ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Ensembling Post-Hoc Image Explanations: When It Works, When It Fails, and How to Tell the Difference

Jinhua Xu[1], Davide Anguita[1], Fabio Roli[1,2], Jing Yuan[3], and Luca Oneto[1] *

1 - University of Genoa, Genova, Italy
2 - University of Cagliari, Cagliari, Italy
3 - Zhejiang Normal University, Jinhua, China

Abstract. Post-hoc explanation methods of image recognition models often exhibit high variance or disagreement across explanations when the input data are perturbed, the underlying models are modified, or different explainability techniques are employed. To mitigate this issue, several approaches have been proposed, among which ensemble strategies that aggregate multiple explanations have attracted particular attention. Although some of these methods demonstrate good empirical performance, most existing works remain largely empirical, with limited theoretical justification or understanding of why ensemble strategies work and when they fail. In this paper, we analyze the factors that influence the success and failure of ensemble strategies that combines multiple explanations, using different datasets, convolutional neural network architectures, post-hoc explanation techniques, and ensembling strategies to identify the most influential image patches. In particular, we compare various ensembling strategies based on distinct voting principles - namely, Borda Count, Kemeny–Young, Reciprocal Rank Fusion, and the Schulze method - and show that the performance of such ensemble methods depends on the degree of satisfaction of their underlying theoretical assumptions.

1 Introduction

In recent years, post-hoc explanation methods for image models have become an increasingly important research area, providing powerful tools for interpreting complex black-box deep learning models in critical domains such as autonomous driving and healthcare [1]. These methods typically aim to quantify the contribution of each input pixel, or image patch, to a model's prediction, following three main paradigms [2]: gradient-based methods (e.g., Guided Backpropagation and Integrated Gradients), game-theoretic methods (e.g., SHAP and GradientSHAP), and perturbation-based methods (e.g., Occlusion and Feature Ablation). However, the practical usefulness of these explanations is limited by a fundamental issue known as the *disagreement problem*, which arises when explanations become inconsistent under changes of the input data, model parameters, or the choice of explainability technique [3]. This inconsistency stems from several factors [4], including the high non-linearity of modern models, the perturbation assumptions made about the data, and the methodological assumptions underlying each explanation approach.

*This work is partially supported by (i) project ELSA – European Lighthouse on Secure and Safe AI funded by the European Union's Horizon Europe under the grant agreement No. 101070617, (ii) project SERICS (PE00000014) and FAIR (PE00000013) under the NRRP MUR program funded by the EU - NGEU, and (iii) project FISA-2023-00128 funded by the MUR program "Fondo italiano per le scienze applicate".

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Several approaches have been proposed to mitigate the disagreement problem [5], among which ensemble strategies - where multiple explanations are aggregated to produce a single consolidated output - appear to be the most promising. Different studies have explored various aggregation schemes [6], including normalization-based techniques, quality-based weighting, and supervised learning approaches. Despite their demonstrated empirical success, these methods are largely driven by heuristic principles. As a result, a fundamental theoretical understanding of the conditions under which an ensemble of explanations can solve or mitigate the above mentioned disagreement problem remains largely absent from the current literature.

In this work, the addressed explanation problem is the identification of the top-k $p\times p$-patches that most strongly influence a model's prediction. We conduct experiments on multiple datasets (ImageNet [7] and MedMNIST [8]), convolutional neural network architectures (ResNet [9] and DenseNet [10]), and pixel-wise attribution techniques [2](Saliency, InputXGradient, Guided Backpropagation, Deconvolution, Integrated Gradients, DeepLIFT, Occlusion, Feature Ablation, GradientSHAP, and DeepLIFT SHAP). Pixel-wise attributions are then aggregated into their corresponding $p\times p$-patch representations through simple averaging. Finally, we compare several ensemble strategies based on distinct voting principles (Simple Averaging [11], Borda Count [12], Kemeny–Young [13], Reciprocal Rank Fusion [14], and the Schulze method [15]). Our results show that these approaches perform well only when their underlying theoretical assumptions are approximately satisfied - specifically, when explanations are filtered to include only those that adhere to the noise model assumed by the corresponding voting principle. To the best of our knowledge, this study provides the first empirical evidence that theory-guided ensemble explanations for image post-hoc interpretation can simultaneously achieve state-of-the-art empirical performance.

