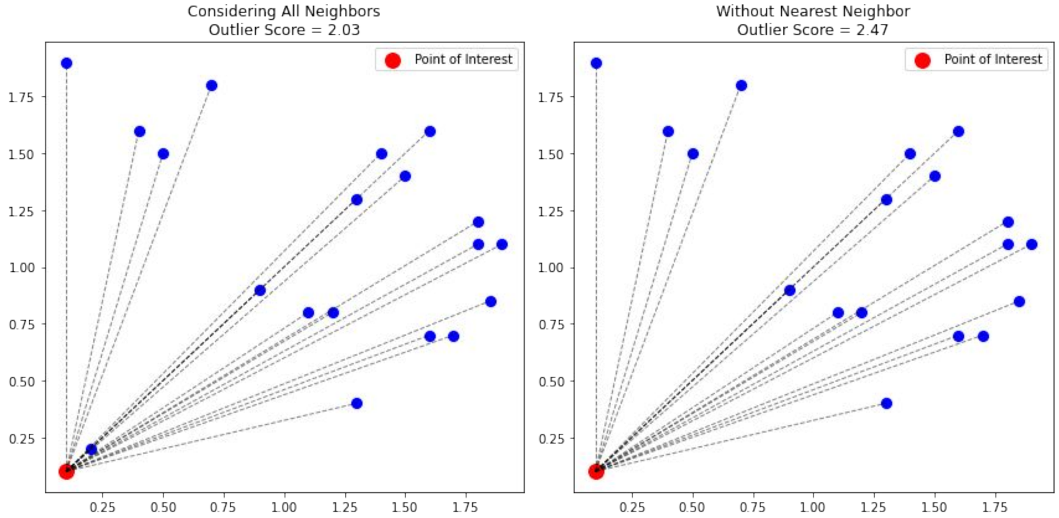


Robust anomaly detection

- Soft minimum approach: factor in all neighbors' contributions
- More robust to dataset variations

$$y_j = \text{soft min}_{k \neq j} z_{jk} = -\frac{1}{\gamma} \log \left(\frac{1}{N-1} \sum_{k \neq j} \exp(-\gamma z_{jk}) \right)$$

- The soft minimum is computed using the log-sum-exp trick to maintain numerical stability.



Selecting a suitable parameter γ

The parameter ***γ*** has to be chosen to maximize distinction between anomalous and non-anomalous instances.

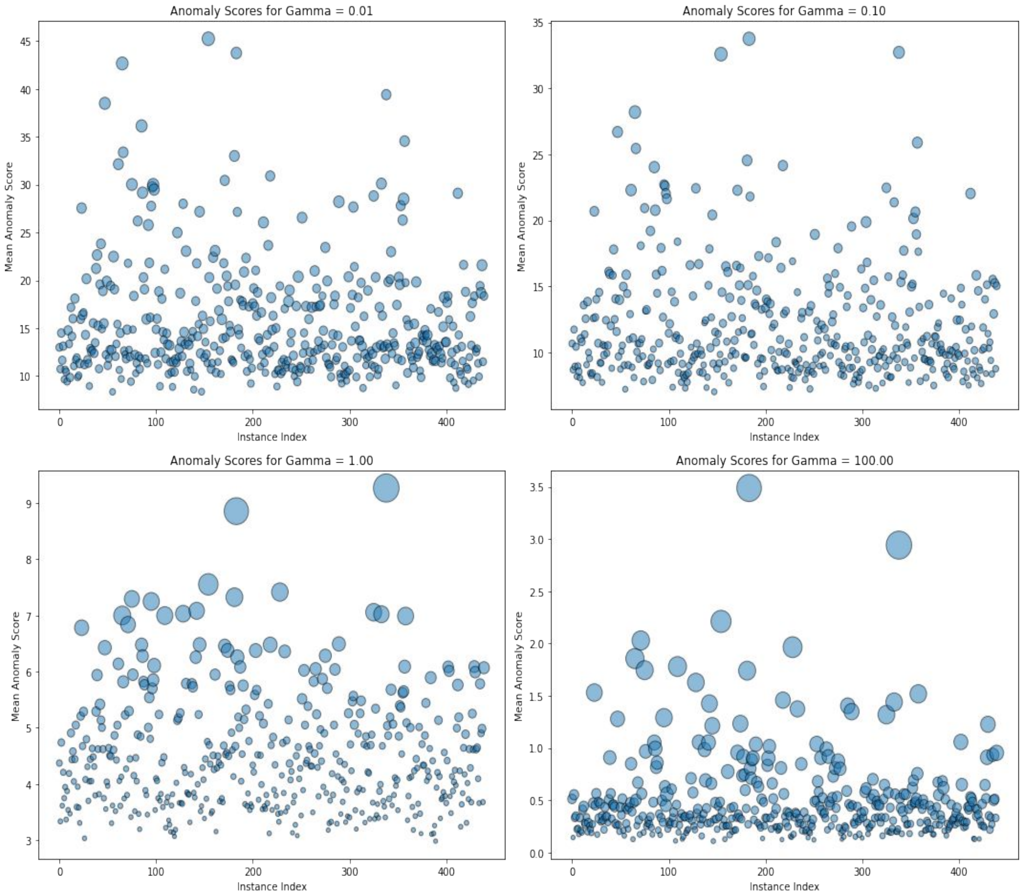
Bootstrapping Approach:

- Simulate several dataset variants with replacement.
- Calculate anomaly scores for each variant.
- Analyze mean and spread of scores.

