Applied Intelligence (2024) 54:1922–1936
https://doi.org/10.1007/s10489-024-05277-5

# Finding the input features that reduce the entropy of a neural network's prediction

**Narbota Amanova**[1] · **Jörg Martin**[1] · **Clemens Elster**[1]

Accepted: 7 January 2024 / Published online: 25 January 2024


**Abstract**
In deep learning-based image classification, the entropy of a neural network's output is often taken as a measure of its uncertainty. We introduce an explainability method that identifies those features in the input that impact most this uncertainty. Learning the corresponding features by straightforward backpropagation typically leads to results that are hard to interpret. We propose an extension of the recently proposed *oriented, modified integrated gradients* (OMIG) technique as an alternative to produce perturbations of the input that have a visual quality comparable to explainability methods from the literature but marks features that have a substantially higher impact on the entropy. The potential benefits of the modified OMIG method are demonstrated by comparison with current state-of-the-art explainability methods on several popular databases. In addition to a qualitative analysis of explainability results, we propose a metric for their quantitative comparison, which evaluates the impact of identified features on the entropy of a prediction.

**Keywords** Deep learning · Explainability · Image classification · Quantitative evaluation metric · Entropy reduction

## 1 Introduction

The number of successful applications of deep learning has increased rapidly in the last years [1], with deep learning being the state-of-the-art in disciplines such as medical applications, speech recognition, autonomous driving, and many more. Unfortunately, most applications focus on the accuracy of the results without examining their trustworthiness. Important aspects of the trustworthiness of neural networks comprise their robustness [2–4], e.g. to pertubations in the input, the quantification of uncertainty and the ability to explain the behavior of a neural network. As the interest in deep learning solutions grows, so does the interest in tools that help to understand and interpret machine learning approaches or evaluate the reliability by quantifying their uncertainty. Various explainability methods [5, 6] and uncertainty estimation techniques [7–10] have been developed to cope with the black-box nature of neural networks [11], and many of the Big Tech companies are creating tools for explainable deep learning [12, 13]. The focus of this work is to provide a new method for explainability linked with the predicted entropy of a neural network.

Any explainability technique should explain or present the decision of machine learning algorithms to humans [14]. In the case of input attribution, we expect an explainability method to highlight input features that are associated with some target value. Many metrics for the evaluation of explainability techniques are based on the manipulation of relevant features that are either removed or added to an analyzed image [15–17]. Thus, it is reasonable to assume that in the context of classification tasks, perturbation in the highlighted input features should be crucial for the predicted class probability.

Since a qualitative analysis of explainability methods has been identified as possibly being biased [18, 19], there is a growing interest in developing quantitative metrics. Quantitative analysis is performed, for example, in [16], where changes in the decision-making process due to the absence of some critical regions in the image in classification tasks are investigated. The authors confirmed that explanations indicate which features are important for the final decision, but the uncertainty of the final decision has not yet been studied in this context. Some papers suggest a re-training strategy to measure a model change when important features are removed [20, 21]. Many of the previous studies in

✉ Narbota Amanova
narbota.amanova@gmail.com

[1] Physikalisch-Technische Bundesanstalt, Abbestraße 2-12, Berlin 10587, Germany

quantitative explainability analysis have only been done on synthetic data, text classification tasks or binary image classification datasets [22–24]. Some of these require the provision of ground-truth heatmaps to evaluate heatmaps quantitatively [25].

In this work, we consider classification problems and propose an explainability technique that aims at highlighting those features that have the most impact on the entropy of the class probabilities predicted by a neural network. To this end we build on a novel explainability method introduced in [26] for the use case of image quality assessment for mammography and showed more meaningful results than other popular explainability methods. The successful employment of the method in a medical application raised the question of whether the proposed method can also be successfully applied in other fields. We address this question here and show how this method, called OMIG, can be used for image classification tasks to detect those input features that have the most impact on the entropy of the output of a neural network classifier.