2 Preliminaries

In this work, we consider a predictive model in the form of an image classifier $f : \mathcal{X} \to \mathcal{Y}$ that maps an RGB input image $x \in \mathcal{X} \subseteq \{0, 1, \ldots, 255\}^{r\times c\times 3}$ to its corresponding class label $y \in \mathcal{Y}$. A pixel-wise explanation for the model f for an image x with respect to the label y is represented by a matrix $e(f, x, y) \in \mathbb{R}^{r\times c\times 3}$, where large positive values indicate pixels that are more important for predicting y, whereas large negative values indicate pixels that contribute more to predicting the opposite class, denoted as $\neg y$, and small (positive or negative) values indicate less important pixels. A top-k $p\times p$-patch explanation $e_{k,p}(f, x, y)$ for the model f is defined as a matrix[1] in $\{0, 1, \cdots, \}^{r/p\times c/p}$, where each elements are are ranked from the top important 1 to least important. These elements correspond to the top-k most important $p\times p$ patches in the image, i.e., the regions that contribute most to the prediction of y (or $\neg y$), thus representing the decisive areas for the model's output. Note that $e(f, x, y)$ provides a more informative and fine-grained explanation compared to $e_{k,p}(f, x, y)$. However, the latter is significantly more interpretable for a human observer and can be derived easily from the former, e.g., by taking the average absolute value of $e(f, x, y)$ in

[1] With slight abuse of notation, we assume that r and c are multiples of p.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

the patch.

The quality of an explanation $e_{k,p}(f, x, y)$ can be assessed using different metrics. *Fidelity* which measures the decrease in classification performance when the top-k patches of the images (or their complements) are masked, denoted as $\mathcal{F}$ (or $\bar{\mathcal{F}}$). Masking consists of replacing the selected patches with an uninformative but non–disruptive value (e.g., class mean $\mathcal{F}^c$ or dataset mean $\mathcal{F}^d$). *Consistency* which quantifies the percentage of prediction changes in $f(x)$ after masking (as in Fidelity). This results in four metrics following the same naming convention as for fidelity: $\mathcal{C}^c$, $\mathcal{C}^d$, $\bar{\mathcal{C}}^c$, and $\bar{\mathcal{C}}^d$. Unlike Fidelity, this metric does not depend on the original accuracy of the model, allowing aggregation across datasets and model architectures. *Robustness* which measures the degradation in Fidelity or Consistency between $e_{k,p}(f, x, y)$ and $e_{k,p}(f, \tilde{x}, y)$, where $\tilde{x}$ is generated by adding noise to x. We adopt the extended naming notation $\mathcal{R}_\star^{\circ,*}$ (and $\bar{\mathcal{R}}_\star^{\circ,*}$), where $\star \in \{\mathtt{F}, \mathtt{C}\}$, $\circ \in \{\mathtt{g}, \mathtt{p}, \mathtt{s}\}$ correspond to Gaussian, salt–and–pepper, and speckle noise, respectively, and $* \in \{c, d\}$.

If multiple explanations are available, i.e., $\{e^{(1)}_{k,p}(f, x, y), e^{(2)}_{k,p}(f, x, y), \cdots\}$, it is possible to construct an ensemble explanation $e^{(e)}_{k,p}(f, x, y)$ using a voting mechanism to get an explanation which is better of any of the original ones.

3 Ensembling explanations

In this section, we describe the available voting mechanisms used to ensemble explanations, together with the underlying hypotheses that, when satisfied, allow us to establish their theoretical effectiveness.

The voting mechanisms for ensembling explanations can be broadly divided into two categories: (i) *theoretically grounded* methods, which offer guarantees of improvement or optimality under specific assumptions - such as the Borda count, Schulze, Reciprocal Rank Fusion, and Kemeny-Young methods; and (ii) *heuristic* methods, which are primarily practical and lack formal guarantees - such as simple averaging techniques.