Optimal γ Selection:

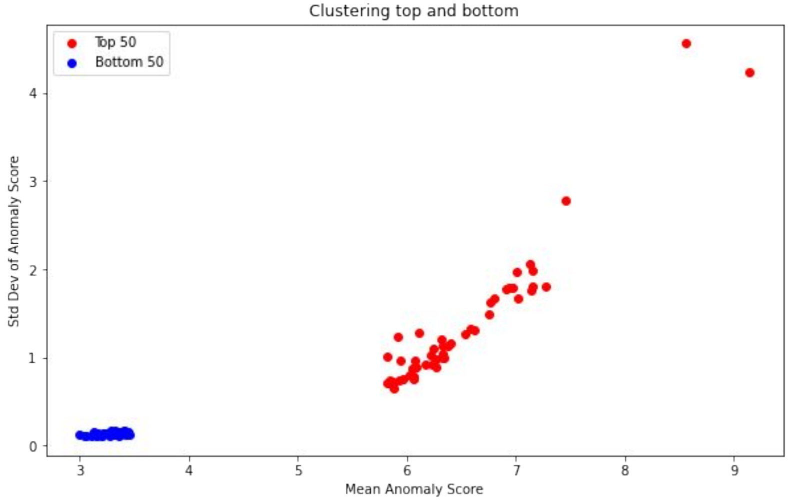
- Balance between separability and spread.
- Visual and metric-based evaluation.

γ=1 seems optimal



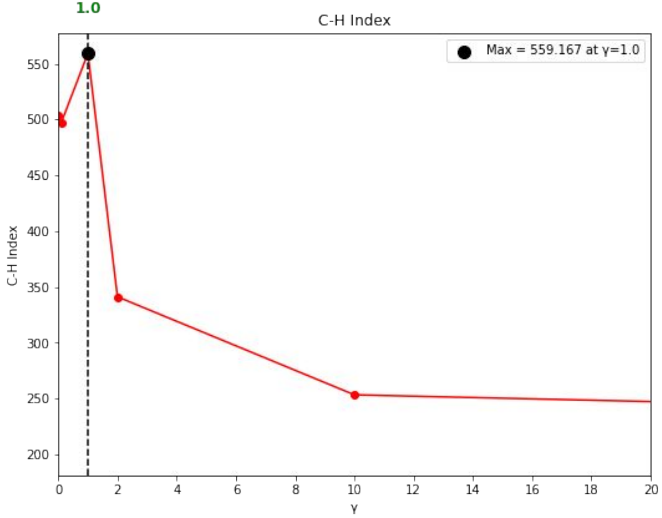
Optimal γ with evaluation metric

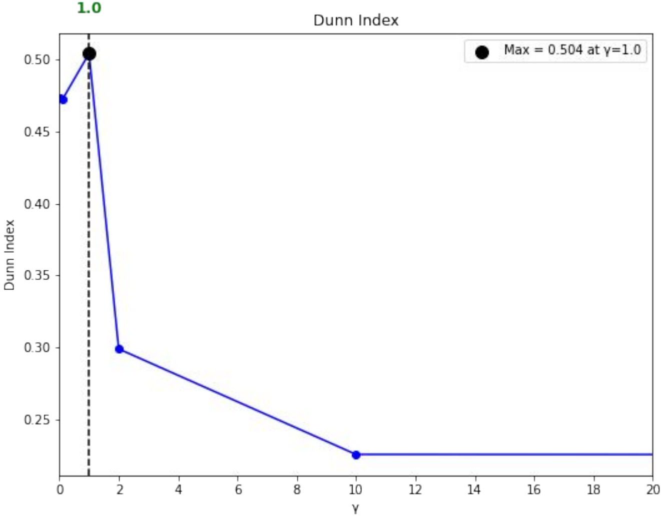
1. Select the top 50 and bottom 50 instances based on their mean anomaly scores.
2. Apply K-means clustering with 2 clusters on the combined dataset.
3. Evaluate clustering quality by C-H, Dunn



$$\text{Calinski-Harabasz} = \frac{\sum_k \sum_{i \in C_k} \|\mu_k - \mu\|^2}{\sum_k \sum_{i \in C_k} \|x_i - \mu_k\|^2} \cdot \frac{N-K}{K-1}$$

$$\text{Dunn} = \frac{\min_{(i,j):c_i \neq c_j} \|x_i - x_j\|}{\max_{(i,j):c_i = c_j} \|x_i - x_j\|}$$