To assess the impact of highlighted features on the entropy, we use a metric that measures how strongly the entropy of predicted class probabilities is affected when the input is modified depending on the heatmap produced by an explainability method. In this way, an uncertainty measure can be taken to assess explainability methods quantitatively.

So far, only limited research has been done directly combining explainability and uncertainty estimation. Examples comprise the sensitivity analysis of input variables for uncertainty quantification in Bayesian neural networks for regression tasks [27] or a deep generative model used to find counterfactual explanations of uncertainty quantification in [28]. In [29] the authors propose interpretable active learning strategies, which are typically built using measures such as uncertainty. In [30], the influence of features in the input on the epistemic uncertainty is assessed. The authors explained Bayesian dropout uncertainty [31] with Local Interpretable Model-Agnostic Explanations [32], and Layer-Wise Relevance Propagation [33] techniques. These previous research works are based on uncertainty quantification techniques, applied to synthetic or partially synthetic data, and they analyze exclusively binary image classification datasets.

This work introduces an extended/modified explainability method suitable for classification tasks in Section 2.1. Additionally, other established explainability methods are briefly described in Section 2.2. Furthermore, we define the proposed metric for quantitavely assessing explainability methods in Section 2.3. The results of both qualitative and quantitative comparisons between the proposed method and baseline explainability methods are presented in Section 3. Potential limitations of the proposed approach are discussed in Section 4, where also an outlook on further research is given.

## 2 Methods

### 2.1 Oriented modified integrated gradients

The OMIG explainability method was first introduced in the field of mammography image quality assessment, where it helped to explain predictions of image quality made by a trained neural network. The presented results demonstrated that OMIG obtains more meaningful heatmaps compared to other popular explainability methods and thus increases the trustworthiness of the employed deep learning model. Originally, OMIG was designed to detect a "natural direction" towards an expected image quality improvement. This is done by repeatedly applying several small changes of the input towards a direction expected to improve image quality. In this work, we adapt this idea to classification tasks and the problem of finding those input features that affect the uncertainty of the network prediction, that is, the entropy. We, therefore, pick the "natural direction" to point towards a lower entropy. The underlying principle behind OMIG, as presented in this work, is illustrated in Fig. 1.

In the following, let $f(x)$ be an output of a trained classifier with an input $x$. We assume that the last activation is given by a *Softmax* layer. This allows us to work with probability distributions to describe the $M$-dimensional predictions

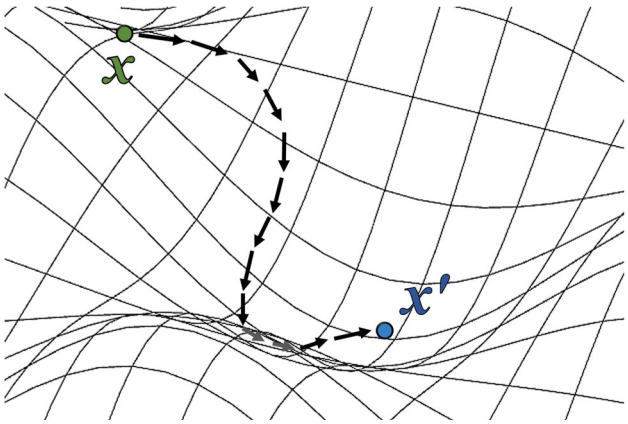



**Fig. 1** The path between the input image $x$ and its updated version $x'$ in the input space. The surface illustrates the entropy of the network prediction for elements from the input space. The path is chosen such that its endpoint $x'$ has a lower entropy compared to the starting point $x$. The position of $x'$ depends on the choice of hyperparameters $c$ and $L$, with $c$ determining the path length, which must be fine-tuned depending on the use case. $\varepsilon = c/L$ is responsible for the step size. Note that $x'$ is not necessarily always placed in a local or global minimum since we want to find an $x'$ close to $x$. $L$ is chosen large enough so that the step size $\varepsilon$ is kept small enough. The OMIG explanation for $x$ is obtained as the difference between $x$ and $x'$ representing the gradients that have been "integrated" along the path. In contrast to OMIG, the integrated gradients method [34], on which OMIG builds, utilizes gradients that are "integrated" along a straight path to a baseline. Figure 1 was created using [35]

