One For All: A Unified Approach to Classification and Self-Explanation

Mehdi Naouar[1,2,*], Yannick Vogt[1,2], Joschka Boedecker[1,2,3], Gabriel Kalweit[1,2], and Maria Kalweit[1,2]

[1] Neurorobotics Lab, University of Freiburg, Freiburg, Germany
[2] Collaborative Research Institute Intelligent Oncology, Freiburg, Germany
[3] IMBIT//BrainLinks-BrainTools, University of Freiburg, Freiburg, Germany
* Corresponding author: naouarm@cs.uni-freiburg.de

Abstract. The integration of deep learning into medical vision applications has led to a growing demand for interpretable predictions. Typically, classification and explainability are treated as separate processes, with explainability methods applied post hoc to pre-trained classifiers. However, this decoupling introduces additional computational costs and may lead to explanations misaligned with the underlying model. In this paper, we propose One For All (OFA), an efficient, single-stage approach that jointly optimizes classification accuracy and self-explanation during training. OFA achieves this through a multi-objective framework, eliminating the need for separate explainability models while ensuring faithful and robust explanations. Extensive experiments on medical datasets confirm that OFA delivers competitive classification performance while providing high-quality, inherently interpretable explanations, making it a scalable and versatile solution for fully explainable classification.

Keywords: Explainability · XAI · Medical Image Classification

1 Introduction

The integration of artificial intelligence (AI) into medical imaging has significantly enhanced diagnostic accuracy, disease detection, and clinical decision support. Deep learning models now assist in identifying pathologies across various imaging modalities, offering substantial benefits in efficiency and diagnostic consistency. However, the widespread adoption of these models in healthcare remains constrained by a critical challenge: the lack of explainability. Trust in AI-driven medical decisions is essential for regulatory approval, clinician acceptance, and ultimately, patient safety.

To address this, explainability methods have been developed to elucidate how AI models generate predictions. Post-hoc approaches, which generate explanations after a model has made its prediction, often rely on perturbation-based surrogate models [8, 19] or sampling techniques [21]. While effective, this multi-step pipeline introduces substantial computational costs and potential misalignment

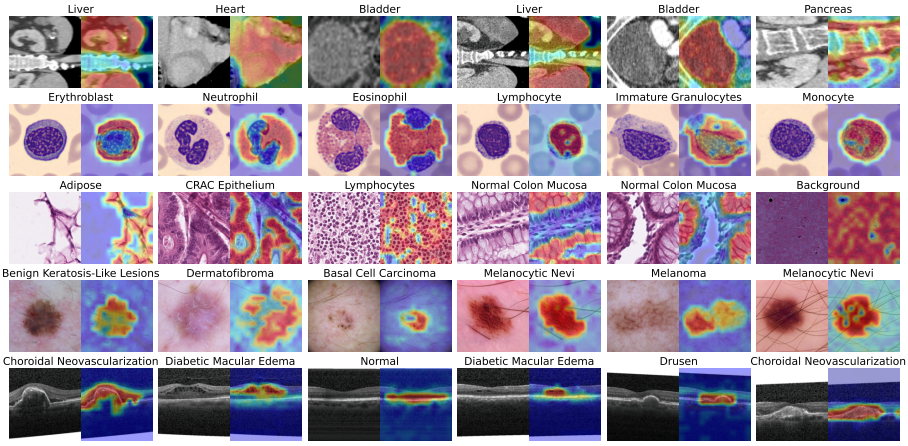

Fig. 1. Predictions and attribution scores of OFA-classifiers trained on the datasets OrganCMNIST+, BloodMNIST+, PathMNIST+, DermaMNIST+ and OctMNIST+.

between predictions and explanations, when generated by two decoupled models. In the context of medical imaging, where efficiency is crucial, these limitations hinder broader applicability. In contrast, on-the-fly explainability approaches integrate explanation generation directly into inference, enabling real-time attribution with minimal additional cost [1, 9, 24]. Attention mechanisms [1, 9] in transformer-based models [11] exemplify this approach by producing attribution scores inherently during prediction. While computationally efficient, the attribution estimates of these methods often lack in faithfulness, as they are not explicitly optimized for explainability [13, 15, 23].