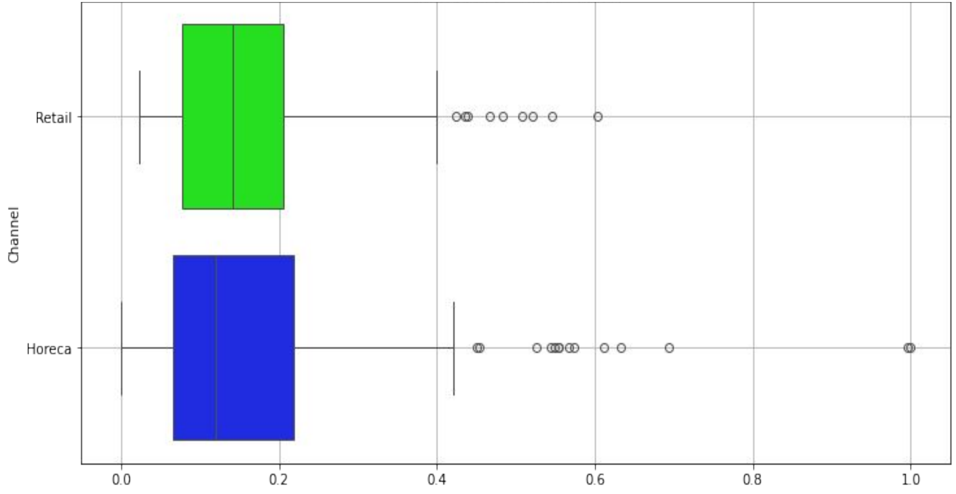




Relation between Anomalies and Metadata

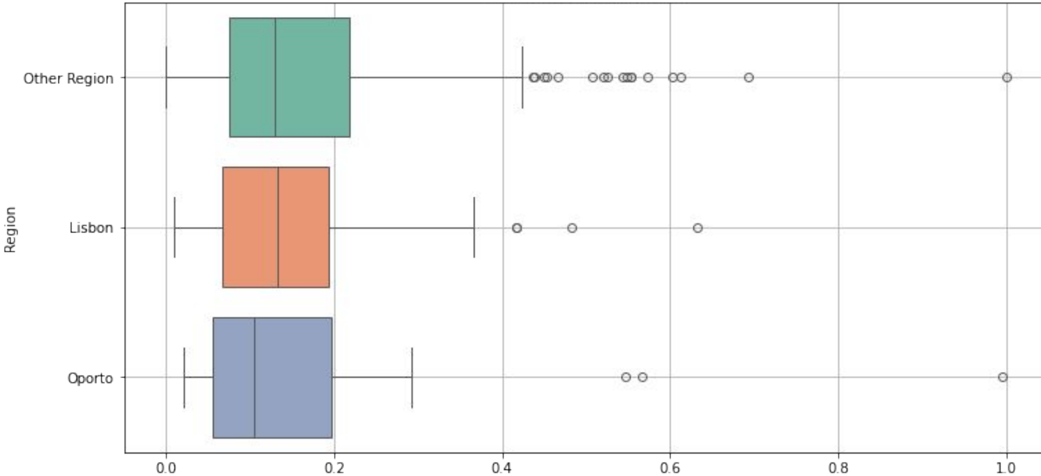
Channel-wise Analysis:

1. **Retail:** Higher median score, Fewer extremes
2. **Horeca(Hotel, Restaurant, Café):** Lower median, more extreme anomalies

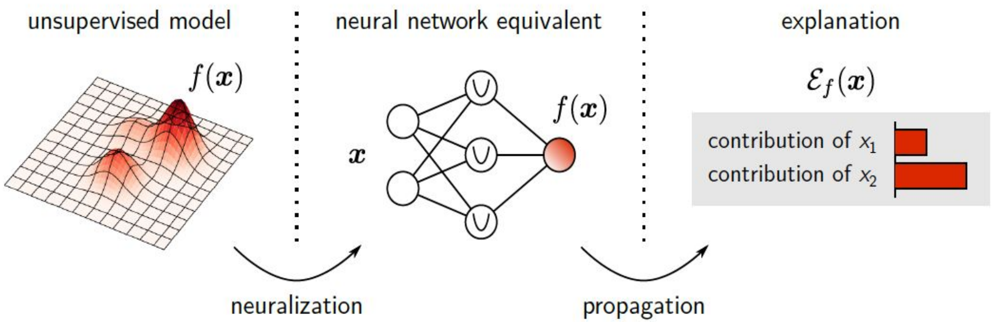


Region-wise Analysis:

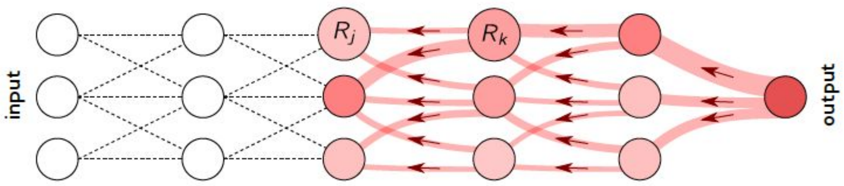
1. **Other Region:** High median, many extreme outliers
2. **Lisbon:** Slightly higher median, fewer extremes
3. **Oporto:** Lowest median, fewest extremes



XAI LRP application for Explanations - NEON approach



Source: Montavon et al., 2022



Source: Montavon et al., 2019

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k$$

NEON Method: Converts an unsupervised model to an equivalent neural network for LRP.

1. **Neuralization:** Rewrites the unsupervised model as a neural network.
2. **Propagation**: Applies LRP to the neural network to explain predictions.

