Suppressive Dropout: A Demonstration of Explainable Channel-Selective Regularization for Preserving Rare Features*

Seunghyun Lee[1,*]

[1]Department of Industrial Management Engineering (Double Major in Mathematics), Korea University, Seoul, Republic of Korea

Abstract

Deep learning models frequently overfit to dominant features, suppressing the learning of rare but informative patterns—a phenomenon analogous to the neurophysiological mechanism of lateral inhibition. Standard dropout discards channels at random, ignoring inter-channel suppressive interactions. We present Suppressive Dropout (SDrop), a biologically inspired, explainable regularization system that selectively removes channels identified as over-inhibitory to their neighbors. The system computes an EGPG score (Energy–Global Peakedness–Grid) through a reinterpretation of Local Response Normalization (LRN) to rank channels by their suppressive influence, and applies SGridLC, a spatially-aware grid-based local control mechanism, to regulate feature density at the sub-image level. Every dropout decision is transparent and visualizable, making SDrop a direct contribution to eXplainable AI (XAI): practitioners can inspect which channels are dropped, why they are identified as harmful suppressors, and where in the spatial domain the suppression occurs. Experiments on CIFAR-100 and TinyImageNet demonstrate accuracy gains of up to +0.77%p over the baseline with improved stability over standard dropout.

Keywords

Suppressive Dropout, Lateral Inhibition, Explainable AI, Channel Regularization, Rare Feature Preservation, LRP Visualization

1. Introduction

Despite rapid advances in deep learning, a persistent failure mode remains: models tend to latch onto *dominant features*—strong textures, backgrounds, or high-energy activations—while neglecting subtle but informative *rare features* [1, 2]. This imbalance degrades out-of-distribution generalization and makes models harder to interpret.

This phenomenon mirrors the neurophysiological mechanism of *lateral inhibition* [3]: strongly activated neurons suppress the activity of their neighbors, sharpening local contrast but potentially silencing weaker, informative signals (Figure 1a). In artificial networks, Local Response Normalization (LRN) [4] formalizes this principle mathematically. When inhibition is excessive, certain channels monopolize the feature map, crowding out minority-class evidence.

A conceptual inversion: targeting the strongest, not the weakest. All prior regularization methods share one implicit assumption: *high activation is a proxy for importance*. Standard dropout [5] ignores activation magnitude entirely, removing channels at random. Guided and Targeted Dropout [6] make the assumption explicit—they prune *weak*, low-magnitude channels to induce sparsity and preserve the "important" high-activation ones. Channel attention mechanisms (SE-Net [7], CBAM [8]) go further by actively *amplifying* dominant channels as a learned importance weight. DropBlock [9] and SpatialDropout [10] add spatial structure but inherit the same retention bias toward high-activation channels. More recently, explainability scores have been used to guide regularization: Sevillano-

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Demo submission to XAI 2026. The system is implemented and tested on CIFAR-100 and TinyImageNet.

*Corresponding author.

sh200411@korea.ac.kr (S. Lee)

https://orcid.org/0009-0006-1926-653X (S. Lee)

0009-0006-1926-653X (S. Lee)

García et al. [11] propose penalizing features identified as low-impact by attribution methods, while Saadallah [12] uses SHAP values to regularize model parameters. Both approaches, however, retain the implicit assumption that low-attribution channels should be suppressed—leaving dominant, over-inhibitory channels untouched. *SDrop inverts this logic entirely.* Motivated by LRP attribution analysis (detailed in Section 2) and lateral inhibition theory, we argue that the most dominant, high-energy channels are precisely the *harmful* ones: they suppress weaker but informative rare-feature channels through a lateral inhibition mechanism, starving minority-class detectors of gradient signal. Dropping the strongest suppressor channels forces the network to activate its previously silenced feature detectors. SDrop is, to our knowledge, the *first dropout method to deliberately target the most activated channels*—a direct inversion of every prior guided dropout strategy—grounded in the EGPG suppression score derived from a reinterpretation of LRN. We introduce Suppressive Dropout (SDrop): rather than random removal, SDrop preferentially drops channels whose inhibitory influence is both *strong* (high energy) and *globally distributed* (low spatial peakedness). SGridLC extends this with spatially-resolved grid control (Figure 1b).

