Assessment of the Reliablity of a Model's Decision by Generalizing Attribution to the Wavelet Domain

Gabriel Kasmi[1,2] **Laurent Dubus**[2,3] **Yves-Marie Saint Drenan**[1] **Philippe Blanc**[1]
[1]MINES Paris, Université PSL Centre Observation Impacts Energie (O.I.E.)
[2]RTE France
[3]WEMC (World Energy & Meteorology Council, UK)
[1]{firstname.lastname}@minesparis.psl.eu
[2]{firstname.lastname}@rte-france.com

Abstract

Neural networks have shown remarkable performance in computer vision, but their deployment in numerous scientific and technical fields is challenging due to their black-box nature. Scientists and practitioners need to evaluate the *reliability* of a decision, i.e., to know simultaneously if a model relies on the *relevant* features and whether these features are *robust* to image corruptions. Existing attribution methods aim to provide human-understandable explanations by highlighting important regions in the image domain, but fail to fully characterize a decision process's reliability. To bridge this gap, we introduce the **W**avelet s**C**ale **A**ttribution **M**ethod (WCAM), a generalization of attribution from the pixel domain to the space-scale domain using wavelet transforms. Attribution in the wavelet domain reveals where *and* on what scales the model focuses, thus enabling us to assess whether a decision is reliable. Our code is accessible here: `https://github.com/gabrielkasmi/spectral-attribution`.

1 Introduction

Deep neural networks have become the standard for numerous computer vision applications. However, there is a growing consensus that these models cannot be safely deployed in real-world applications [5] as models are not reliable enough, owing to two reasons [43]. Firstly, the black-box nature of deep neural networks motivates using explainable AI (XAI) techniques to generate a human-understandable explanation of a model's decision [14]. Secondly, distribution shifts [28] are ubiquitous in real-world cases and cause models to fail unpredictibly [17]. Safely deploying deep learning in real-world settings requires at least tools that enable auditing the *relevance* and *robustness* to distribution shifts of a model's decision.

Attribution methods [3], which consist of identifying the most important features in the input, have improved the understanding of the decision process of deep learning models. On the other hand, Fourier analysis has been extensively used to analyze the robustness of models [60]. [55, 4, 63] showed that robust models rely on low-frequency components. Existing attribution methods only represent the model decision in the pixel (space) domain, while Fourier only provides a decomposition in the frequency (scale) domain. To the best of our knowledge, no work has yet expanded attribution in the space-scale domain, which enables the audit of both the relevance *and* robustness of a model's decision.

We introduce the **W**avelet s**C**ale **A**ttribution **M**ethod (WCAM). This novel attribution method represents a model decision in the space-scale (or wavelet) domain. The decomposition in the space-scale domain highlights which structural components (textures, edges, shapes) are important for the prediction, allowing us to assess the *relevance* of a decision process. Moreover, as scales correspond to frequencies, we simultaneously evaluate whether this decision process is robust. We discuss the potential of the WCAM for application in expert settings (e.g., medical imaging or remote sensing), show that we can quantify the robustness of a prediction with the WCAM and highlight concerning phenomenon regarding the consistency of the decision process of deep learning models.

2 Related works

Explainability Explainability in computer vision typically quantifies the contribution of an image's pixel or region to a model's prediction. Saliency [47] was among the first methods to identify such regions. The approach used the model's gradients and the classification score. A line of works improved this approach: instead of using the model's gradients, other works used the model's activations to generate explanations. It is the principle behind the class activation map (CAM, [62]), which has also been further refined [46, 44, 52]. These methods quickly compute an explanation but require access to the model's gradients or activation. We often refer to these methods as "white-box" explanation methods. By contrast, "black-box" methods are model agnostic. Explanations are computed by perturbing (e.g., occluding parts of the image) the inputs and computing a score that reflects the model's sensitivity to the perturbation. The various proposed methods, e.g., Occlusion [61], LIME [41], RISE, [36], Sobol [7], HSIC [35] or EVA [8] differ in that they use different sampling strategies to explore the space of perturbations and can be seen as special cases of a more general approach based on Shapley values [29]. However, the main limitation of these methods is that they only explain *where* the model focuses and are therefore not informative enough in many settings where one wants to assess *what* the model sees [1].

To begin addressing the *what*, [10] recently introduced CRAFT. This method combines matrix factorization for concept identification and Grad-CAM [44] for concept localization on the input image. Another line of work focused on identifying the most significant points in the training dataset through influence functions [27]. However, such approaches require access to the model and the training data and are, therefore, hard to implement in applied settings.

