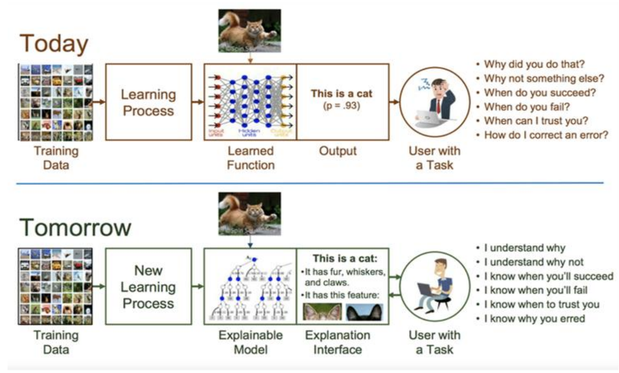
Towards Explainable Image Classification

Vahidin Hasić, University of Sarajevo, Sarajevo, Bosnia and Herzegovina,
vahidin.hasic@etf.unsa.ba

Motivation

While Deep Neural Networks (DNNs) excel in image classification, their black-box nature necessitates the development of Explainable AI (XAI) methods. Existing XAI techniques face limitations in balancing explainability, fidelity, and efficiency.



Lin, Cheng-Jian & Jhang, Jyun-Yu. (2021). Bearing Fault Diagnosis Using a Grad-CAM-Based Convolutional Neuro-Fuzzy Network. Mathematics. 9. 1502. 10.3390/math9131502.

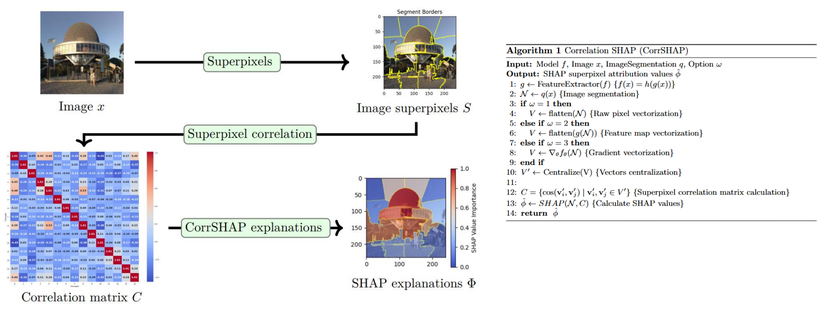
Thesis Contributions:

- **Perturbation-based explanations –** explaining model predictions by perturbing the input.
- **Concept-based explanations -** explaining model predictions by leveraging human-understandable concepts.
- **Sample-based explanations -** explaining model predictions by leveraging training data.

CLEAR-VISION: Credible Learning for Explainable and Reliable Visual Recognition

CLEAR-VISION

Correlation SHAP (CorrSHAP)



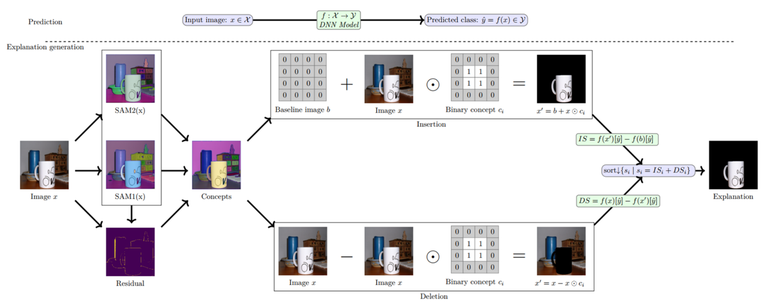
Algorithm 1 Correlation SHAP (CorrSHAP)