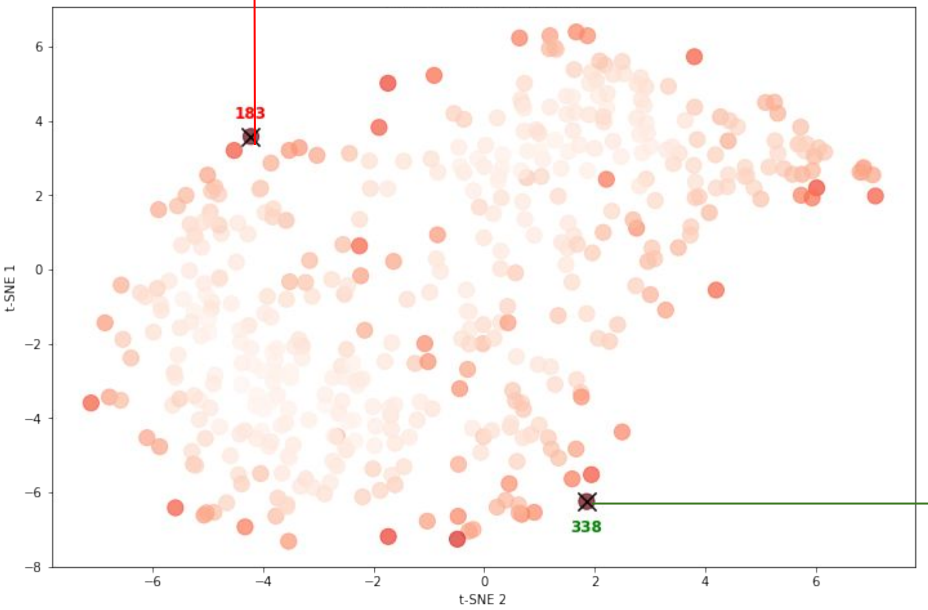
Layer 1: $z_{jk} = \|x_j - x_k\|^2$

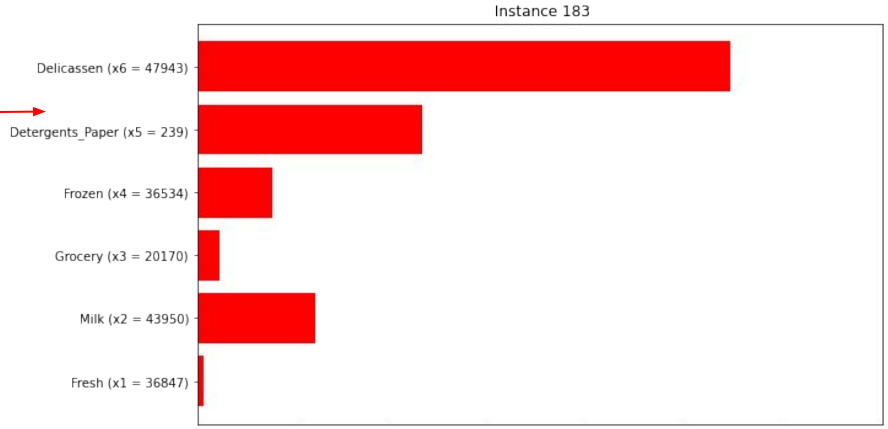
Layer 2: $y_j = \text{soft min}_{k \neq j}\, z_{jk} = -\frac{1}{\gamma} \log \left(\frac{1}{N-1} \sum_{k \neq j} \exp(-\gamma z_{jk}) \right)$

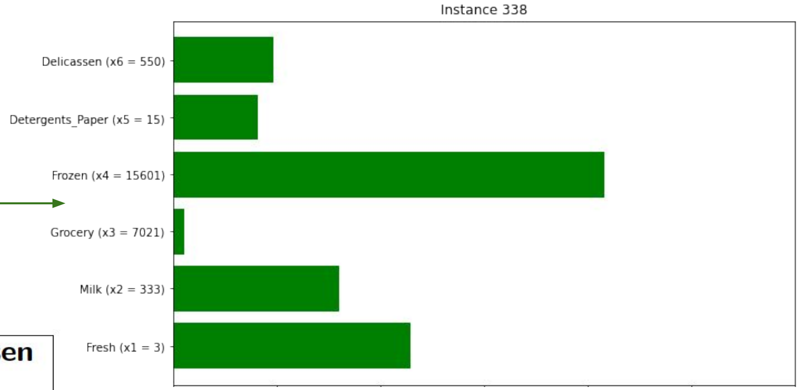
Layer 2 to Layer 1: $R_k^{(j)} = \frac{\exp(-\gamma z_{jk})}{\sum_{k \neq j} \exp(-\gamma z_{jk})} \cdot y_j$

Layer 1 to Input Features: $R_i^{(j)} = \sum_{k \neq j} \frac{(x_{ki} - x_{ji})^2}{\|x_k - x_j\|^2} \cdot R_k^{(j)}$