In this work, we operate within the domain of on-the-fly explainability, and introduce One For All, a unified approach to image classification and self-explanation. Our method leverages principles from approximation-based removal techniques to generate faithful explanations while eliminating the need for a separate explainer model. Specifically, we employ a multi-objective training strategy to train a single model that simultaneously performs classification and generates consistent attribution scores. In a comprehensive evaluation, we conducted multiple studies across seven medical imaging datasets spanning different modalities from MedMNIST+ [28] assessing both predictive accuracy and explanation quality. Our results demonstrate that our pipeline maintains predictive accuracy comparable to standard classification optimization. Moreover, our method significantly outperforms both gradient-based and attention-based approaches in attribution quality, as evidenced by both quantitative and qualitative analyses. Overall, our approach provides a unified solution for explainable medical image classification, combining computational efficiency and high-quality explanations without compromising predictive accuracy. By aligning predictions and explanations within a single model, it enhances interpretability, simplifies deployment, and strengthens the trust and reliability essential for clinical decision-making.

2 Related work

The growing demand for explainable AI has led to the exploration of several on-the-fly attribution methods, each addressing different aspects of model interpretability. Class Activation Maps (CAM) were initially developed for Convolutional neural networks (CNNs) and focus on generating feature importance maps by weighting the activation maps of the final convolutional layer [29]. Grad-CAM and its variants [5, 10, 22] offer more flexibility but still face limitations when applied to transformer-based architectures [8]. Attention-based methods, which take advantage of the attention mechanism in transformer models, attempt to track the flow of information between input tokens and class tokens [1, 7]. While attention scores can offer some insights, they often focus on less informative regions, limiting their effectiveness [8, 9]. Gradient-based methods, such as Saliency Maps [24] and Integrated Gradients [27], compute gradients of the class scores relative to the input, with newer techniques like SmoothGrad [25] and Integrated Gradients offering more stable and refined results. Finally, Layer-Wise Relevance Propagation (LRP) methods [3, 18], which propagate model predictions back to the input to highlight relevant features, have proven effective for CNNs but encounter challenges with transformers due to their architecture [2, 6], prompting various adaptations to stabilize the explanations.

In contrast to these methods, OFA leverages principles of post-hoc removal-based approaches [8, 19] by training a segmentation model for classification alongside an auxiliary loss for self-explanations. However, unlike traditional post-hoc methods, our approach integrates classification and explanation within a single model, enabling both tasks to be performed in a single inference step.

3 Background

3.1 Shapley Values

Shapley values, originally introduced in cooperative game theory, provide a principled approach to fairly distributing an outcome among a set of contributors [8]. In the context of explainable AI, they quantify the contribution of each feature to a model's prediction. Given a feature set N and a function $v(S)$ representing the model's prediction for a subset $S \subset N$, the Shapley value ϕ_i of a feature i is computed as the weighted average of its marginal contributions across all possible subsets of features, expressed as:

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

An alternative perspective on Shapley values was introduced by [16], which avoids explicit marginal contribution terms. Instead, Shapley values are expressed as the difference between two weighted expectations—one when the

feature is present and one when it is absent:

$$\phi_i = \underbrace{\sum_{S \subseteq N \setminus \{i\}} w_S \cdot v(S \cup \{i\})}_{\phi_i^+} - \underbrace{\sum_{S \subseteq N \setminus \{i\}} w_S \cdot v(S)}_{\phi_i^-} \tag{2}$$

where ϕ_i^+ and ϕ_i^- correspond to expected model predictions conditioned on the inclusion and exclusion of feature i, respectively. This formulation enables more flexible optimization strategies for Shapley value estimation.

While both formulations consider all possible subsets of features, ensuring a theoretically sound measure of contribution, computing exact Shapley values is computationally infeasible for high-dimensional feature spaces due to the exponential growth in the number of subsets.