Input: Model f, Image x, ImageSegmentation q, Option ω
Output: SHAP superpixel attribution values ϕ

```
g ← FeatureExtractor(f) {f(x) = h(g(x))}
N ← q(x) {Image segmentation}
if ω = 1 then
    V ← flatten(N) {Raw pixel vectorization}
else if ω = 2 then
    V ← flatten(g(N)) {Feature map vectorization}
else if ω = 3 then
    V ← ∇x f_θ(N) {Gradient vectorization}
end if
V' ← Centralize(V) {Vectors centralization}
C = {cos(v'_i, v'_j) | v'_i, v'_j ∈ V'} {Superpixel correlation matrix calculation}
φ ← SHAP(N, C) {Calculate SHAP values}
return φ
```

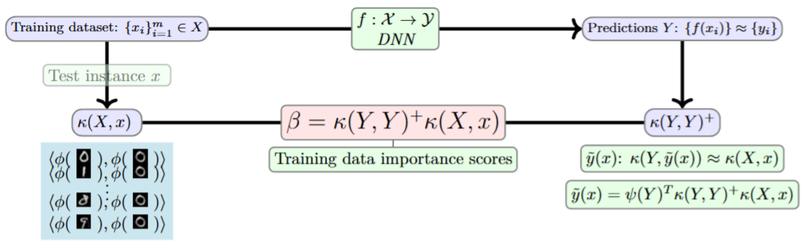
Framework of the proposed CorrSHAP method. The input image is segmented into superpixels. Superpixels are vectorized and centralized into vectors. The correlation matrix between superpixels is calculated using cosine similarity between individual vectors. For each superpixel, we take correlated superpixels, where correlation is higher than the threshold, and perform perturbations on all combinations to calculate superpixel attribution. [1]

Any Segment Explanation (ASE)



Any Segment Explanation (ASE) overview. An input image is classified, yielding a certain prediction. It is segmented with the Segment Anything Model. Segments are treated as concepts and are transformed into binary masks. Perturbed images are generated by inserting/deleting concepts. Insertion/Deletion Scores (IS/DS) measure model prediction change. The concepts are ranked by combining IS and DS scores, where the k top-scoring concepts are shown as the explanation. [2]

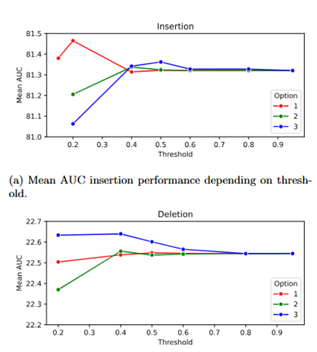
Kernel Sample Based Explanations (K-SBE)



A pretrained model generates predictions Y for a training dataset X, where each image is flattened into a feature vector, forming a matrix X. The output matrix Y has rows for instances and columns for output dimensions. For a test instance x, the kernel matrix $\kappa(X, x)$ and the pseudoinverse $\kappa(Y, Y)^+$ are computed. The attribution scores for the training instances are calculated using the dot product $\kappa(Y, Y)^+ \cdot \kappa(X, x)$. Sorting this vector reveals the importance of each training instance x_i related to the test instance x. [3]

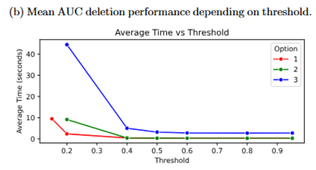
Results

Correlation SHAP (CorrSHAP)

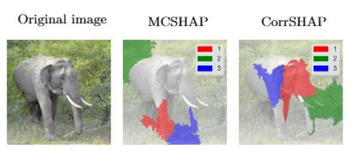


(a) Mean AUC insertion performance depending on threshold.

(b) Mean AUC deletion performance depending on threshold.



(c) Execution time depending on threshold.