Frequency-centric perspective on model robustness A line of works aimed at explaining the behavior of neural networks through the lenses of frequency analysis. Several works showed that convolutional neural networks (CNNs) are biased towards high frequencies [55, 60] and that robust methods tend to limit this bias [63, 4]. Other works highlighted a so-called spectral bias [40, 59, 24], showing that CNNs learn the input image frequencies from the lowest to the highest. More recently, [57] leveraged Fourier analysis to characterize learning shortcuts [12]: this work showed that learning shortcuts are context-dependent as models tend to favor the most distinctive frequency to make a prediction.

These works showed that the decomposition of a model's decision in the Fourier (frequency) domain characterizes what models see on the input. However, the Fourier domain only represents the frequencies, while the Wavelet (space-scale) domain provides an assessment in both space and frequency simultaneously [32]. Expanding attribution into the wavelet domain could provide a more comprehensive assessment of what models see on the input, enabling the practitioner to gauge whether the model's decision is reliable.

3 Methods

3.1 Background

Wavelet transform A wavelet is an integrable function $\psi \in L^2(\mathbb{R})$ with zero average, normalized and centered around 0. Unlike a sinewave, a wavelet is localized in space and in the Fourier domain. It implies that dilatations of this wavelet enable to scrutinize different frequencies (scales) while translations enable to scrutinize spatial location. To compute an image's (continuous) wavelet transform (CWT), one first defines a filter bank $\mathcal{D}$ from the original wavelet ψ with the scale factor s and the 2D translation in space u. We have

$$\mathcal{D} = \left\{ \psi_{s,u}(x) = \frac{1}{\sqrt{s}} \psi\left(\frac{x-u}{s}\right) \right\}_{u \in \mathbb{R}^2,\ s \geq 0}, \quad (1)$$

where $|\mathcal{D}| = J$, and J denotes the number of levels. The computation of the wavelet transform of a function $f \in L^2(\mathbb{R})$ at location x and scale s is given by

$$\mathcal{W}(f)(x,s) = \int_{-\infty}^{+\infty} f(u) \frac{1}{\sqrt{s}} \psi^* \left(\frac{x-u}{s}\right) \mathrm{d}u, \tag{2}$$

which can be rewritten as a convolution [32]. Computing the multilevel decomposition of f requires applying Equation 2 J times with all dilated and translated wavelets of $\mathcal{D}$. [31] showed that one could implement the multilevel dyadic decomposition of the discrete wavelet transform (DWT) by applying a high-pass filter H to the original signal f and subsampling by a factor of two to obtain the *detail* coefficients and applying a low-pass filter G and subsampling by a factor of two to obtain the *approximation* coefficients. Iterating on the approximation coefficients yields a multilevel transform where the j^{th} level extracts information at resolutions between 2^j and 2^{j-1} pixels. The detail coefficients can be decomposed into horizontal, vertical, and diagonal components when dealing with 2D signals (e.g., images).

Sobol sensitivity analysis The Sobol sensitivity analysis consists of decomposing the variance of the output of a model into fractions that can be attributed to a set of inputs. Let $(X_1, \ldots, X_K)$ be independent random variables and $\mathcal{K} = \{1, \ldots, K\}$ denote the set of indices. Let f be a model, X an input, and $f(X)$ the model's decision (e.g., the output probability). We denote $f_\kappa = f_\kappa(X_\kappa)$ the partial contributions of the variables $(X_k)_{k\in\kappa}$ to the score $f(X)$. The Sobol-Hoeffding decomposition [20] decomposes the decision score $f(X)$ into summands of increasing dimension

$$f(X) = f_\emptyset + \sum_{\kappa \in \mathcal{P}(\mathcal{K}) \setminus \{\emptyset\}} f_\kappa(X_\kappa), \tag{3}$$

Where $f_\emptyset$ denotes the prediction with no features, $\mathcal{P}(\mathcal{K})$ denotes the power set of $\mathcal{K}$ and $\emptyset$ the empty set. Then, $\forall (u,v) \in \mathcal{K}^2$ such that $u \neq v$, $\mathbb{E}\left[f_u(X_u) f_v(X_v)\right] = 0$, we derive from Equation 3 the variance of the model's score

$$Var(f(X)) = \sum_{\kappa \in \mathcal{P}(\mathcal{K})} Var(f_\kappa(X_\kappa)), \tag{4}$$