3.2 Approximate Shapley Value Estimation

To address the computational complexity of Shapley values, various approximation methods have been proposed. One such approach, FastShap [14], estimates Shapley values by solving a least-squares problem over a subset distribution $p_w(S) \propto w_S$, where samples are drawn proportionally to the weighting term w_S:

$$\mathbb{E}_{p_w(S)} \left[(v(S) - \sum_{i \in S} \phi_i)^2 \right] \tag{3}$$

Despite improving computational efficiency, using the Mean Squared Error (MSE) to approximate probabilistic model outputs introduces limitations. MSE is designed for scalar comparisons and does not inherently account for the probabilistic nature of model predictions. It fails to capture uncertainties, lacks scale invariance, and does not enforce probability constraints such as normalization, potentially leading to inaccurate or inconsistent attribution estimates.

To address this issue, a recent work [19] proposes an alternative approximation approach. Instead of directly minimizing MSE, feature subset values are mapped to a probability distribution, which is then optimized by minimizing its divergence from the true prediction distribution. Given a feature subset S, the sum of the positive Shapley values ϕ^+ of the included features and the sum of the negative Shapley values ϕ^- of the masked features are combined into a probability distribution, referred to as the Shapley distribution $u(S)$:

$$u(S) = \sigma \left(\sum_{i \in S} \phi_i^+ + \sum_{i \notin S} \phi_i^- \right) \tag{4}$$

where σ represents a softmax function for multiclass settings or a sigmoid function for binary classification. The Shapley distribution is optimized during training by minimizing the Jensen–Shannon (JS) divergence between the estimated distribution $u(S)$ and the target model's predicted distribution $v(S)$.

4 Method: One For All

To achieve effective and efficient explainability, OFA integrates the principles of Shapley distribution estimation [19] directly into the classification training framework. The input image is partitioned into n patches, which are processed to estimate the positive and negative Shapley values, ϕ^+ and ϕ^-, both of shape $n \times C$, where C denotes the number of classes. OFA is trained end-to-end with three loss terms that jointly optimize both classification and explanation.

4.1 Classification

OFA optimizes classification by maximizing the log-likelihood of the Shapley distribution over all image patches N with respect to the ground-truth class y, using a cross-entropy loss:

$$l_1 = CE\left(\sigma\left(\sum_{i\in N} \phi_i^+ - \phi_i^-\right), y\right) \tag{5}$$

4.2 Robustification to Masking

OFA's self-explanation mechanism builds on removal-based attribution methods, which state that masking important regions should strongly affect predictions, while masking irrelevant ones should not. However, this principle holds only if the model is inherently robust to masking—that is, if its predictions remain stable when irrelevant regions are removed. To enforce this robustness, we encourage consistency in masked-image predictions by maximizing the log-likelihood of the ground-truth class. This is achieved by uniformly sampling one mask for each image and minimizing the cross-entropy loss between the Shapley distribution of masked images S, defined in Equation (4), and the ground-truth class y:

$$l_2 = CE\left(\sigma\left(\sum_{i\in S} \phi_i'^+ + \sum_{i\notin S} \phi_i'^-\right), y\right) \tag{6}$$

4.3 Self-Explanation

To ensure that the model produces meaningful explanations, we introduce a consistency constraint on the Shapley distribution. The underlying idea is that the contribution of image patches to the prediction should remain stable across different perturbations. Specifically, when an image is masked, the Shapley values should be consistent with those computed for the unmasked image. To enforce this consistency, we sample n=32 masks from the informed mask distribution presented in [19] and minimize the Kullback-Leibler (KL) divergence between the Shapley distributions of the original and masked images:

$$l_3 = D_{KL}\left(\sigma\left(\sum_{i\in S} \phi_i^+ + \sum_{i\notin S} \phi_i^-\right) \middle\| \sigma\left(\sum_{i\in S} \phi_i'^+ + \sum_{i\notin S} \phi_i'^-\right)\right) \tag{7}$$

4.4 Overall Pipeline

One For All is trained following the three objectives—classification, robustness to masking, and self-explanation—into a unified training objective $L = l_1 + l_2 + l_3$. During inference, the (unmasked) input image is passed to OFA to compute the positive and negative shapley values (ϕ^+ and ϕ^-). These values are then used to derive the classification prediction through the Shapley distribution: $\sigma\left(\sum_{i \in N} \phi_i^+ - \phi_i^-\right) \in [0,1]^C$ where the contribution of each patch i to each class is inherently captured by its Shapley value $\phi_i = \phi_i^+ - \phi_i^- \in C$.