Explanation for an instance

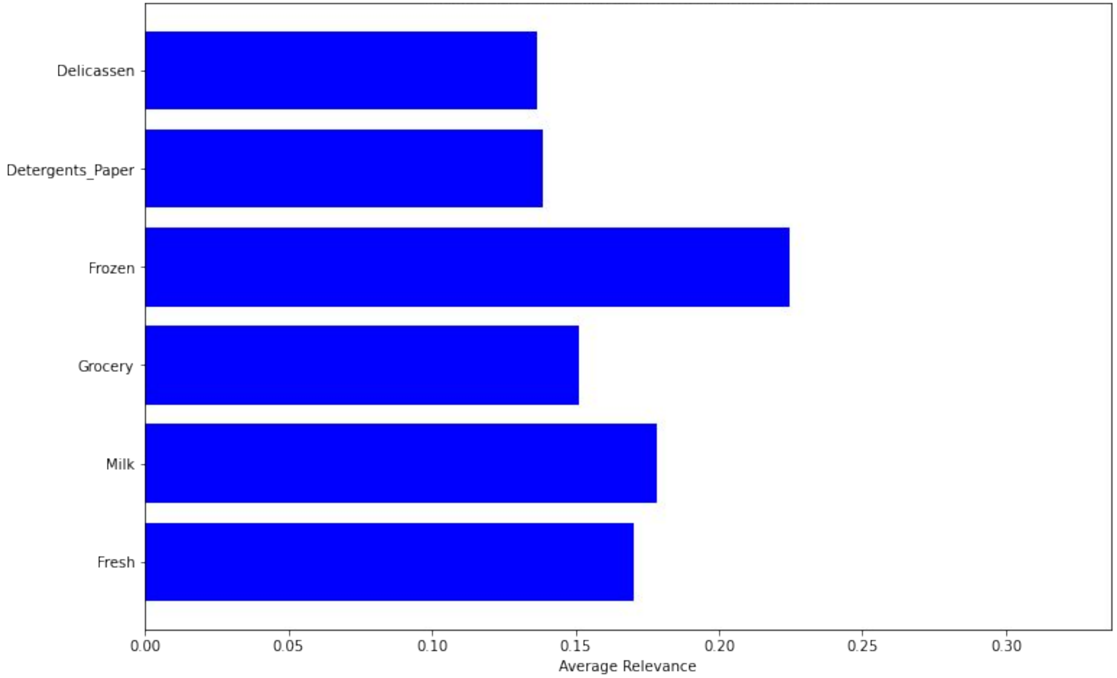






instance	fresh	milk	grocery	frozen	detergents paper	delicatessen
183	36847	43950	20170	36534	239	47943
338	3	333	7201	15601	15	550
AVG	12000	5796	7951	3072	2881	1525

Feature Relevance Analysis for Top 10 Outliers



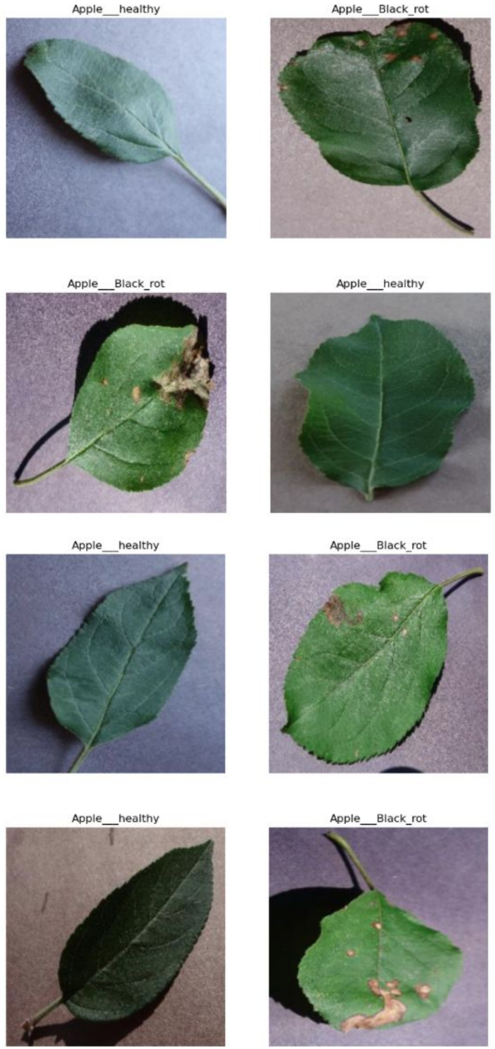
- Normalized contributions to compare across varied outlierness scores
- Frozen category most significant for top outliers

III. Getting Insights into Images and their Metadata

Presented by: **Jakub Tarka**

Introduction

- **Dataset**: Images of apple leaves in two classes: healthy and rotten
- **Objective**: determine relation between images and their metadata
- Instead of working in pixel space, use a pre-trained image classification neural network: **VGG-16**
- Extract more general lower-level feature representation

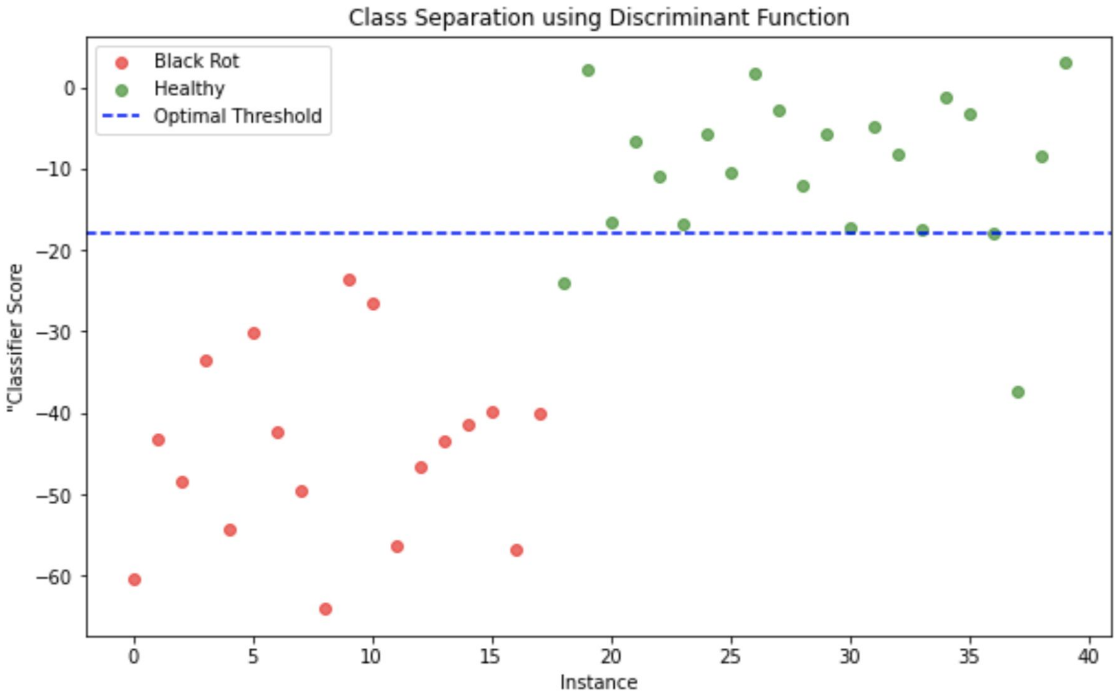


Difference of Means

- Difference of Means applied on feature representation
- ***g(x)*** contains classifier scores for each instance
- Clear separation of the two classes
- Youden's J statistic to determine optimal separation threshold