Equation 4 enables us to describe the influence of a subset X_κ of features as the ratio between its own and the total variance. This corresponds to the first order **Sobol index** given by

$$S_\kappa = \frac{Var(f_\kappa(X_\kappa))}{Var(f(X))}. \tag{5}$$

S_κ measures the proportion of the output variance $Var(f(X))$ explained by the subset of variables X_κ [50]. Focusing on single features, S_k captures the *direct* contribution of the feature X_k to the model's decision. To capture the indirect effect, due to the effect of X_k on the other variables, **total Sobol indices** S_{T_k} [21] can be computed as

$$S_{T_k} = \sum_{\kappa \in \mathcal{P}(\mathcal{K}),\, k \in \kappa} S_\kappa. \tag{6}$$

Total Sobol indices (TSIs) measure the contribution of the k^{th} feature, taking into account both its *direct* effect and its *indirect* effect through its interactions with the other features.

Efficient estimation of Sobol indices As seen from Equation 5, estimating the impact of a feature k on the model's decision requires recording the partial contribution $f_k(X_k)$. This partial contribution corresponds to a *forward*. Estimating Sobol indices requires computing variances by drawing at least N samples and computing N forwards to estimate a first-order Sobol index S_k of a single feature k. As we are interested in the TSI of a feature k, we need to estimate the Sobol index of all sets of features $\kappa \in \mathcal{K}$ such that $k \in \kappa$. To minimize the computational cost of this computation, [7] introduced an efficient sampling strategy based on Quasi-Monte Carlo methods [34] to generate the N perturbations of dimension K applied to the input and used Jansen's estimator [23] to estimate the TSIs given the models' outputs and the quasi-random perturbations. Their approach requires $N(K+2)$ forwards [7].

To estimate the TSIs, they draw two matrices from a Quasi-Monte Carlo sequence of size $N \times K$ and convert them into perturbations, which they apply to X. The perturbated input yields two matrices, A and B. a_{jk} (resp. b_{kj}) is the element of A (resp. B) corresponding to the k^{th} feature and the j^{th} sample. For the k^{th} feature, they define $C^{(k)}$ in the same way as A, except that the column corresponding to feature k is replaced by the column of B. They then derive an empirical estimator for the Sobol index and TSI as

$$\hat{S}_k = \frac{\hat{V} - \frac{1}{2N}\sum_{j=1}^{N}\left[f(B_j) - f\left(C_j^{(k)}\right)\right]^2}{\hat{V}}, \qquad \hat{S}_{T_k} = \frac{\frac{1}{2N}\sum_{j=1}^{N}\left[f(A_j) - f\left(C_j^{(k)}\right)\right]^2}{\hat{V}}, \quad (7)$$

where $f_{\emptyset} = \frac{1}{N}\sum_{j=1}^{N} f(A_j)$ and $\hat{V} = \frac{1}{N-1}\sum_{j=0}^{N}\left[f(A_j) - f_{\emptyset}\right]^2$. Further implementational details can be found in [7].

3.2 The wavelet scale attribution method (WCAM)

Overview The **W**avelet s**C**ale **A**ttribution **M**ethod (**WCAM**) is an attribution method that quantifies the importance of the regions of the wavelet transform of an image to the predictions of a model. Figure 1 depicts the principle of the WCAM. The importance of the regions of the wavelet transform of the input image is estimated by **(1)** generating masks from a Quasi-Monte Carlo sequence, **(2)** evaluating the model on perturbed images. We obtain these images by computing the DWT of the original image, applying the masks on the DWT to obtain perturbed DWT,[1] and inverting the perturbed DWT to generate perturbed images. We generate $N(K+2)$ perturbed images for a single image. **(3)** We estimate the total Sobol indices of the perturbed regions of the wavelet transform using the masks and the model's outputs using Jansen's estimator [23]. [7] introduced this approach to estimate the importance of image regions in the pixel space. We generalize it to the wavelet domain.