5 Experiments

5.1 Experimental setup

To evaluate OFA, we conduct experiments on a diverse set of medical imaging datasets, covering a range of classification tasks: PathMNIST+, DermaMNIST+, OCTMNIST+, PneumoniaMNIST+, RetinaMNIST+, BreastMNIST+, BloodMNIST+, OrganCMNIST+, and OrganSMNIST+ [28]. For each dataset, models are trained for approximately 10,000 iterations, with a maximum of 100 epochs. Both OFA and Segmenter adapt the Segmenter architecture from [26], utilizing a ViT-B backbone and a single transformer block as decoder. All Vision Transformer [11] classifiers and backbone models are based on the ViT-B configuration, with an image size of 224 and a patch size of 16 (a total of 196 patches), and utilizing registers as described in [9]. These models are initialized with the pre-trained weights from DINOv2 [20]. Training is conducted with a batch size of 64, using the AdamW optimizer [17] with a learning rate of $1e-5$ and a weight decay of $1e-5$.

5.2 Classification performance

We evaluate the classification performance of OFA by comparing it to ResNet-18 [12], ResNet-50 [12], and a ViT-B [11] classifier. Additionally, we compare the results of our method to a baseline model (Segm-B) that adopts the same neural architecture [26] but is trained solely using the cross-entropy classification objective Equation (5). This comparison serves to assess the impact of the auxiliary loss terms—specifically, the robustness to masking and the self-explanation objectives—on the classification performance of the model. We adopt the masking strategy from Dinov2 using a learned mask embedding, which is assigned to all masked patches.

The results in Table 1 highlight three main points. First, ResNet architectures are outperformed by all transformer models across most datasets, particularly in AUC and accuracy, demonstrating the advantage of transformer-based models in capturing global context. Second, using a segmentation model for classification does not seem to affect performance when using (nearly) the same architecture, as ViT-B and Segm-B show similar performance on average. Finally, OFA-B performs similarly to Segm-B, indicating that the auxiliary losses, specifically

Table 1. Comparison of classification performance across different models (ResNet-18, ResNet-50, ViT-B, Segm-B, and OFA-B) on various datasets from MedMNIST+. The table reports AUC and accuracy for each model.

	Path		Derma		OCT		Pneumonia		Retina	
Model	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-18	98.9	90.9	92.0	75.4	95.8	76.3	95.6	86.4	71.0	49.3
ResNet-50	98.9	89.2	91.2	73.1	95.8	77.6	96.2	88.4	71.6	51.1
ViT-B	99.7	**97.5**	98.2	90.9	99.7	**93.0**	**98.8**	91.0	85.9	64.0
Segm-B	**99.8**	96.6	98.5	90.4	99.6	92.8	92.9	**91.7**	86.5	66.3
OFA-B	99.6	97.0	**99.0**	**91.6**	**99.8**	92.7	96.6	91.3	**88.7**	**68.3**

	Breast		Blood		OrganC		OrganS	
Model	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-18	89.1	83.3	99.8	96.3	**99.4**	**92.0**	97.4	77.8
ResNet-50	86.6	84.2	99.7	95.0	99.3	91.1	97.5	78.5
ViT-B	91.9	89.7	**100.0**	**99.2**	98.2	85.8	97.4	83.4
Segm-B	92.5	**91.0**	99.9	**99.2**	98.6	85.7	98.1	**83.8**
OFA-B	**93.8**	88.5	**100.0**	**99.2**	98.7	84.7	**98.3**	83.2

robustness to masking and self-explanation, do not compromise classification performance.

5.3 Explanation Evaluation

We perform both quantitative and qualitative analyses across nine different datasets, comparing our method with two baselines: Vanilla gradient (Grad) [24] and the attention scores from the last transformer block (Attn) [1]. Quantitative performance is evaluated using two metrics: SRG (Symmetric Relevance Gain) [4] and R-SRG (Relatively Symmetric Relevance Gain) [19].