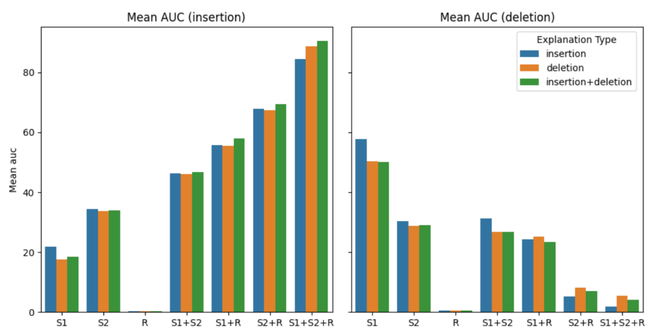


Area Under the Curve (AUC) Insertion ↑					
Model	Superpixels	CorrSHAP 1	CorrSHAP 2	CorrSHAP 3	MCSHAP
MobileNet-v2	Quickshift	80.4	80.37	80.29	**80.89**
	SLIC	**78.12**	78.13	78.15	77.79
ResNet-18	Quickshift	60.21	60.23	60.21	**61.60**
	SLIC	54.66	54.65	54.65	**55.41**
ResNet-50	Quickshift	**82.63**	82.61	82.63	82.27
	SLIC	**81.20**	81.20	81.20	80.66
ViT-b16	Quickshift	80.83	80.83	80.74	**81.84**
	SLIC	76.36	76.33	76.41	**76.82**
Area Under the Curve (AUC) Deletion ↓					
MobileNet-v2	Quickshift	**20.14**	20.14	20.16	20.93
	SLIC	**19.48**	19.48	19.48	21.40
ResNet-18	Quickshift	8.25	8.25	8.25	**8.03**
	SLIC	9.06	9.06	9.06	**9.01**
ResNet-50	Quickshift	**22.79**	22.79	22.79	24.16
	SLIC	**22.36**	22.36	22.39	23.79
ViT-b16	Quickshift	17.20	17.17	17.13	**16.86**
	SLIC	20.53	20.53	20.53	**20.48**
Execution Time (seconds) ↓					
MobileNet-v2	Quickshift	**0.42**	0.54	1.15	16.13
	SLIC	**0.74**	0.89	1.86	14.98
ResNet-18	Quickshift	**0.43**	0.50	1.09	7.85
	SLIC	0.66	**0.51**	1.91	13.82
ResNet-50	Quickshift	**0.46**	0.58	2.76	25.16
	SLIC	**0.78**	0.93	5.93	36.49
ViT-b16	Quickshift	**0.40**	0.52	5.41	7.26
	SLIC	**0.73**	1.01	10.53	18.00

Performance analysis as a function of the threshold. **Option 1** - raw pixel vectorization, **Option 2** - feature map vectorization, and **Option 3** - gradient vectorization. Lower AUC deletion, higher AUC insertion scores, and lower execution time indicate superior performance. The execution time decreases sharply until 0.5 threshold, without sacrificing method faithfulness. Further reductions of threshold yields diminishing returns in performance.

Any Segment Explanation (ASE)

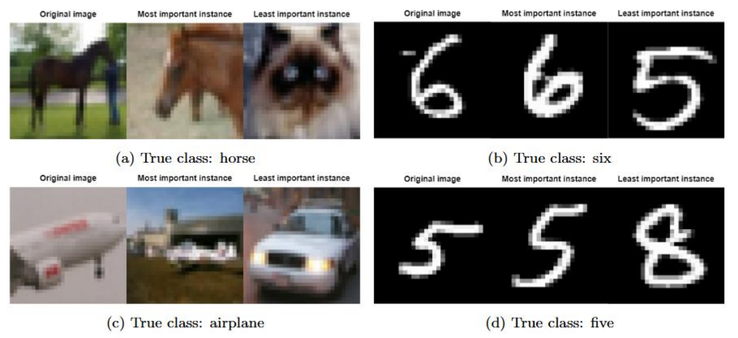
		ASE(ours)	EAC*	DeepLIFT*	GradSHAP*	IntGrad*	KernelSHAP*	FeatAbl*	LIME*
Insertion ↑	ResNet-50	**91.10**	83.400	75.235	64.658	68.772	64.544	70.187	76.638
	MobileNet-v2	**90.67**	74.651	34.197	47.848	48.662	60.837	59.197	61.282
	ViT-b16	**89.86**	89.594	54.455	68.125	69.480	75.152	65.656	76.161
	ResNet-18	**78.03**	73.558	47.799	38.877	36.806	50.547	43.448	50.592
Deletion ↓	ResNet-50	**8.99**	23.799	25.262	40.996	36.214	26.583	37.332	25.307
	MobileNetv2	6.61	**6.002**	26.381	14.679	13.382	7.766	8.866	7.344
	ViT-b16	**6.24**	17.298	40.784	30.948	29.903	21.825	34.191	19.254
	ResNet-18	**2.57**	6.596	8.588	11.273	11.555	6.638	8.352	6.776