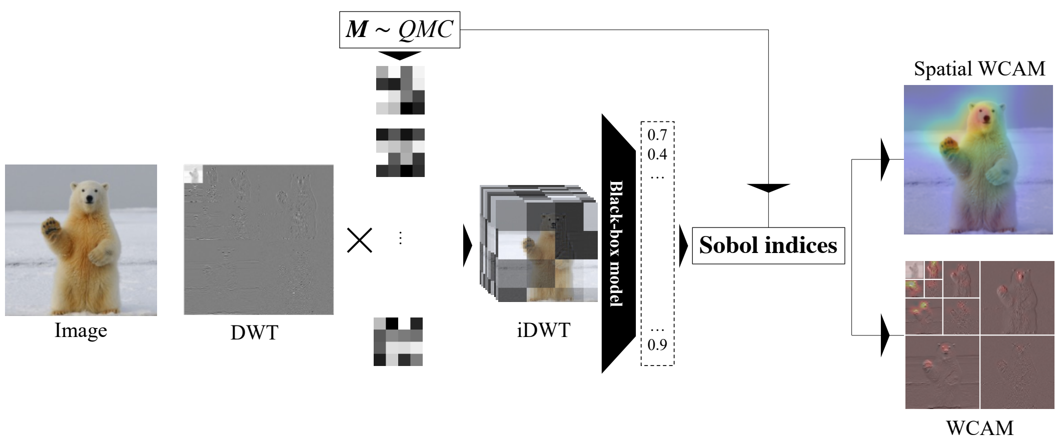

Figure 1: Flowchart of the wavelet scale attribution method (WCAM).

Generation of the masks We follow the sampling procedure introduced by [7] to generate the masks. Their approach involves drawing two independent matrices of size $N \times K$ from a Sobol low discrepancy sequence. N is the number of designs necessary to estimate the variance, and K is the sequence dimension. We reshape this sequence as a two-dimensional mask to generate our perturbation. By default, we perturb the wavelet transform with a mask of size 28 $\times$ 28 to balance between the dimensionality of the sequence and the accuracy of the perturbation. We reshape the 784-dimensional sequence to a grid of 28 $\times$ 28 to define our perturbation masks. We tried alternative mapping from the unidimensional sequence to the mask but with limited effect on the dimensionality reduction and at the expense of the meaningfulness of the perturbation in the wavelet domain. Figure 2 illustrates our workflow for one mask to generate the images that are then passed to the model.

[1] On an RGB image, we apply the DWT channel-wise and apply the same perturbation to each channel.

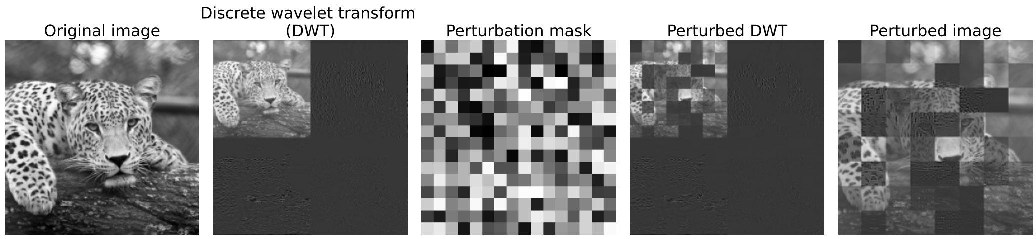

Figure 2: Workflow on a grayscale image and for a 1-level wavelet transform. We first compute the discrete wavelet transform of the image, then apply a mask on the DWT. It yields the perturbed DWT, which we invert to generate the perturbed image. We evaluate the model on the perturbed image.

The WCAM expands attribution to the space-scale domain The WCAM decomposes a prediction into the wavelet domain. As Figure 3 depicts, highlighting an important area in the pixel domain (i) does not provide information on *what* the model sees. By decomposing the prediction into the wavelet domain (ii), the WCAM represents the important features of a prediction in terms of structural components. In the example of Figure 3, we can see two important areas for predicting the fox: the hind leg and the ear. We can see that three distinct components contribute to the prediction for the ear. Areas **(a)**, **(b)**, **(c)** and **(d)** highlight these components. **(a)** corresponds to details at the 1-2 pixel scale, i.e., fine-grained details such as the fur in the ear. **(b)** corresponds to details at the 2-4 pixel scale, i.e., larger details such as the shape of the ear. We can see that both vertical (**(b)**) and horizontal (**(c)**) components of the shape of the ear contribute to the prediction. On the other hand, for the hind leg, only the overall shape (4-8 pixel size, **(d)**) contributes to the prediction.

