Salvage: Shapley-distribution Approximation Learning Via Attribution Guided Exploration for Explainable Image Classification

Mehdi Naouar[1,2,*], Hanne Raum[1], Jens Rahnfeld[1], Yannick Vogt[1,2], Joschka Boedecker[1,2,3], Gabriel Kalweit[1,2], Maria Kalweit[1,2]

[1]University of Freiburg, [2]CRIION, [3]BrainLinks-BrainTools

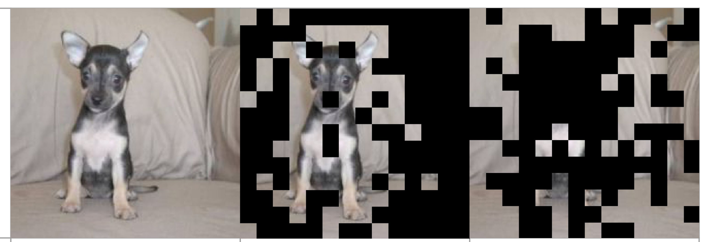
Motivation and Background

Removal-based Principle:

Masking portions of the input image to observe the resulting changes in the model's prediction.

Removal of important regions → Significant drop in the classification probability

Removal of irrelevant regions → Little effect on the models prediction



Input Image			
Classification probability (Chihuahua)	1.00	0.95	0.15

Shapley Values Estimation

Let N be a set of features and $v(S)$ the prediction outcome given a feature subset $S \subset N$. The Shapley value ϕ_i of a feature i is obtained as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \underbrace{\frac{|S|!\,(|N|-|S|-1)!}{|N|!}}_{w_S} (v(S \cup \{i\}) - v(S))$$

$$= \underbrace{\sum_{S \subseteq N \setminus \{i\}} w_S \cdot v(S \cup \{i\})}_{\phi_i^+} - \underbrace{\sum_{S \subseteq N \setminus \{i\}} w_S \cdot v(S)}_{\phi_i^-}$$

FastShap (Jethani et al., 2022) suggests a Least Square objective for the approximation of the Shapley value over the random mask distribution $p_w(S) \propto w_S$:

$$\mathbb{E}_{p_w(S)}\left[\left(v(S) - \sum_{i \in S} \phi_i\right)^2\right]$$

Issue: The Mean Square error is designed to approximate scalars, not probability distributions.

Method

Shapley Distributions

Our solution: the sum of the Shapley is mapped to a probability distribution using Softmax/Sigmoid (σ)

The resulting Shapley Distribution is given by: $u(S) = \sigma(\sum_{i \in S} \phi_i^+ + \sum_{i \notin S} \phi_i^-)$

The Shapley distribution is optimized by minimizing its Jensen–Shannon divergence to the target distribution $v(S)$:

$$\arg\min_{\phi^+, \phi^-} \mathbb{E}_{p_w(S)} \left[D_{JS}(u(S) \,||\, v(S)) \right]$$

Attribution Guided Sampling

Problem with random mask distribution: high unbalance between masks yielding high vs low prediction likelihoods

Our solution: Importance sampling during training. Two mask splits are sampled:

Split 1: proportional to ϕ (masks with high likelihood)   Split 2: proportional to $-\phi$ (masks with low likelihood)

Classification Aggregation

The attribution scores can be directly mapped to a classification prediction using the Shapley distribution of all image patches.

$$u(N) = \sigma(\textstyle\sum_{i \in N} \phi_i^+) \approx v(N)$$

Corresponding author*:
Mehdi Naouar
naouarm@cs.uni-freiburg.de

universität freiburg

CRIION COLLABORATIVE RESEARCH INSTITUTE INTELLIGENT ONCOLOGY

// BrainLinks BrainTools

Paper and References:

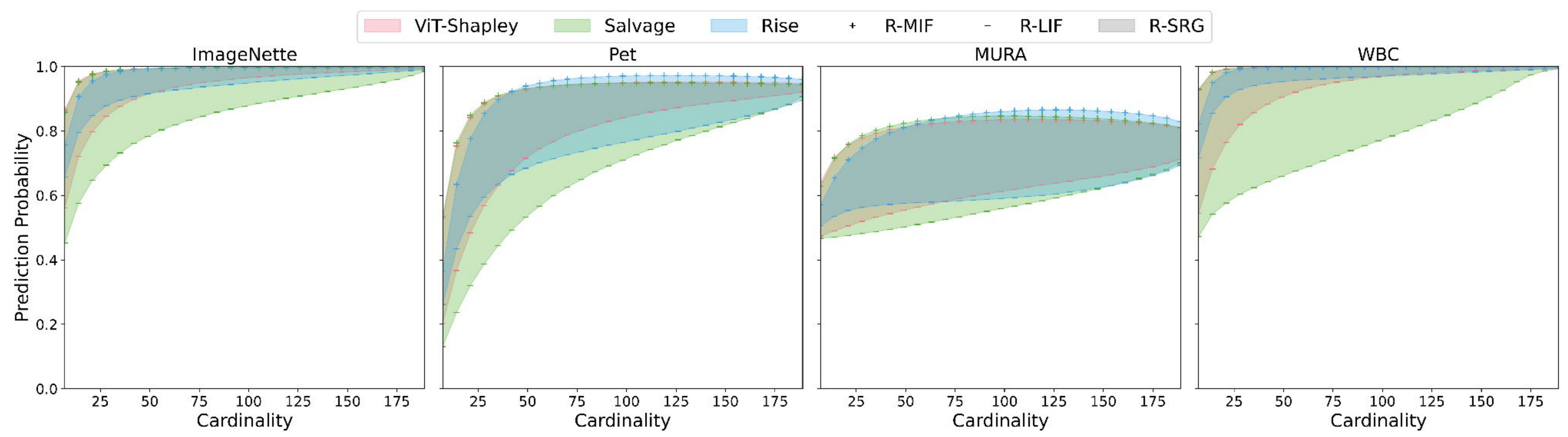
Quantitative Results

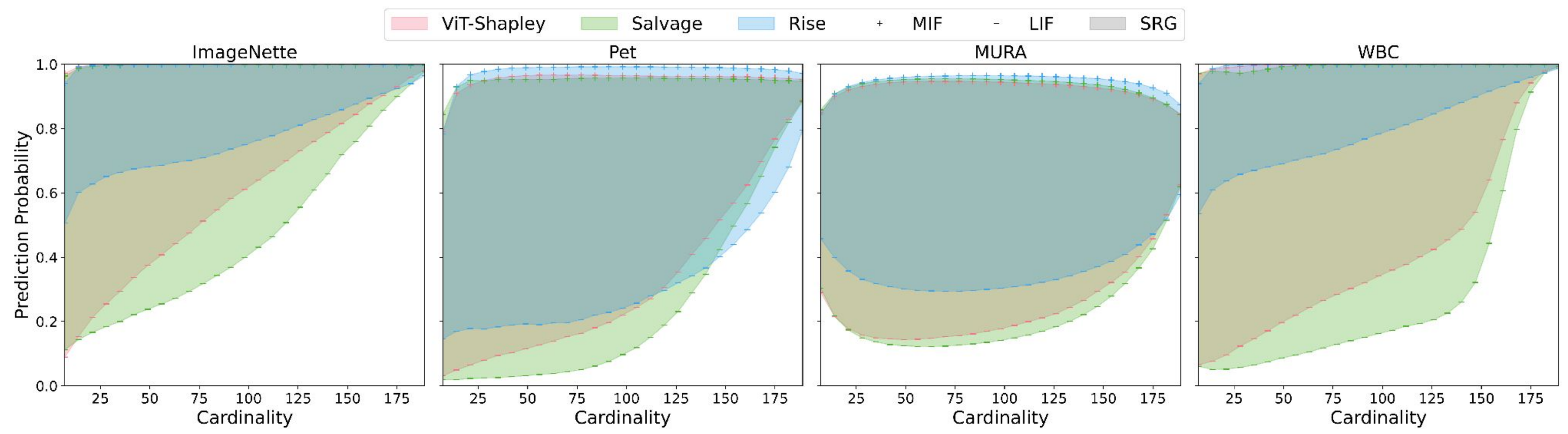
Metric Scores

Table 1: Quantitative results computed on the Pets, ImageNette, WBC, and MURA datasets. The performance of 10 baseline methods is measured in terms of SRG, R-SRG, RMA, and RRA.

	Pets				ImageNette		MURA		WBC	
Method	SRG	R-SRG	RMA	RRA	SRG	R-SRG	SRG	R-SRG	SRG	R-SRG
GradCam	10.6	3.5	48.1	42.7	-1.9	-3.3	16.2	10.1	-18.5	-20.2
EigenCam	27.4	3.2	48.9	62.9	13.2	-3.1	0.1	-4.5	22.9	-7.0
Attn. last	47.9	9.6	61.1	70.1	27.0	3.0	22.4	7.0	42.2	1.6
Attn. Roll.	52.0	11.2	51.5	74.6	32.0	3.4	17.6	6.3	48.1	2.5
ViT-CX	50.2	17.6	30.6	67.5	29.9	7.5	19.8	9.1	41.6	7.3
Sal. Maps	51.1	10.8	52.7	**76.3**	27.7	2.8	25.3	8.5	42.8	2.1
IntGrad	27.4	7.9	51.5	58.8	11.0	2.2	13.9	6.1	11.8	1.6
LRP	49.5	9.2	63.9	71.8	27.9	3.0	19.3	6.8	37.0	1.7
RISE	63.7	18.5	30.1	47.8	22.9	5.4	56.5	22.1	20.7	3.5
ViT-Shap	61.1	14.7	52.7	69.0	40.3	6.2	65.3	20.6	57.4	7.3
Salvage	**68.5**	**26.3**	**64.9**	73.5	**51.3**	**14.9**	**68.6**	**25.3**	**69.7**	**22.6**
Random	0.0	0.0	30.0	29.4	0.0	0.0	0.0	0.0	0.0	0.0