$f(x) = \left(f^1(x), \ldots, f^M(x)\right)$ for $M$ number of classes. The entropy is then given by

$$H(f(x)) = -\sum_{m=1}^{M} f^m(x) \log\left(f^m(x)\right). \tag{1}$$

As the entropy is a measure of the uncertainty associated with a prediction, it is natural to ask the following question:

*How do you modify a given input image x to a new image x' so that the neural network can predict the same class as for x, but with less entropy?*

We therefore want to construct an $x'$, close to $x$, such that $H(f(x')) < H(f(x))$. To this end, we use the gradients of the entropy $H(f(x))$ w.r.t. the input:

$$R(x) = \frac{\partial H(f(x))}{\partial x}. \tag{2}$$

We set $x_0 = x$, choose $c > 0$ and $L > 0$, set $\varepsilon = c/L$ and successively update $x$ using the gradients:

$$x_{l+1} = x_l - \varepsilon \cdot R(x_l) \tag{3}$$

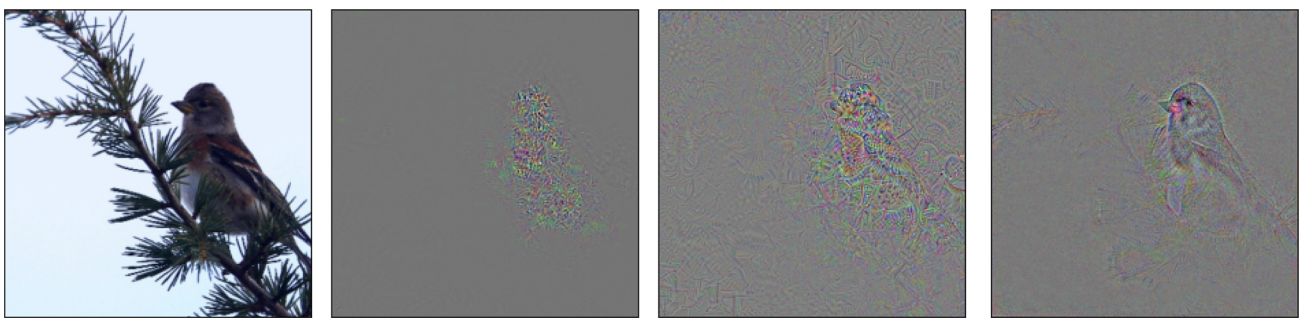
for $l \in [0, L-1]$. Finally, we set $x' = x_L$. The difference between the start and the end point of the path, as illustrated in Fig. 1, is denoted as a preliminary explainability approach, namely

$$\Delta x = x' - x. \tag{4}$$

The results of the described preliminary approach are presented in Fig. 2(b). This heatmap highlights the object itself but does not highlight any particular features of the object. We observed a visual improvement of the results when the gradients $R(x)$ are being normalized by dividing $R(x)$ through its 2-norm, so that $R(x)$ in (3) is replaced with

$$\widetilde{R}(x) = \frac{R(x)}{\frac{1}{\sqrt{N}}\|R(x)\|_2} = \frac{R(x)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} R(x)_i^2}}, \tag{5}$$