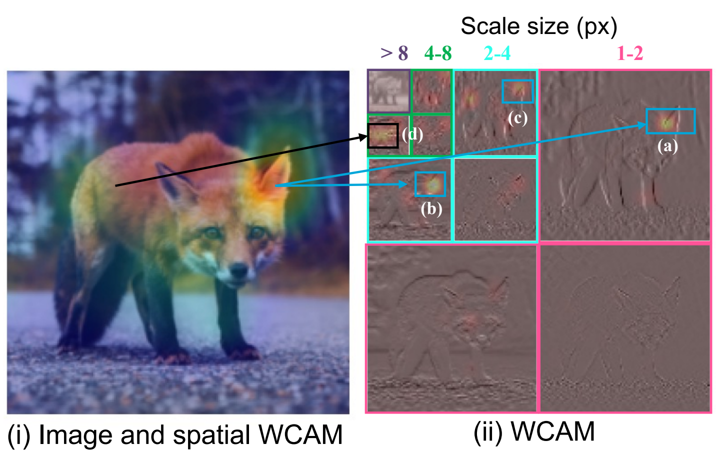

Figure 3: Decomposition of a prediction from the pixel domain (i) into the wavelet domain (ii) with the WCAM

4 Results and use cases

4.1 Results on evaluation benchmarks

μ-Fidelity We evaluate our method the μ-Fidelity, introduced by [2]. Contrary to insertion and deletion, which are area-under-curve metrics, the μ-Fidelity is a correlation metric. It measures the correlation between the decrease of the predicted probabilities when features are in a baseline state and the importance of these features. We have

$$\mu\text{-Fidelity} = \underset{\substack{u \subseteq \{1,\dots,K\}, \\ |u|=d}}{\text{Corr}} \left(\sum_{i \in u} g(x_i), f(x) - f(x_{x_u = x_0}) \right) \tag{8}$$

where g is the explanation function (*i.e.,* the explanation method), which quantifies the importance of the set of features u.

Results In Table 1, we evaluate our method against a range of popular methods and across various model architectures. Results show that we outperform existing black-box methods and are competitive with white-box attribution methods. The projection in the space-scale domain is the cause for the superiority of our method: we can see that the WCAM shows that the coarser scales are essential for a prediction. When flattening the WCAM accross scales (see appendix B.1 for more details) to derive the Spatial WCAM, our method performs similarly to other attribution methods. In appendix B.1, we provide additional evaluation results using the insertion and deletion scores [36].

Table 1: μ-**Fidelity** scores obtained on 100 ImageNet validation set images (higher is better). The best results are **bolded** and second best <u>underlined</u>. All benchmarks use the Xplique library [9].

	Method	*VGG16* [48]	*ResNet50* [16]	*MobileNet* [22]	*EfficientNet* [53]
μ-Fidelity (↑)					
White-box	Saliency [47]	0.043	0.060	-0.002	0.052
	Grad.-Input [46]	<u>0.105</u>	0.051	0.023	0.030
	Integ.-Grad. [52]	**0.137**	**0.112**	<u>0.130</u>	**0.134**
	GradCAM++ [44]	0.089	0.083	-0.001	0.063
	VarGrad [44]	0.054	<u>0.099</u>	**0.279**	<u>0.093</u>
Black-box	RISE [36]	0.020	0.074	-0.025	<u>0.042</u>
	Sobol [7]	0.095	<u>0.108</u>	-0.036	0.013
	Spatial WCAM (ours)	0.016	-0.037	<u>0.020</u>	-0.016
	WCAM (ours)	**0.197**	**0.191**	**0.105**	**0.187**

4.2 Assessing the robustness of a prediction

Scales, frequencies, and robustness Scales in the wavelet domain correspond to dyadic frequency ranges in the Fourier domain. The smallest scales correspond to the highest frequencies. Therefore, the WCAM connects attribution with frequency-centric approaches to model robustness. Figure 4 uses the WCAM to quantify the model's robustness. Following [4], we distinguish models trained with adversarial training ("AT," [30, 39, 45]), robust training ("RT," [18, 19, 13]) and standard training ("ST," e.g., ERM [54]). We can see that AT and RT models favor coarse scales (i.e., low-frequencies) over fine scales (i.e., high-frequencies). The WCAM characterizes robust models by estimating the importance of each frequency component in the final prediction. We can see that the ordering from the detail coefficients corresponding to the largest scales from those corresponding to the highest remains the same. This can be further assessed by the cumulative frequencies (right axis), where the most robust models have a more concentrated cumulative curve than the least robust models. These results are in line with existing works [63, 4, 55, 60] and show that the WCAM correctly estimates the robustness of a model.