SRG and R-SRG Curves



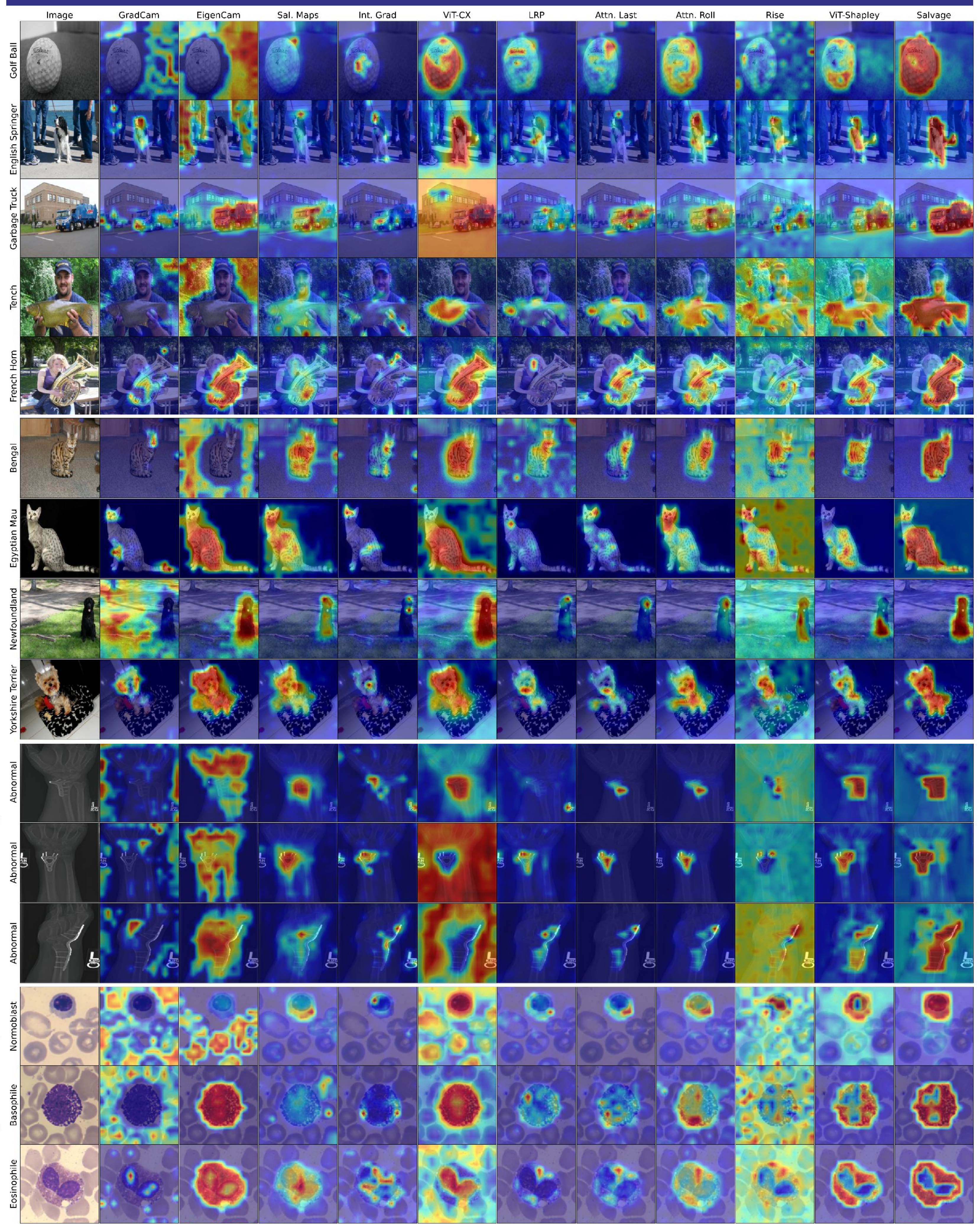


Classification Performance

Table 2: An overview of the classification performance of the original classifier, ViT-Shapley, and Salvage computed on Pet, ImageNette, WBC and MURA.

Model	Pets Accuracy	ImageNette Accuracy	WBC Accuracy	MURA Precision	MURA Recall	MURA F1-score	MURA MCC
Classifier	95.91%	99.64%	99.75%	84.64%	78.88%	81.66%	0.66
ViT-Shapley	0.00%	0.05%	0.69%	59.03%	92.74%	72.14%	0.39
Salvage	93.61%	98.88%	99.75%	80.31%	80.52%	80.41%	0.62

Qualitative Examples



Image

Predicted class: Church, English Springer, Garbage Truck, Garbage Truck

Ground-truth Class: Parachute, French Horn, Gas Pump, Chain Saw