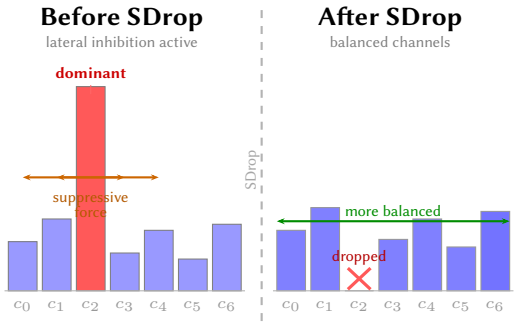
2. Related Work

Dropout variants. Standard dropout [5] deactivates each unit independently with probability p. SpatialDropout [10] drops entire feature channels to decorrelate spatially correlated activations. DropBlock [9] removes contiguous spatial blocks, improving regularization in convolutional networks. Targeted Dropout [6] prunes low-magnitude weights to induce structured sparsity. Sevillano-García et al. [11] propose a regularization term that penalizes low-impact features using attribution scores, and Pochinkov et al. [13] show that ablating neurons with peak activations in attention heads reveals dominant units that monopolize representational capacity. All these methods either operate randomly or focus on *weak, low-activation* units. SDrop inverts this strategy: it preferentially removes the *most dominant, energetic* channels—those identified as over-inhibitory by the EGPG score.

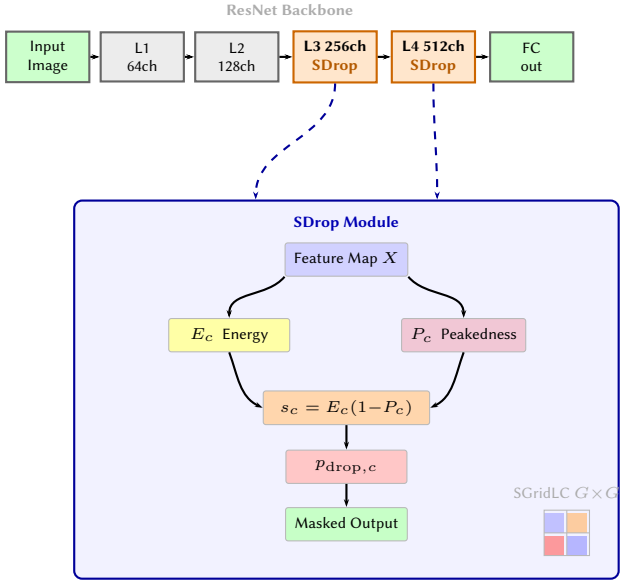
Channel attention and feature recalibration. SE-Net [7] and CBAM [8] learn to recalibrate channel responses by modeling inter-channel dependencies, typically *amplifying* informative (dominant) channels. This reinforces the very channels SDrop suppresses. SDrop can be seen as a complementary training-phase intervention: by preventing dominant channels from monopolizing the feature space, it allows post-hoc attention mechanisms to converge on more diverse and interpretable features.