SRG is a ranking-based metric that assesses attribution methods by measuring the change in model performance when features are ranked by importance and either added or removed. It focuses on the consistency of feature rankings and their impact on model predictions. R-SRG extends SRG by incorporating weighted sampling based on feature importance. Instead of merely ranking features, R-SRG evaluates the relative importance of each feature through sampling, offering a more accurate assessment of attributions by accounting for the magnitude of feature contributions.

As shown in Table 2, OFA consistently outperforms both gradient- and attention-based methods across all datasets. This is evident from the significantly higher SRG and R-SRG scores achieved by OFA, demonstrating its superior ability to generate accurate and reliable attributions. Additionally, we present qualitative results in Figure 2 in the form of attribution maps corresponding to the predicted class. These results further highlight the effectiveness of OFA in producing clearer and more interpretable explanations for the predicted class.

Table 2. SRG and R-SRG scores for the proposed OFA method compared to gradient-based (Grad) and attention-based (Attn) methods across nine different datasets.

	Path		Derma		OCT		Pneumonia		Retina	
Method	SRG	R-SRG	SRG	R-SRG	SRG	R-SRG	SRG	R-SRG	SRG	R-SRG
Attn	-4.95	-0.88	-8.83	-3.98	-18.35	-8.31	-1.31	-0.01	2.78	1.51
Grad	-1.69	-1.02	12.16	3.05	43.17	10.50	1.01	0.30	18.28	5.96
OFA	**35.26**	**12.47**	**22.91**	**10.43**	**55.32**	**23.95**	**36.22**	**4.19**	**20.28**	**10.98**

	Breast		Blood		OrganC		OrganS	
Method	SRG	R-SRG	SRG	R-SRG	SRG	R-SRG	SRG	R-SRG
Attn	-0.04	0.06	-7.36	-4.63	1.62	0.69	3.23	0.45
Grad	19.06	2.14	36.45	3.41	5.55	0.30	9.22	1.13
OFA	**22.15**	**7.11**	**65.35**	**28.80**	**29.42**	**11.05**	**32.37**	**13.95**

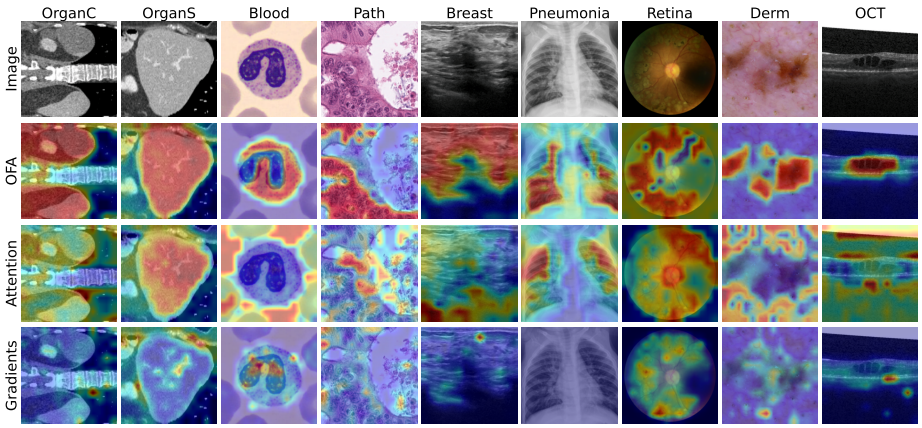

Fig. 2. Attribution maps for the predicted class generated by the proposed OFA method, compared to gradient-based and attention-based methods.

6 Conclusion

In this work, we presented One For All, a novel and effective approach to integrating explainability directly into the image classification process. By combining classification and self-explanation into a single model with multi-objective training, we eliminate the need for a post-hoc computation, significantly reducing computational costs without sacrificing predictive performance. Our extensive evaluation across diverse medical imaging datasets demonstrates that One For All not only maintains high classification accuracy but also delivers more faithful and reliable explanations compared to existing methods. This unified approach has the potential to improve the interpretability and trustworthiness of AI systems in medical contexts, ultimately advancing the role of explainable AI in clinical decision-making.