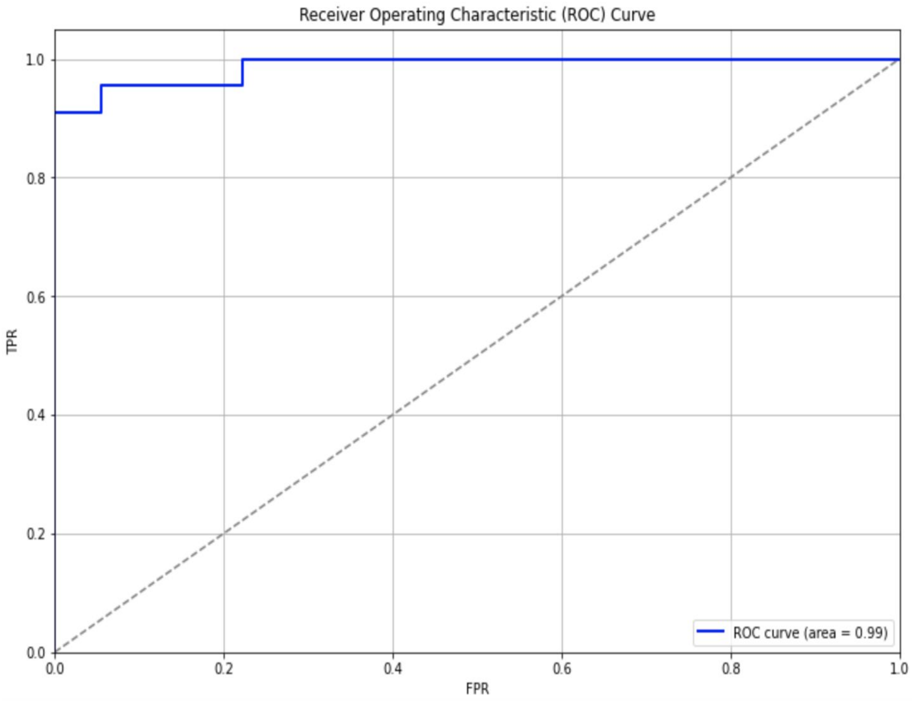
$$w = \frac{\mu_2 - \mu_1}{\|\mu_2 - \mu_1\|} \qquad \mu_1 = \frac{1}{|\mathcal{C}_1|} \sum_{i \in \mathcal{C}_1} \Phi(x_i)$$

$$g(x) = w^\top \Phi(x) \qquad \mu_2 = \frac{1}{|\mathcal{C}_2|} \sum_{i \in \mathcal{C}_2} \Phi(x_i)$$



ROC Curve and ROC-AUC Score

- Metric to quantify the quality of the Difference of Means classifier
- Test instances: 18 healthy, 22 black rot
- AUC metric implies near perfect classification (small variations)



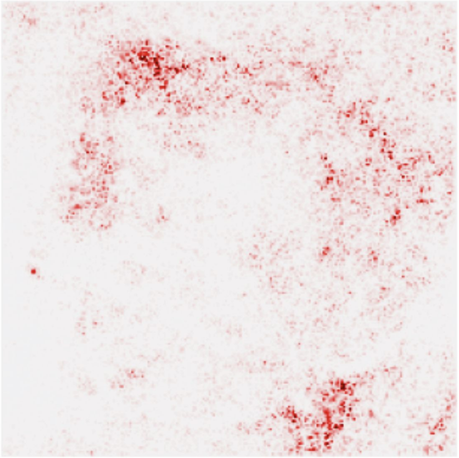
Sensitivity Analysis

- Classes are separable, but why?
- Find which input features contribute the most to the classification score
- Gradient-based approach:
  - Pass image through the network
  - Perform backward pass to compute gradients
  - Compute norm over all 3 channels
  - Output is a 224x224 tensor which can be rendered as a heatmap
- Result is noisy and misses key features.