LRP and explainability-motivated design. LRP [14, 15] and GradCAM [16] are post-hoc attribution methods that reveal which input regions and feature channels drive a model's prediction. Santos and Santos [17] demonstrate LRP's ability to trace per-channel relevance in complex neural networks, while Pe et al. [18] extend gradient-based localization beyond standard GradCAM to richer signal-level explanations. Shahid et al. [19] further show that higher-order partial derivatives reveal non-trivial feature interactions in neural networks that first-order attribution methods miss—interactions of precisely the kind SDrop addresses at the channel level through the EGPG score. LRP was the *direct empirical motivation* for SDrop: applied channel-wise, LRP backpropagates prediction scores to per-channel attributions and consistently reveals that relevance mass concentrates on a small subset of high-energy channels, while sparse, informative channels receive negligible attributed relevance [20, 21]. This concentration pattern—a direct LRN-type signature visible in explanation maps—confirms that dominant channels are not merely co-adapting; they are actively suppressing minority-class evidence through lateral inhibition [3]. If those channels are identified and removed during training, the model is forced to develop richer, more diverse feature representations. SDrop operationalizes this LRP-revealed suppression structure as a dropout criterion via the EGPG score, making every channel removal decision traceable to a quantified suppression score and enabling the XAI analyses in Section 4.

3. System Description



(a) Lateral inhibition concept. Left: dominant channel c_2 (red) suppresses neighbors. Right: after SDrop drops c_2, remaining channels are more balanced.



(b) SDrop system architecture. The ResNet backbone applies SDrop at L3 and L4 (orange). Inside the SDrop module, Energy E_c and Peakedness P_c are computed *independently* from the feature map X (branching arrows), then combined as the EGPG suppression score $s_c = E_c(1 - P_c)$. High-s_c channels (high energy, diffuse spread) receive the highest drop probability. SGridLC (lower-right) applies independent masking per cell of a $G \times G$ spatial grid.

Figure 1: SDrop overview: (a) lateral inhibition motivation and (b) system architecture.

3.1. Mathematical Foundation

Traditional LRN normalizes each channel activation x_k by its neighbors' energy:

$$y_k = x_k \cdot \left(1 + \alpha \sum_{j \in N(k)} x_j^2\right)^{-\beta} \tag{1}$$

To quantify how strongly a neighbor x_j suppresses the target activation x_k, we compute:

$$\frac{\partial y_k}{\partial x_j^2} = -\frac{\alpha\beta x_k}{\left(1 + \alpha \sum_{i \in N(k)} x_i^2\right)^{\beta+1}} \tag{2}$$

The magnitude of this negative quantity represents the *suppressive force* exerted by channel j on channel k.

3.2. EGPG Score: Identifying Over-Inhibitory Channels

Channel Energy E_c captures suppression potential:

$$E_c = 1 - \left(1 + \gamma \cdot \tfrac{1}{HW} \sum_{h,w} x_{c,h,w}^2\right)^{-\delta} \tag{3}$$

Spatial Peakedness P_c measures activation concentration:

$$P_c = \frac{\max_{h,w} |x_{c,h,w}|}{\sum_{h,w} |x_{c,h,w}| + \epsilon} \tag{4}$$

High peakedness indicates a localized, informative activation; low peakedness indicates diffuse, redundant coverage. The *EGPG suppression score* and resulting drop probability are:

$$s_c = E_c \cdot (1 - P_c), \qquad p_{\text{drop},c} = p_{\text{base}} \cdot \frac{s_c}{\max_c s_c} \tag{5}$$

3.3. SGridLC: Spatially-Aware Local Feature Density Control

SGridLC partitions the feature map $X \in \mathbb{R}^{C \times H \times W}$ into a $G \times G$ spatial grid. For each cell (i, j): local energy $E_c^{(i,j)}$ and peakedness $P_c^{(i,j)}$ are computed within the cell; a local score $s_c^{(i,j)}$ drives an independent drop mask per cell. This separates object and background treatment (Figure 2).

3.4. XAI Integration

Every SDrop decision is interpretable:

- *Score maps*: EGPG heatmaps show which channels are over-suppressive (Figure 2).
- *Spatial drop masks*: SGridLC overlays reveal *where* regularization is most active.
- *LRP relevance*: Before/after LRP heatmaps show relevance redistribution toward rare-feature regions [14].