Theoretically grounded ensemble methods typically rely on two main hypotheses. The first is the *independence assumption*, which requires that each ranking is generated independently of the others. The second is the *noise model* assumption, which states that the observed rankings π follow an exponential family distribution of the form

$$\mathbb{P}\{\pi \mid \pi^*\} \propto \exp\big[-\gamma\, d(\pi, \pi^*)\big], \tag{1}$$

where π^* denotes the (unobserved) ground-truth optimal ranking - namely, the oracle we aim to recover - d is a distance measure between rankings (e.g., an ℓ_1 distance), and γ is the dispersion parameter governing the noise level. In practice, of course, neither of the two hypotheses is fully satisfied.

Real independence is never fully achieved in practical applications, especially in *model explanation* or *data discovery* settings. In *model explanation*, all explanations are generated from that same model, violating the independence assumption. Similarly, in the data-discovery setting - where we examine explanations from multiple models trained on the same dataset to extract meaningful insights - the models themselves are trained on the same (or overlapped subset) data and are therefore dependent, leading to dependent explanations as well. In

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

data discovery, independence can be encouraged by training one or more models on different subsets of the data, thus allowing the generation of multiple, diverse, and approximately independent explanations. Nevertheless, as we will show with a simple experiment in Section 4, the lack of independence does not substantially affect the quality of the ensembling in practice.

What instead has a much deeper impact on the quality of the ensembled explanations is the extent to which the assumed noise model is actually satisfied. Ensembling explanations that do not conform to the hypothesized noise model can lead to outcomes worse than those of the best individual explanation in the set. Conversely, when we ensemble explanations that satisfy the noise model underlying the voting mechanism, the aggregated explanation consistently provides an improvement over the best single explanation. Note that we cannot directly verify the noise model, as doing so would require knowledge of unknown quantities such as the ground-truth ranking or the true noise distribution. Moreover, we do not even know whether the explanations are genuinely generated from an underlying ground-truth process. Nevertheless, we can estimate their empirical distribution, substituting the deterministic quantities to their empirical estimators. For this purpose, we proceeded as follows. Let us consider the *model explanation* setting where several explanations for multiple inputs of on model are available. We then select a subset of these explanations - for example, the top-n ones - and ensemble them using the techniques under study (i.e., Borda count, Schulze, Reciprocal Rank Fusion, or Kemeny-Young). In this step, we assume that only a subset of these explanations satisfies the noise model. Next, for each image, we compute the distance $d(\pi, \pi^*)$ between each individual explanation (ranking) π and the ensembled explanation π^*, and use these distances to estimate the empirical version of $\mathbb{P}\{\pi \mid \pi^*\}$ in Eq. (1), denoted by $\hat{\mathbb{P}}$. Note that $\hat{\mathbb{P}}$ converges to $\mathbb{P}$ as the number of examples increases. At this point, we show that the ensembled explanation begins to outperform the best individual explanation in the set precisely when the top-n explanations exhibit an empirical distribution $\hat{\mathbb{P}}$ that aligns with the noise model assumed by the ensembling strategy.

4 Empirical Evidences

In this work, we considered four distinct datasets: the natural image dataset *ImageNet* [7] and three medical image datasets from *MedMNIST* [8] (*BloodMNIST*, *DermaMNIST*, and *BreastMNIST*). The *ImageNet* dataset consists of 50,000 images across 1,000 classes. The *MedMNIST* datasets include *DermaMNIST* (10,015 images, 7 classes; 7,007 train / 1,003 validation / 2,005 test), *BloodMNIST* (17,092 images, 8 classes; 11,959 train / 1,712 validation / 3,421 test), and *BreastMNIST* (780 images, 2 classes; 546 train / 78 validation / 156 test). We then fine-tuned two well-known architectures on these datasets: *ResNet-18* [9] and *DenseNet-121* [10]. For the *ImageNet* experiments, we used the official pre-trained weights provided by Torchvision. For the *MedMNIST* experiments, the models were initialized from Torchvision and fine-tuned for 100 epochs using the Adam optimizer with a batch size of 128 and a learning rate of 1×10^{-5}. Explanations for the test data were generated using ten attribution methods: Saliency, InputXGradient, Guided Backpropagation, Deconvolution,

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Table 1: EXP2: Impact of noise model on ensemble performance (see Section 2 for the naming convention).