Relevance of the important scales in expert settings The wavelet transform provides a multi-resolution decomposition of an image [32]. In other words, it isolates textures, edges, and shapes with respect to their scales and locations. The WCAM lets us see whether a model relies more on textures or shapes, which is impossible with existing attribution methods. In many settings, such information is valuable as it is interpreted in terms of specific features of the object of interest. Therefore, the WCAM links visual features (the model relies on) and semantic features (from which humans draw conclusions). For instance, in brain tumor classification from magnetic resonance imaging (MRI), small-scale details correspond to interpapillary capillary loop patterns [11]. In natural sciences, small scales can correspond to veins or small patterns on specific species of leaves [58, 51]. Finally, in remote sensing, scales correspond to different

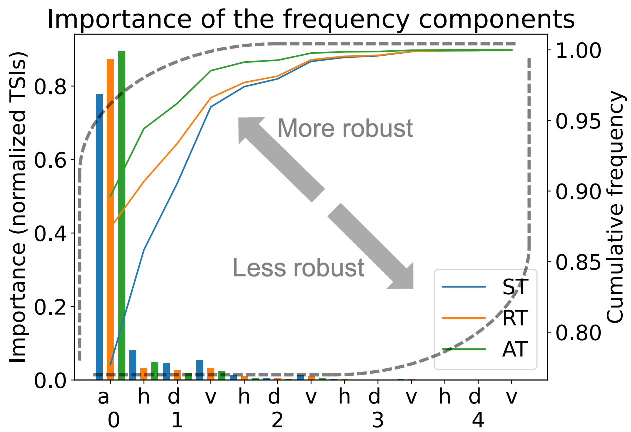

Figure 4: Representation of the scales of the WCAM as frequencies. Levels (numbered from (0 to 4) indicate the scales, from the coarser (i.e., lowest frequencies) to the finest (i.e., highest frequencies. The level 0 or "a" corresponds to the approximation coefficients. Labels "h," "v," and "d" correspond to the horizontal, vertical, and diagonal details, respectively. The rightmost index plots the cumulative curve. "AT," "RT," and "ST" stand for adversarial, robust, and standard training, respectively.

structural properties of small objects [26, 25], and information in the finer scales is more sensitive to image quality. If the latter varies too much (e.g., due to different acquisition conditions and signal-to-noise ratios), then information at these scales may disappear.

Towards assessing the reliability of a prediction We argue that a reliable prediction should be both relevant and robust. It is relevant as it relies on expected factors (or more generally on non-spurious factors) and robust as we want our prediction to be invariant to perturbations that can occur during the data acquisition process (e.g., heterogeneous properties of the optical transfer function or alterations of the signal-to-noise ratio) and limit the generalization capabilities of the model. Some relevant factors may be unreliable: in this case, the practitioner needs to be aware of that and adjust the data acquisition process. On the other hand, some models that are provably more robust may not be reliable if they rely on spurious components. The WCAM is a first step towards increasing the reliability of deep models in real-life settings as it unveils what scales are important and whether they are robust or not.

5 Conclusions and future work

We introduce the **W**avelet s**C**ale **A**ttribution **M**ethod (**WCAM**), a generalization attribution to the space-scale domain. The WCAM highlights the important regions in the space-scale domain using efficient perturbations of the wavelet transform of the input image. We estimate the contribution of the regions of the wavelet transform using total Sobol indices. Compared to existing attribution methods, the WCAM identifies what scales, which correspond to semantic features, are important for a prediction, thus providing more guidance regarding the relevance of the decision process of a model. Moreover, the WCAM connects attribution with robustness as scales in the wavelet domain correspond to frequencies in the Fourier domain. Applications of the WCAM lie in expert settings where practitioners need to evaluate the reliability of the prediction made by the model: does it rely on *relevant* components, and are these components *robust* to input perturbations that can occur in real life? Such assessment is crucial to improve the trustworthiness of deep learning models, especially as these models seem to be scale-inconsistent and rely on factors for contextual rather than semantical reasons.

Limitations and future works The main limitation of our approach is that it is computationally more expensive than existing black-box attribution methods. We plan to evaluate the benefits of the WCAM in an expert setting: the remote sensing of rooftop photovoltaic (PV) installations, where current models lack reliability due to their sensitivity to acquisition conditions [26].

6 Acknowledgements

This work is funded by RTE France, the French transmission system operator, and benefited from CIFRE funding from the ANRT. The authors gratefully acknowledge the support of this project. The authors would like to thank Thomas Fel, the first author of the original article *"Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis"* for his insightful comments and advice while preparing this work. The authors would also like to thank Hugo Thimonier for his helpful advice and feedback and Thomas Heggarty for proofreading the final manuscript.