4. Demonstration

Previous editions of this conference have featured interactive XAI demonstration systems: Jullum et al. [22] presented eXplego, a tool for selecting appropriate XAI methods for a given task; Schürmann and Sager-Müller [23] demonstrated a real-time XAI dashboard for semantic segmentation; and Werner et al. [24] showed ClustXRAI for interactive exploration of process mining explanations. These systems are designed as *post-hoc* explanation tools applied to fixed, already-trained models. SDrop's demo is qualitatively different: the explainability mechanism (the EGPG score) is *integral to the training process itself*, and the demo exposes in real-time how the suppression scores drive each channel dropout decision. Users observe not only what the model learns, but why specific channels are removed during training.

The demo runs on a laptop with a pre-loaded interactive interface (Python/PyTorch + Streamlit). Users can interact with the following features:

1. *Image selection*: Upload or select from CIFAR-100 / TinyImageNet.
2. *EGPG score inspection*: Per-channel bar charts and spatial heatmaps show which channels are over-inhibitory.
3. *Activation comparison*: Toggle between baseline and SDrop activations in real-time to observe feature representation changes.
4. *Hyperparameter control*: Adjust p_{base}, grid size $G \in \{2, 4\}$, and insertion layer (L3 or L4) with immediate visual feedback.
5. *LRP heatmaps*: Compare relevance maps before/after SDrop, demonstrating redistribution away from dominant channels.
6. *SGridLC overlay*: View spatial drop masks overlaid on the input, showing targeted regularization relative to object boundaries.

5. Experimental Evaluation

5.1. Setup

Experiments used CIFAR-100 (32×32, 50 000 / 10 000 images) with modified ResNet-18 [25], and TinyImageNet (64×64, 100 000 / 10 000 images) with ResNet-50 [25]. SDrop was inserted at Layer 3 and Layer 4. Training: SGD (momentum $= 0.9$, weight decay $= 5 \times 10^{-4}$), cosine annealing from lr $= 0.1$, batch size 128, standard augmentation. Results are mean $\pm$ std over 3 runs.

5.2. Results

On CIFAR-100, SDrop_Energy (Rate 0.1, L4) achieves 77.42%, surpassing the baseline (+0.77%p) and standard dropout at the same setting (+0.64%p) with low std of 0.16. Standard dropout at Rate 0.3 degrades below baseline (-0.24%p), whereas SDrop at L4 consistently improves over it. SGridLC (Rate 0.1, L3, $G = 2$) also gains +0.11%p. On TinyImageNet, SGridLC (Rate 0.3, L3, $G = 4$) achieves 48.66% vs. baseline 48.62%, while standard dropout at L4 drops to 48.34% (-0.28%p). Notably, SGridLC at Rate 0.1, $G = 4$, L3 produces severe instability (32.34% $\pm$ 27.58): fine-grained grids with low drop rates can disrupt spatial continuity in early training, a diagnostically valuable finding visible through score logging.

Tables 1 and 2 present the full experimental results. The complete sweep across rates, layers, and grid sizes confirms the robustness of the conceptual inversion: across nearly all configurations, SDrop variants maintain or exceed baseline accuracy while standard dropout at higher rates consistently degrades.

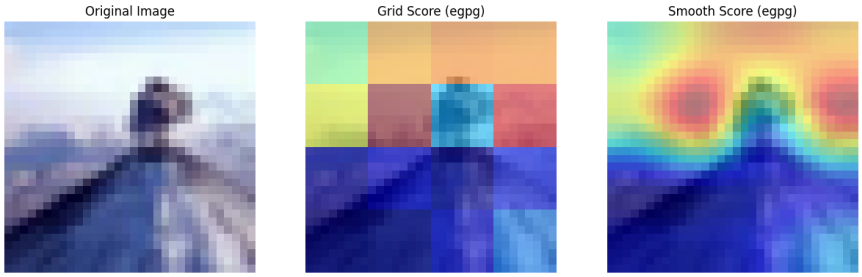
Table 1
CIFAR-100 results (Test Acc. %, mean $\pm$ std, 3 runs).