	$\mathcal{F}^d\downarrow$	$\bar{\mathcal{F}}^d\uparrow$	$\mathcal{C}^d\downarrow$	$\bar{\mathcal{C}}^d\uparrow$	$\mathcal{R}_{\mathsf{F}}^{\mathsf{g},d}\downarrow$	$\mathcal{R}_{\mathsf{C}}^{\mathsf{g},d}\downarrow$	$\mathcal{R}_{\mathsf{F}}^{\mathsf{p},d}\downarrow$	$\mathcal{R}_{\mathsf{C}}^{\mathsf{p},d}\downarrow$	$\mathcal{R}_{\mathsf{F}}^{\mathsf{s},d}\downarrow$	$\mathcal{R}_{\mathsf{C}}^{\mathsf{s},d}\downarrow$
ImageNet, ResNet-18										
Best Explanation	0.34	0.62	**0.34**	0.86	0.172	0.160	0.332	0.068	0.056	0.047
SimpleAvg	0.37	0.61	0.45	0.83	0.187	0.174	0.360	0.028	0.062	0.051
Borda (NAIVE)	0.35	**0.62**	0.43	0.87	**0.184**	0.173	0.346	0.045	0.059	0.048
Borda (THEORY)	**0.31**	0.62	0.37	**0.87**	0.170	0.157	0.301	0.048	**0.050**	**0.040**
RRF (NAIVE)	0.37	0.62	0.45	0.87	0.194	0.174	0.362	0.026	0.066	0.053
RRF (THEORY)	0.30	**0.62**	0.37	0.86	0.168	**0.156**	**0.297**	0.025	0.061	0.052
Schulze (NAIVE)	0.39	0.62	0.47	0.85	0.196	0.181	0.379	0.011	0.065	0.047
Schulze (THEORY)	0.31	0.62	0.37	0.86	0.173	0.160	0.305	**0.004**	0.059	0.046
Kemeny-Young (NAIVE)	0.35	0.62	0.42	0.87	0.184	0.170	0.345	0.048	0.063	0.049
Kemeny-Young (THEORY)	0.32	0.62	0.43	0.87	0.174	0.161	0.313	0.064	0.052	0.053
DermaMnist, ResNet-18										
Best Explanation	0.52	0.73	0.59	0.98	0.011	0.003	**0.033**	0.018	0.002	0.002
SimpleAvg	0.57	0.71	0.69	0.96	0.053	0.036	0.131	0.004	0.007	0.001
Borda (NAIVE)	0.51	0.73	0.59	**0.98**	0.001	0.008	0.082	0.054	0.014	0.019
Borda (THEORY)	0.49	0.73	0.55	0.98	0.005	**0.001**	0.099	0.005	0.005	0.007
RRF (NAIVE)	0.53	0.73	0.61	0.98	0.007	0.005	0.072	0.071	**0.001**	0.004
RRF (THEORY)	**0.48**	0.73	**0.54**	**0.98**	**0.001**	0.007	0.066	0.001	0.003	0.002
Schulze (NAIVE)	0.55	0.73	0.66	0.98	0.019	0.018	0.063	0.063	0.698	0.018
Schulze (THEORY)	0.49	**0.74**	0.55	0.98	0.004	0.010	0.068	0.030	**0.001**	**0.002**
Kemeny-Young (NAIVE)	0.51	0.73	0.58	0.98	0.010	0.026	0.099	0.048	0.015	0.007
Kemeny-Young (THEORY)	0.49	0.72	0.57	0.98	0.005	0.008	0.099	0.023	0.013	0.007

Integrated Gradients, DeepLIFT, Occlusion, Feature Ablation, GradientSHAP, and DeepLIFT SHAP, all implemented via the Captum [16] library with standard hyperparameters. Finally, the explanations were ensembled using Simple Averaging, Borda Count, Kemeny–Young, Reciprocal Rank Fusion (RRF), and the Schulze method. Then we performed two main experiments: EXP1 and EXP2.