Ablation study of concept insertion and deletion using MobileNet v2 on 100 randomly sampled ImageNet-1k images. Performance is measured by the area under the curve (AUC). S1 and S2 refer to segmentation algorithms SAM1 and SAM2, respectively, while R denotes the residual region. Results demonstrate improved performance when combining segmentation methods and concept insertion and deletion.

Kernel Sample Based Explanations (K-SBE)

	Execution time (s) ↓					Mean AUC DEL accuracy proponent ↑				
	K-SBE (ours)	TracIn	Gen-Rep NTK-final (tracing)	Influence Function	Random	K-SBE (ours)	TracIn	Gen-Rep NTK-final (tracing)	Influence Function	Random
CIFAR-10	**0.00926**	4144.23	104.72	124.25	*0.00079*	2.60 ± 5.00	**3.80 ± 5.22**	1.00 ± 3.22	-4.00 ± 5.26	-3.60 ± 5.15
CIFAR-100	**0.02838**	3944.29	398.54	234.49	*0.00035*	**3.80 ± 6.44**	2.40 ± 5.44	2.20 ± 3.47	3.00 ± 4.36	1.40 ± 4.90
FMNIST	**0.00805**	2794.05	394.94	265.05	*0.00046*	**2.60 ± 5.39**	-7.00 ± 5.30	-2.00 ± 2.33	-4.40 ± 3.84	-3.80 ± 3.89
MNIST	**0.00355**	2793.02	125.66	139.20	*0.00031*	**0.00 ± 0.79**	-2.60 ± 3.55	-0.60 ± 1.42	-1.40 ± 3.01	-0.60 ± 2.92



(a) True class: horse

(b) True class: six

(c) True class: airplane

(d) True class: five

Quantitative comparison of proposed K-SBE against: TracIn, Generalized Representer and Influence Function. We evaluate efficiency (execution time in seconds, lower is better) and explanation faithfulness (delta accuracy, higher is better). Random data attribution (italicized) provides the lowest possible execution time. K-SBE drastically outperforms other methods in execution time, approaching the theoretically lowest possible execution time, while being comparable to other state-of-the-art methods in the faithfulness of the explanations.

Open research questions

- How can compensatory mechanisms restore SHAP completeness and symmetry in non-exhaustive sampling?
- How can contextualized verbal and visual explanations be generated for multi-class image classification?
- How can custom kernels be designed and developed to optimize the performance of sample-based explanations?

Supervisor

- Asst. Prof. Dr. Senka Krivic, Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

References

[1] Hasić, Vahidin, Amar Halilović, and Senka Krivić. "Superpixel correlation for explainable image classification." World Conference on Explainable Artificial Intelligence. Cham: Springer Nature Switzerland, 2025.
[2] Hasić, Vahidin, and Senka Krivić. "Understanding Image Classification Prediction with Any Segment Explanation." International Conference on Artificial Neural Networks. Cham: Springer Nature Switzerland, 2025.
[3] Under review

EEML WORKSHOPS
Montenegrin Machine Learning Workshop
8 November 2025, Podgorica, Montenegro