Method	Rate	Layer	Grid	Acc. (%)
Baseline	0.0	–	–	76.65±0.11
Dropout	0.1	L3	–	76.56±0.06
Dropout	0.1	L4	–	76.78±0.11
Dropout	0.3	L3	–	76.53±0.39
Dropout	0.3	L4	–	76.41±0.09
SDrop	0.1	L3	–	76.41±0.14
SDrop	0.1	L4	–	76.91±0.13
SDrop	0.3	L3	–	76.07±0.16
SDrop	0.3	L4	–	76.23±0.36
SDrop_Energy	0.1	L3	–	76.23±0.40
SDrop_Energy	**0.1**	**L4**	**–**	**77.42±0.16**
SDrop_Energy	0.3	L3	–	76.05±0.19
SDrop_Energy	0.3	L4	–	76.20±0.17
SGridLC	0.1	L3	2	76.76±0.20
SGridLC	0.1	L3	4	76.76±0.20
SGridLC	0.1	L4	2	76.29±0.20
SGridLC	0.1	L4	4	76.29±0.20
SGridLC	0.3	L3	2	76.19±0.32
SGridLC	0.3	L3	4	76.19±0.32
SGridLC	0.3	L4	2	76.52±0.08
SGridLC	0.3	L4	4	76.52±0.08

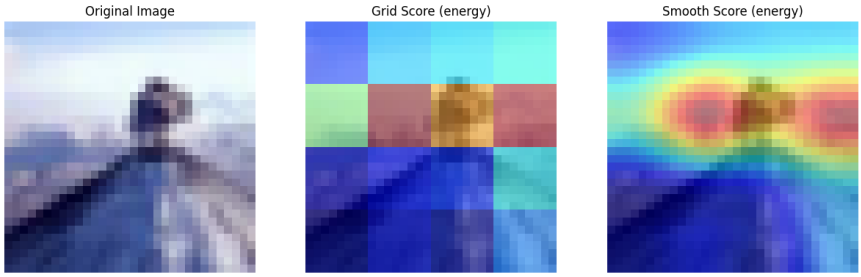
Table 2
TinyImageNet results (Val Acc. %, mean $\pm$ std, 3 runs).

Method	Rate	Layer	Grid	Acc. (%)
Baseline	0.0	–	–	48.62±0.03
Dropout	0.3	L3	–	48.55±0.25
Dropout	0.3	L4	–	48.34±0.06
SDrop	0.1	L3	–	48.60±0.18
SDrop	0.1	L4	–	48.28±0.39
SDrop	0.3	L3	–	48.55±0.22
SDrop	0.3	L4	–	48.37±0.27
SDrop_Energy	0.1	L3	–	48.54±0.21
SDrop_Energy	0.1	L4	–	48.28±0.32
SDrop_Energy	0.3	L3	–	48.18±0.19
SDrop_Energy	0.3	L4	–	48.13±0.42
SGridLC	0.1	L3	2	48.51±0.22
SGridLC	0.1	L3	4	32.34±27.58
SGridLC	0.1	L4	2	48.41±0.06
SGridLC	0.3	L3	2	48.53±0.15
SGridLC	**0.3**	**L3**	**4**	**48.66±0.42**
SGridLC	0.3	L4	2	48.28±0.15