where $N$ denotes the number of elements of $R(x)$. The normalization effectively increases the step size once the gradient becomes small and thus avoids getting stuck too early in the update (3). Figure 2(c) demonstrates how the gradient normalization influences the heatmap. Including the normalization highlights the bird silhouette, making it more visible. But we now observe additional unstructured features that are not associated with the target class. These artifacts are reminiscent of the results of DeepDream [36], VQGAN-CLIP [37] or Open-Edit [38], which are, in fact, conceptually close to the proposed preliminary approach since they also modify the input and hereby amplify specific patterns. To remove the mentioned artifacts and further increase the visual representation we add one further ingredient. We reduce noise with SmoothGrad [39], as illustrated in Fig. 2(d). This is done by adding small perturbations $\epsilon = \mathcal{N}(0, s^2)$ to $n_{\text{samples}}$ copies of an image of interest, estimating heatmaps for each perturbed copy, and computing the average of these $n_{\text{samples}}$ heatmaps. The quality of the heatmap improves with SmoothGrad using only 20 samples with noise whose standard deviation $s$ is chosen to be 10% of the pixel range, as can be seen in Fig. 2(d). Since gradient normalization and SmoothGrad techniques significantly improve the results of the preliminary approach, we call their combination together with the latter the *oriented, modified integrated gradients method for classification problems*, which is generalized in Algorithm 1 and is hereafter referred to as *OMIG*. In the interest of easier understanding, Algorithm 1 shows the analysis of a single perturbed



(a) Input image (b) preliminary approach from (4) (c) Normalized $R(x)$ (d) Normalized $R(x)$, with SmoothGrad

**Fig. 2** Comparison of the heatmaps for the input image (a). The heatmaps are estimated by applying combinations of $R(x)$ normalization and SmoothGrad

copy of an image of interest $x$, which represents a single implementation of SmoothGrad, meaning $n_{\text{samples}} = 1$. The final explainability heatmap $\Delta x$ is obtained by averaging over $n_{\text{samples}}$.

While the OMIG method introduced in this work resembles the OMIG method proposed in [26], there are some substantial differences, namely:

- We use the entropy $H(f(x))$ to give the classification problem a natural direction. In [26] a different direction was used instead.
- We found the gradient normalization from (5) to have a better impact on the performance than the $Z$-score based gradient filtering applied in [26].
- We use SmoothGrad to reduce noise in the produced heatmaps. In a way, this has some resemblance with the upsampling procedure applied in [26].

In the form presented in this article, OMIG has 3 hyperparameters: $\varepsilon$, $c$ and $n_{\text{samples}}$. Finetuning these parameters in dependence on the considered classification task and dataset is recommended. For the examples shown in this paper, finetuning has been done qualitatively by looking for a good trade-off between entropy reduction and heatmap quality. In Appendix C we provide a recommendation for the procedure of hyperparameter finetuning and give some examples. Note, that the computational time of OMIG is also dependant on the the employed hyperparameters, especially on $n_{\text{samples}}$. To obtain informative and smooth heatmaps, OMIG relies, as explained above, on optimization and smoothing over $n_{\text{samples}}$ inputs. This makes the method computationally more expensive than methods that only use a single backward pass (such as, e.g., Guided Backpropagation or Saliency) which can typically be performed in a fraction of a second for the examples used in this work. However, the computational time of OMIG of a few seconds per image (for $n_{\text{samples}} = 100$) for the examples in this work is sufficient from a practical perspective and substantially less than the time needed for pertubation based approaches such as occlusion sensitivity which requires several minutes for an image.

## 2.2 Further explainability methods

The model interpretability library Captum [40] for PyTorch [41] was used to implement other popular explainability techniques that are described in the following.

**Sensitivity analysis** (SA) [42] is a common technique where the gradients

$$R(x) = \frac{\partial \max_{m=1,\dots,M} (\log f^m(x))}{\partial x}$$

**Algorithm 1** OMIG explainability for classification tasks, $n_{\text{samples}} = 1$.

**Input:** input image $x$, trained neural network, perturbation $\epsilon$, parameters $\varepsilon > 0$, and an integer $L > 0$
**Output:** OMIG explainability for classification tasks $(x' - x_\epsilon)$
1: Perturb an input image $x_\epsilon = x + \epsilon$
2: Obtain $R(x_\epsilon)$ with (2)
3: **for** each $l$ in $0, \dots, L-1$ **do**
4: Obtain normalized gradients $\widetilde{R}(x_l)$ with (5)
5: $x_{l+1} = x_l - \varepsilon \cdot \widetilde{R}(x_l)$
6: Obtain $R(x_{l+1})$ with (2)
7: **end for**
8: $x' = x_L$

at the data point $x$ with $\max_{m=1,\dots,M} (\log f^m(x))$ as a maximal predicted probability of $M$ classes are computed, and their size used to indicate the importance. This is implemented in [40]. This method has become very popular because it is easy to implement via backpropagation. We additionally apply the SmoothGrad technique (with same parameters as for OMIG) to SA and refer to it as SA-SG.