References

[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From "Where" to "What": Towards Human-Understandable Explanations through Concept Relevance Propagation, June 2022. arXiv:2206.03208 [cs].

[2] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and Aggregating Feature-based Model Explanations, May 2020. arXiv:2005.00631 [cs, stat].

[3] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, July 2019.

[4] Yiting Chen, Qibing Ren, and Junchi Yan. Rethinking and Improving Robustness of Convolutional Neural Networks: a Shapley Value-based Approach in Frequency Domain. October 2022.

[5] Hervé Delseny, Christophe Gabreau, Adrien Gauffriau, Bernard Beaudouin, Ludovic Ponsolle, Lucian Alecu, Hugues Bonnin, Brice Beltran, Didier Duchel, Jean-Brice Ginestet, Alexandre Hervieu, Ghilaine Martinez, Sylvain Pasquet, Kevin Delmas, Claire Pagetti, Jean-Marc Gabriel, Camille Chapdelaine, Sylvaine Picard, Mathieu Damour, Cyril Cappi, Laurent Gardès, Florence De Grancey, Eric Jenn, Baptiste Lefevre, Gregory Flandin, Sébastien Gerchinovitz, Franck Mamalet, and Alexandre Albore. White Paper Machine Learning in Certified Systems, March 2021. arXiv:2103.10529 [cs] version: 1.

[6] Manik Dhar, Aditya Grover, and Stefano Ermon. Modeling Sparse Deviations for Compressed Sensing using Generative Models, July 2018. arXiv:1807.01442 [cs, stat].

[7] Thomas Fel, Remi Cadene, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis, November 2021. arXiv:2111.04138 [cs].

[8] Thomas Fel, Melanie Ducoffe, David Vigouroux, Remi Cadene, Mikael Capelle, Claire Nicodeme, and Thomas Serre. Don't Lie to Me! Robust and Efficient Explainability with Verified Perturbation Analysis, March 2023. arXiv:2202.07728 [cs].

[9] Thomas Fel, Lucas Hervier, David Vigouroux, Antonin Poche, Justin Plakoo, Remi Cadene, Mathieu Chalvidal, Julien Colin, Thibaut Boissin, Louis Bethune, Agustin Picard, Claire Nicodeme, Laurent Gardes, Gregory Flandin, and Thomas Serre. Xplique: A Deep Learning Explainability Toolbox, June 2022. arXiv:2206.04394 [cs].

[10] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. CRAFT: Concept Recursive Activation FacTorization for Explainability, March 2023. arXiv:2211.10154 [cs].

[11] Luis C. García-Peraza-Herrera, Martin Everson, Laurence Lovat, Hsiu-Po Wang, Wen Lun Wang, Rehan Haidry, Danail Stoyanov, Sébastien Ourselin, and Tom Vercauteren. Intrapapillary capillary loop classification in magnification endoscopy: open dataset and baseline methodology. *International Journal of Computer Assisted Radiology and Surgery*, 15(4):651–659, April 2020.

[12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. arXiv:2004.07780 [cs, q-bio].

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. April 2023.

[14] Hani Hagras. Toward Human-Understandable, Explainable AI. *Computer*, 51(9):28–36, September 2018. Conference Name: Computer.

[15] J. H. Halton. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, 2(1):84–90, December 1960.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, March 2019. arXiv:1903.12261 [cs, stat].

[18] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty, February 2020. arXiv:1912.02781 [cs, stat].

[19] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures, March 2022. arXiv:2112.05135 [cs].

[20] Wassily Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Foundations and Basic Theory*, Springer Series in Statistics, pages 308–334. Springer, New York, NY, 1992.

[21] Toshimitsu Homma and Andrea Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17, April 1996.

[22] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, April 2017. arXiv:1704.04861 [cs].

[23] Michiel J. W. Jansen. Analysis of variance designs for model output. *Computer Physics Communications*, 117(1):35–43, March 1999.

[24] Jason Jo and Yoshua Bengio. Measuring the tendency of CNNs to Learn Surface Statistical Regularities, November 2017. arXiv:1711.11561 [cs, stat].

[25] Gabriel Kasmi, Laurent Dubus, Yves-Marie Saint-Drenan, and Philippe Blanc. Can We Reliably Improve the Robustness to Image Acquisition of Remote Sensing of PV Systems?, September 2023. arXiv:2309.12214 [cs].