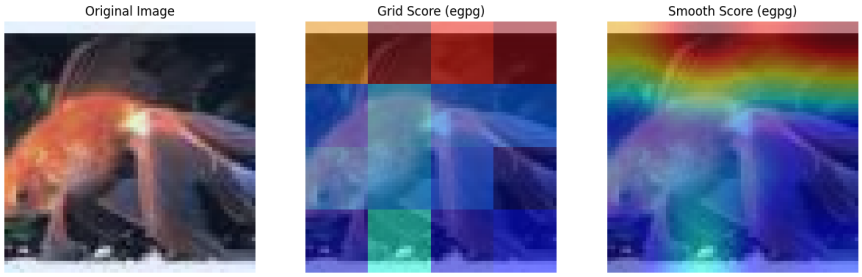
5.3. Feature Map Visualizations



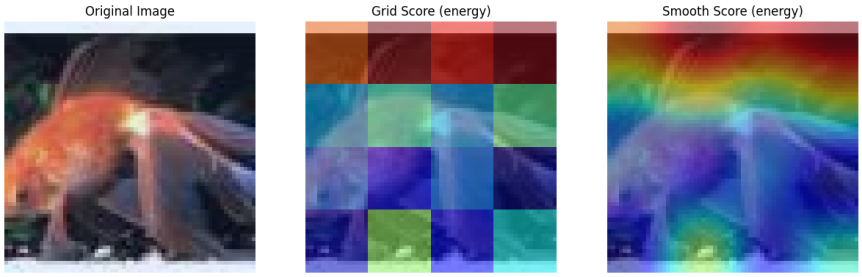
(a) CIFAR-100 L3, Grid 4 — EGPG score. Warm regions (object boundary) mark over-suppressive channels.



(b) CIFAR-100 L4, Grid 4 — Energy score. A single dominant centre-right region drives the best SDrop_Energy (L4) result (+0.77%p).



(c) TinyImageNet L3, Grid 4 — EGPG score. Suppressive channels concentrate in the background (upper-left); the foreground goldfish shows low suppression.



(d) TinyImageNet L3, Grid 4 — Energy score. Broader background suppression; SGridLC (Rate 0.3, $G = 4$) achieves 48.66% ± 0.42.

Figure 2: EGPG and Energy score visualizations. Each panel: *Original Image* | *Grid Score* (per-cell s_c heatmap) | *Smooth Score* (Gaussian-smoothed map). Warm colours = over-inhibitory channels (SDrop targets); cool colours = low-suppression rare-feature channels (preserved). Top two panels: CIFAR-100. Bottom two: TinyImageNet, where the 4×4 SGridLC grid separates foreground from background.

6. Conclusion

All prior dropout and regularization methods share a common implicit assumption: *high activation signals importance*, so either regularize randomly or prune weak, low-magnitude units. SDrop challenges this assumption at its root.

LRP attribution analysis (Section 2) and lateral inhibition theory jointly reveal that dominant high-energy channels actively *suppress* rare-feature activations across neighbors—their strength is a symptom

of harmful monopolization, not of genuine informativeness. SDrop operationalizes this insight through the EGPG score $s_c = E_c \cdot (1 - P_c)$: channels with high energy *and* diffuse spatial spread are identified as over-inhibitory suppressors and preferentially removed. This is precisely the opposite of what Targeted Dropout [6] and similar methods do—they would *keep* those channels as the most important ones. This *conceptual inversion—dropping the strongest channels rather than the weakest—is the central contribution of this work.* It is not a heuristic; it is grounded in a mechanistic account of how dominant channels harm minority-class learning, making every dropout decision traceable, visualizable, and explainable. The experimental results confirm the significance of the inversion: SDrop_Energy achieves $+0.77$%p over the baseline on CIFAR-100 while standard dropout at higher rates *degrades* performance (-0.24%p), and SGridLC achieves the best TinyImageNet result while spatially distinguishing foreground from background suppression. SDrop is the first method where the *reason* for dropping a channel—its suppressive dominance over neighbors—is quantified, visualized, and directly tied to the accuracy gain, bridging regularization and explainability in a unified framework.

Future work will extend SDrop to Vision Transformer attention maps and validate on medical imaging benchmarks where rare-feature preservation carries clinical significance.

Acknowledgments

This work was supported by Korea University.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization, in: Proceedings of the International Conference on Learning Representations (ICLR), 2017, pp. 1–15.

[2] A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images, Technical Report, University of Toronto, 2009.