Acknowledgments. This project was funded by the Mertelsmann Foundation. This work is part of BrainLinks-BrainTools which is funded by the Federal Ministry of Economics, Science and Arts of Baden-Württemberg within the sustainability program for projects of the excellence initiative II.

Disclosure of Interests. The authors declare no competing interests.

References

1. Abnar, S., Zuidema, W.H.: Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020. pp. 4190–4197. Association for Computational Linguistics (2020)
2. Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K., Wolf, L.: XAI for transformers: Better explanations through conservative propagation. In: International Conference on Machine Learning, ICML 2022. Proceedings of Machine Learning Research, vol. 162, pp. 435–451. PMLR (2022)
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**(7), 1–46 (07 2015)
4. Bluecher, S., Vielhaben, J., Strodthoff, N.: Decoupling pixel flipping and occlusion strategy for consistent XAI benchmarks. Trans. Mach. Learn. Res. **2024** (2024)
5. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018. pp. 839–847. IEEE Computer Society (2018)
6. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. pp. 782–791. Computer Vision Foundation / IEEE (2021)
7. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does BERT look at? an analysis of bert's attention. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019. pp. 276–286. Association for Computational Linguistics (2019)
8. Covert, I.C., Kim, C., Lee, S.: Learning to estimate shapley values with vision transformers. In: 11th International Conference on Learning Representations, ICLR 2023 (2023)
9. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: 12th International Conference on Learning Representations, ICLR 2024 (2024)
10. Desai, S., Ramaswamy, H.G.: Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In: IEEE Winter Conference on Applications of Computer Vision, WACV 2020. pp. 972–980. IEEE (2020)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021 (2021)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016. pp. 770–778. IEEE Computer Society (2016)

13. Jain, S., Wallace, B.C.: Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers). pp. 3543–3556. Association for Computational Linguistics (2019)
14. Jethani, N., Sudarshan, M., Covert, I.C., Lee, S., Ranganath, R.: Fastshap: Real-time shapley value estimation. In: 10th International Conference on Learning Representations, ICLR 2022 (2022)
15. Kindermans, P., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (un)reliability of saliency methods. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Lecture Notes in Computer Science, vol. 11700, pp. 267–280. Springer (2019)
16. Kolpaczki, P., Bengs, V., Muschalik, M., Hüllermeier, E.: Approximating the shapley value without marginal contributions. In: 38th AAAI Conference on Artificial Intelligence, AAAI 2024. pp. 13246–13255. AAAI Press (2024)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019 (2019)
18. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognit. **65**, 211–222 (2017)
19. Naouar, M., Raum, H., Rahnfeld, J., Vogt, Y., Boedecker, J., Kalweit, G., Kalweit, M.: Salvage: Shapley-distribution approximation learning via attribution guided exploration for explainable image classification. In: 13th International Conference on Learning Representations, ICLR 2025 (2025)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. **2024** (2024)
21. Petsiuk, V., Das, A., Saenko, K.: RISE: randomized input sampling for explanation of black-box models. In: British Machine Vision Conference, BMVC 2018. p. 151. BMVA Press (2018)
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision, ICCV 2017. pp. 618–626. IEEE Computer Society (2017)
23. Serrano, S., Smith, N.A.: Is attention interpretable? In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers. pp. 2931–2951. Association for Computational Linguistics (2019)
24. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: 2nd International Conference on Learning Representations, ICLR 2014 Workshop Track Proceedings (2014)
25. Smilkov, D., Thorat, N., Kim, B., Viégas, F.B., Wattenberg, M.: Smoothgrad: removing noise by adding noise (2017)
26. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: IEEE/CVF International Conference on Computer Vision, ICCV 2021. pp. 7242–7252. IEEE (2021)

27. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR (2017)
28. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. CoRR **abs/2110.14795** (2021)
29. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016. pp. 2921–2929. IEEE Computer Society (2016)