EXP1 evaluates the impact of independence assumptions in a data discovery scenario. We split the dataset into disjoint subsets and train separate models, ensuring independence. Each model is then explained using different attribution methods, producing independent sets of explanations. These explanations are subsequently combined using various ensembling strategies. The results indicate no meaningful differences compared to training all models on the same data. We do not report the detailed results here due to space constraints, but they are available in the paper's repository[2].

EXP2 shows that a naive application of ensemble methods, without verifying the noise assumptions, does not always guarantee an improvement over a single explanation. For this analysis, we consider all combinations of datasets and models, apply all explanation methods to obtain patch-level explanations, and then combine them either naively (NAIVE) using the previously described ensembling strategies or by selecting the top-n explanations, with the minimal n required to satisfy the underlying noise model (THEORY), namely the exponential distribution of Eq. (1), using the Kolmogorov-Smirnov test. Table 1 reports the results for some representative datasets and models (the complete one are available in the paper's repository[2]), using $k = 20$ and $p = 14$. From Table 1, it is possible to observe that (i) empirical approaches such as Simple

[2] `https://github.com/Xu-jinhua/Theoretical-Ensemble-Strategies-for-XAI`

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Averaging and naive ensembling do not always provide benefits compared to selecting the best individual explanation and (ii) when explanations are ensembled using a theoretically guided strategy, there is a clear improvement that is both practically useful and theoretically grounded. Further investigation is needed to study the effect of k, p, the hyperparameters of the explanation methods, and the number and type of datasets and models. Nevertheless, these preliminary results indicate that this research direction is promising and sheds light on both practical and theoretical challenges related to improving and reconciling different explanations.

5 Conclusions and Future Work

In this work, we investigated the theoretical foundations of ensemble strategies for image post-hoc explanation and provided the first empirical evidence that theory-guided aggregation can outperform the best individual explanation. Our results show that naive ensembling does not consistently improve over single explanations, and that the key determinant of success is the extent to which the underlying noise model assumed by the voting mechanism is approximately satisfied. When explanations are filtered according to this criterion, theoretically grounded methods such as Borda Count, Kemeny-Young, Reciprocal Rank Fusion, and the Schulze method provide reliable and robust improvements across datasets and architectures.

References

[1] M. Mersha, K. Lam, and Others. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599:128111, 2024.
[2] D. Minh, H. X. Wang, and Others. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 55:3503–3568, 2022.
[3] S. Krishna, T. Han, and Others. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.
[4] S. Goethals, D. Martens, and T. Evgeniou. Manipulation risks in explainable ai: The implications of the disagreement problem. In *ECML PKDD*, 2023.
[5] U. Bhatt, A. Weller, and J. M. F. Moura. Evaluating and aggregating feature-based model explanations. In *IJCAI*, 2021.
[6] G. Laberge, Y. B. Pequignot, M. Marchand, and F. Khomh. Tackling the xai disagreement problem with regional explanations. In *AISTAT*, 2024.
[7] O. Russakovsky, J. Deng, and Others. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
[8] J. Yang, R. Shi, and Others. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.
[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE CVPR*, 2016.
[10] G. Huang, Z. Liu, L. Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE conference on computer vision and pattern recognition*, 2017.
[11] W. Hryniewska-Guzik, B. Sawicki, and Others. Normensemblexai: Unveiling the strengths and weaknesses of xai ensemble techniques. *arXiv preprint arXiv:2401.17200*, 2024.
[12] D. G. Saari. *The optimal ranking method is the Borda count*. Northwestern University, 1985.
[13] J. G. Kemeny. *Mathematical models in the social sciences*. Cambridge, Mass., MIT, 1962.
[14] G. V. Cormack, C. L. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *ACM SIGIR*, 2009.
[15] M. Schulze. The schulze method of voting. *arXiv preprint arXiv:1804.02973*, 2018.
[16] N. Kokhlikyan, V. Miglani, and Others. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.