[3] H. K. Hartline, F. Ratliff, Inhibitory interaction of receptor units in the eye of limulus, Journal of General Physiology 40 (1957) 357–376.

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NeurIPS), volume 25, 2012, pp. 1097–1105.

[5] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.

[6] A. N. Gomez, I. Zhang, K. Swersky, Y. Gal, G. E. Hinton, Learning sparse networks using targeted dropout, in: Workshop on Science of Deep Learning at ICLR, 2019.

[7] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.

[8] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[9] G. Ghiasi, T.-Y. Lin, Q. V. Le, DropBlock: A regularization method for convolutional networks, in: Advances in Neural Information Processing Systems (NeurIPS), volume 31, 2018, pp. 10727–10737.

[10] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 648–656.

[11] I. Sevillano-García, J. Luengo, F. Herrera, Low-impact feature reduction regularization term: How to improve artificial intelligence with explainability, in: Late-breaking Work, Demos and Doctoral Consortium, co-located with the 1st World Conference on eXplainable Artificial Intelligence (xAI-2023), volume 3554 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

[12] A. Saadallah, SHAP-guided regularization in machine learning models, in: Late-breaking Work, Demos and Doctoral Consortium, co-located with the 3rd World Conference on eXplainable Artificial Intelligence (xAI-2025), volume 4017 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[13] N. Pochinkov, B. Pasero, S. Shibayama, Investigating neuron ablation in attention heads: The case for peak activation centering, in: Late-breaking Work, Demos and Doctoral Consortium, co-located with the 2nd World Conference on eXplainable Artificial Intelligence (xAI-2024), volume 3793 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024.

[14] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS ONE 10 (2015) e0130140.

[15] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise relevance propagation: An overview, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer Nature, 2019, pp. 193–209.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[17] J. D. M. dos Santos, J. P. M. dos Santos, Explaining ANN-modeled fMRI data with path-weights and layer-wise relevance propagation, in: Late-breaking Work, Demos and Doctoral Consortium, co-located with the 1st World Conference on eXplainable Artificial Intelligence (xAI-2023), volume 3554 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

[18] S. Pe, T. Buonocore, G. Nicora, E. Parimbelli, SignalGrad-CAM: Beyond image explanation, in: Late-breaking Work, Demos and Doctoral Consortium, co-located with the 3rd World Conference on eXplainable Artificial Intelligence (xAI-2025), volume 4017 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[19] Z. Shahid, Y. Rahulamathavan, S. Dogan, Second glance: A novel explainable AI to understand feature interactions in neural networks using higher-order partial derivatives, in: Late-breaking Work, Demos and Doctoral Consortium, co-located with the 2nd World Conference on eXplainable Artificial Intelligence (xAI-2024), volume 3793 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024.

[20] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, volume 11700, Springer Nature, 2019.

[21] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, in: Digital Signal Processing, volume 73, 2018, pp. 1–15.

[22] M. Jullum, J. Sjødin, R. Prabhu, A. Løland, eXplego: An interactive tool that helps you select appropriate XAI-methods for your explainability needs, in: Late-breaking Work, Demos and Doctoral Consortium, co-located with the 1st World Conference on eXplainable Artificial Intelligence (xAI-2023), volume 3554 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023.

[23] F. Schürmann, S. D. Sager-Müller, Interactive xAI-dashboard for semantic segmentation, in: Late-breaking Work, Demos and Doctoral Consortium, co-located with the 2nd World Conference on eXplainable Artificial Intelligence (xAI-2024), volume 3793 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024.

[24] C. L. Werner, J. Amling, C. Dormagen, S. Scheele, ClustXRAI: Interactive cluster exploration and explanation for process mining with generative AI, in: Late-breaking Work, Demos and Doctoral Consortium, co-located with the 3rd World Conference on eXplainable Artificial Intelligence (xAI-2025), volume 4017 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.