[26] Gabriel Kasmi, Yves-Marie Saint-Drenan, David Trebosc, Raphaël Jolivet, Jonathan Leloux, Babacar Sarr, and Laurent Dubus. A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata. *Scientific Data*, 10(1):59, January 2023.

[27] Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions, December 2020. arXiv:1703.04730 [cs, stat].

[28] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, and others. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[29] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks, September 2019. arXiv:1706.06083 [cs, stat].

[31] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[32] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[33] M. D. McKay, R. J. Beckman, and W. J. Conover. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, 21(2):239–245, 1979. Publisher: [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality].

[34] William J. Morokoff and Russel E. Caflisch. Quasi-Monte Carlo Integration. *Journal of Computational Physics*, 122(2):218–230, December 1995.

[35] Paul Novello, Thomas Fel, and David Vigouroux. Making Sense of Dependence: Efficient Black-box Explanations Using Dependence Measure, September 2022. arXiv:2206.06219 [cs, stat].

[36] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models, September 2018. arXiv:1806.07421 [cs].

[37] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient Starvation: A Learning Proclivity in Neural Networks, November 2021. arXiv:2011.09468 [cs, math, stat].

[38] Aahlad Puli, Lily Zhang, Yoav Wald, and Rajesh Ranganath. Don't blame Dataset Shift! Shortcut Learning due to Gradients and Cross Entropy, August 2023. arXiv:2308.12553 [cs, stat].

[39] Zhipeng Qu, Armel Oumbe, Philippe Blanc, Bella Espinar, Gerhard Gesell, Benoît Gschwind, Lars Klüser, Mireille Lefèvre, Laurent Saboret, Marion Schroedter-Homscheidt, and Lucien Wald. Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method. *Meteorologische Zeitschrift*, 26(1):33–57, February 2017.

[40] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the Spectral Bias of Neural Networks, May 2019. arXiv:1806.08734 [cs, stat].

[41] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, August 2016. arXiv:1602.04938 [cs, stat].

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, January 2015. arXiv:1409.0575 [cs].

[43] Suchi Saria and Adarsh Subbaswamy. Tutorial: Safe and Reliable Machine Learning, April 2019. arXiv:1904.07204 [cs].

[44] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020. arXiv:1610.02391 [cs].

[45] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial Training for Free!, November 2019. arXiv:1904.12843 [cs, stat].

[46] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences, April 2017. arXiv:1605.01713 [cs].

[47] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, April 2014. arXiv:1312.6034 [cs].

[48] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, April 2015. arXiv:1409.1556 [cs].

[49] I. M Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, January 1967.

[50] Ilya Meerovich Sobol. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990. Publisher: Russian Academy of Sciences, Branch of Mathematical Sciences.

[51] Edward J. Spagnuolo, Peter Wilf, and Thomas Serre. Decoding family-level features for modern and fossil leaves from computer-vision heat maps. *American Journal of Botany*, 109(5):768–788, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajb2.1842.

[52] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, June 2017. arXiv:1703.01365 [cs].

[53] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, September 2020. arXiv:1905.11946 [cs, stat].

[54] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[55] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8681–8691, Seattle, WA, USA, June 2020. IEEE.

[56] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. Frequency Shortcut Learning in Neural Networks. 2022.

[57] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. What do neural networks learn in image classification? A frequency shortcut perspective, July 2023. arXiv:2307.09829 [cs].

[58] Peter Wilf, Shengping Zhang, Sharat Chikkerur, Stefan A. Little, Scott L. Wing, and Thomas Serre. Computer vision cracks the leaf code. *Proceedings of the National Academy of Sciences of the United States of America*, 113(12):3305–3310, March 2016.

[59] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency Principle: Fourier Analysis Sheds Light on Deep Neural Networks. *Communications in Computational Physics*, 28(5):1746–1767, June 2020. arXiv:1901.06523 [cs, stat].

[60] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A Fourier Perspective on Model Robustness in Computer Vision, September 2020. arXiv:1906.08988 [cs, stat].

[61] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks, November 2013. arXiv:1311.2901 [cs].

[62] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[63] Zhuang Zhang, Dejian Meng, Lijun Zhang, Wei Xiao, and Wei Tian. The range of harmful frequency for DNN corruption robustness. *Neurocomputing*, 481:294–309, April 2022.