$$S_i = \left\| \frac{\partial g}{\partial x_i} \right\|^2$$

Apple Black Rot - Sample Image

Heatmap



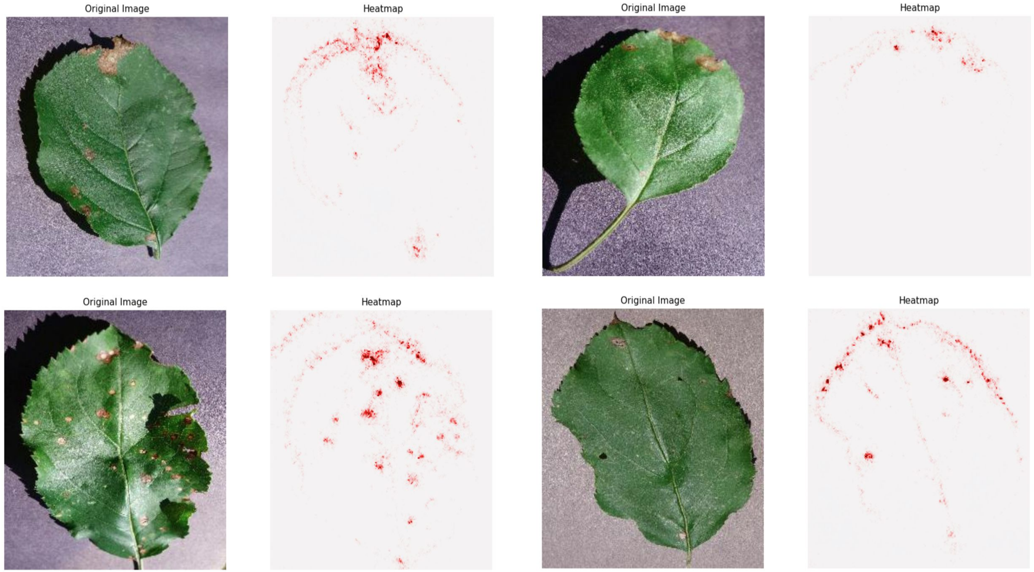
Robustification - Biased Layer

$$z_k = \Big(\sum_j a_j w^{\uparrow}_{jk} + b^{\uparrow}_k\Big) \cdot \left[\frac{\sum_j a_j w_{jk} + b_k}{\sum_j a_j w^{\uparrow}_{jk} + b^{\uparrow}_k}\right]_{\text{cst.}}$$

$$w^{\uparrow}_{jk} = w_{jk} + 0.25 \max(0, w_{jk})$$

$$b^{\uparrow}_k = b_k + 0.25 \max(0, b_k).$$

- **Idea:** Bias the gradient to highlight excitatory effects.
- Create Biased Layers which amplify those weights which have a positive value.
- Replace Convolutional and Linear layers with Biased Layers
- Visible improvement: less noise, more accuracy
- High level of inaccuracy remains



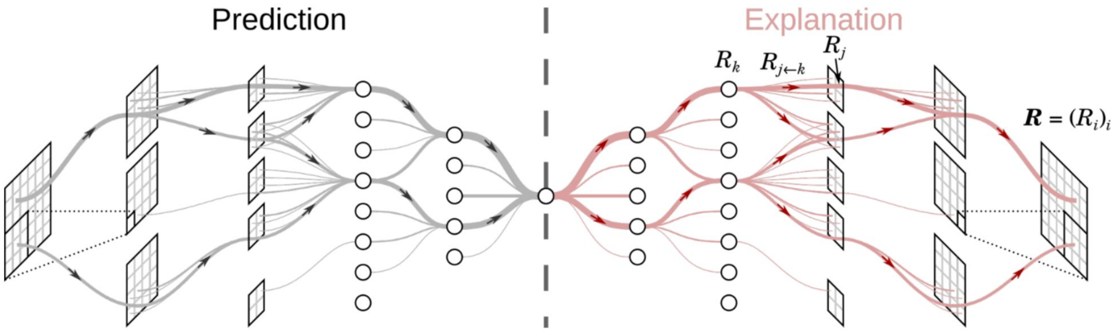
Robustification - Discussion

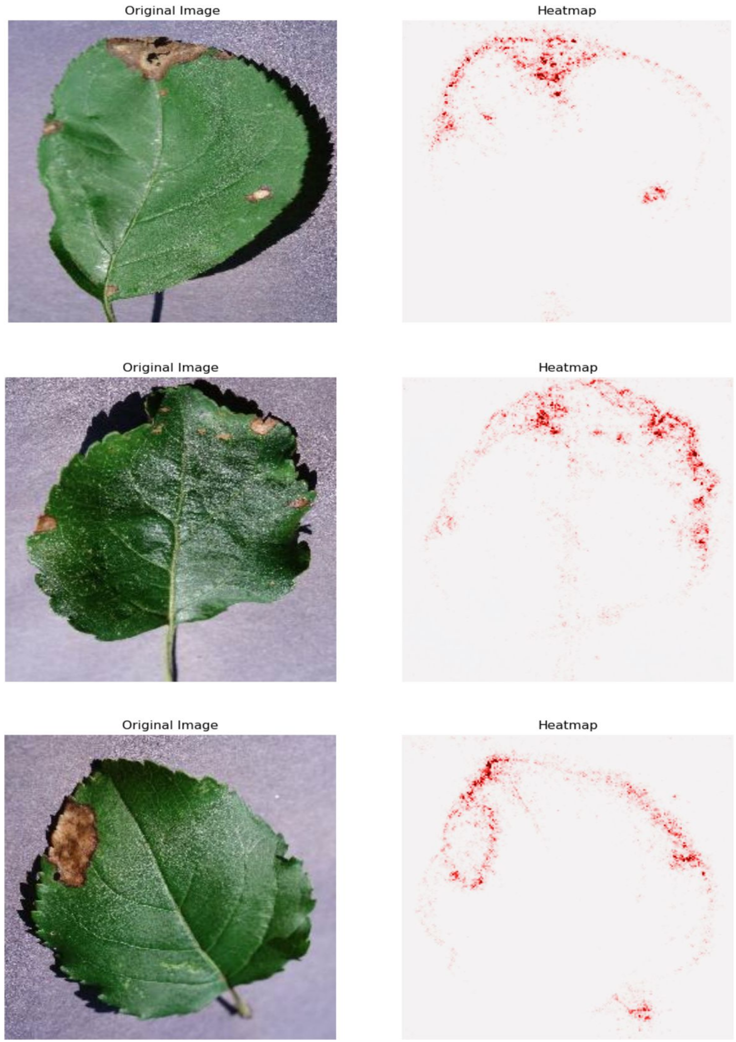
1. Explanation Method

- Gradient-based methods are noisy

Solutions:

- More robust methods available, e.g. *SmoothGrad* - reduces noise by adding noise
- Propagation-based methods: *LRP*

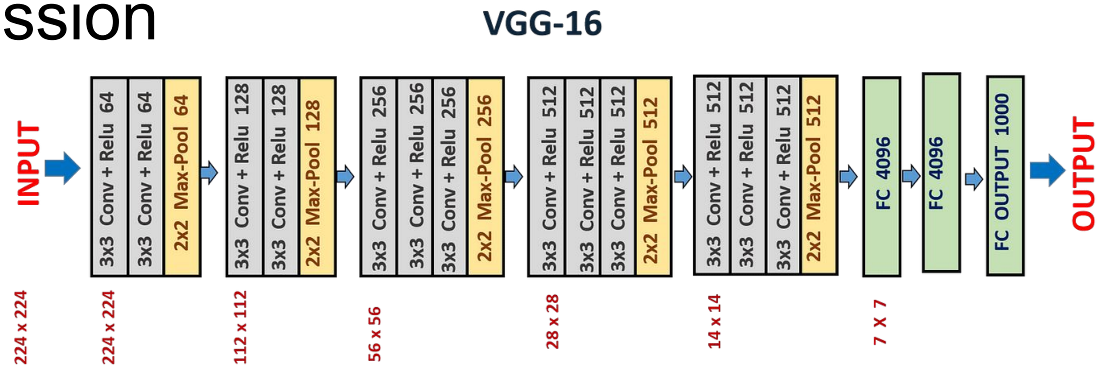




Robustification - Discussion

2. Pre-Trained Model

- VGG-16 not trained on plants, trees, leaves
- No understanding of what a leaf is supposed to look like



Solutions:

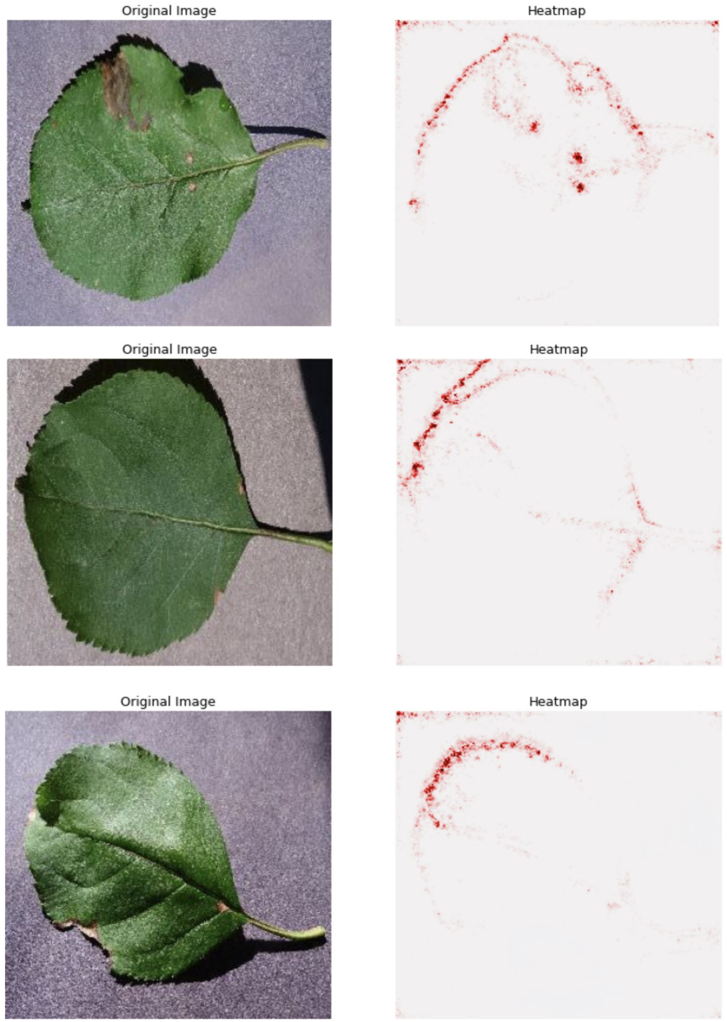
- Fine-Tuning the Model - add layers and train on the given leaf dataset
- Domain-Specific Pre-Trained Models are available, trained on *PlantVillage* dataset
- Segmentation-Based Models

Robustification - Discussion

3. **Data Quality**
    - Leaf edges and shadows are often highlighted due to abrupt pixel changes
    - Possible training data bias between class metadata

Solutions:

- Extract bounding box of leaf parts only
- Augment dataset:
    - Different backgrounds
    - Rotation/Flipping
    - Light Intensity
    - Scaling and cropping
    - Random noise
- Blur / reduce image size



References

Project 1:

1. G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller. "Layer-Wise Relevance Propagation: An Overview," Lecture Notes in Computer Science, vol. 11700, pp. 193–209, Springer, 2019.
2. G. Montavon, J. R. Kauffmann, W. Samek, and K.-R. Müller. "Explaining the Predictions of Unsupervised Learning Models," Lecture Notes in Computer Science, vol. 13200, pp. 117–138, Springer, 2020.

Project 2:

1. Fundamentals of Chemical Bonding, LibreTexts Chemistry
2. Bindungslängen - Internetchemie

Project 3:

1. K. Simonyan, A. Vedaldi, and A. Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," Visual Geometry Group, University of Oxford.
2. D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. "SmoothGrad: removing noise by adding noise."

Thank you!

Do you have any question?