**Guided backpropagation** (GB) [43] represents a concept close to the sensitivity analysis acting different in ReLU (rectified linear unit) non-linearities. Here, only those features with positive values are used for backpropagation.

**Integrated gradients** (IG) [34] are estimated by cumulating gradients along the straight line path from a baseline $x'$ to the input $x$. The authors assign the baseline $x'$ as a black image (a zero-intensity image).

**Gradients SHAP values** (GS) [44] is an approach close to integrated gradients and SmoothGrad. Here, the output gradients are calculated with respect to randomly chosen points between the input samples, which are perturbed by Gaussian noise, and the pre-defined baselines. SHAP values are then represented by the expectation of gradients × (inputs - baselines) [45].

## 2.3 Quantitative assessment of explainability

The main objective of this paper is to find those features in an image that have most impact on the entropy predicted by a neural network. We therefore want to assess a method with respect to the significance of the highlighted features for the entropy. To this end, we introduce the following quantitative criterion:

$$\frac{H(f(x_{\text{modified}}))}{H(f(x_{\text{original}}))} = \frac{H\left(f\left(x_{\text{original}} + \delta \cdot \sigma \cdot \Delta x / \widetilde{\|\Delta x\|_2}\right)\right)}{H(f(x_{\text{original}}))}, \quad (6)$$

where $\Delta x$ is the heatmap of some explainability method and $x_{\text{original}} + \delta \cdot \sigma \cdot \Delta x / \widetilde{\|\Delta x\|_2}$ is created from a perturbed image $x_{\text{original}}$. To allow for comparability between heatmaps

$\Delta x$ produced by different methods we first normalize the heatmaps by dividing them by

$$\widetilde{\|\Delta x\|_2} = \frac{\Delta x}{\frac{1}{\sqrt{N}}\|\Delta x\|_2} = \frac{\Delta x}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\Delta x_i^2}}$$

and then consider the quotient above for different scalings $\delta$. The variability of the analyzed dataset is taken into consideration within the image perturbation by scaling $\Delta x/\widetilde{\|\Delta x\|_2}$ with $\sigma$, where $\sigma$ is the standard deviation of all pixel values in the analyzed dataset.

Note, that the design of the criterion (6) is motivated by the underlying question of this work: which features impact the entropy? Neither the criterion (6) nor our proposed OMIG method from Section 2.1 should be understood as indicating that entropy yields a universal description of the behavior of the network. To capture the behavior of a network typically some sort of specification of what is meant by behavior is required. In our case this focus will mainly be laid on the uncertainty described by the predicted entropy.

## 3 Results

The analyzed models and datasets are given in Section 3.1. Section 3.2 presents the qualitative evaluation of the studied explainability techniques. Here, the predictions for randomly selected test images from the datasets mentioned in Section 3.1 are explained and compared with the results obtained by the OMIG method for classification problems. The results of the quantitative comparison of the explainability methods are then shown in Section 3.3.

### 3.1 Analyzed datasets and employed models

Four popular datasets are evaluated, which differ in image size, number of target classes, and consist of black and white and color images. First we analyze the **MNIST** dataset [46] which is a set of black and white images of handwritten digits. The model architecture from [47] is implemented. Further we evaluate the **fashion-MNIST** dataset [48] that is based on the articles of the Europe's largest online fashion platform and has the same number of target classes as MNIST. The model architecture from [49] is implemented for this dataset. In addition, **caltech 256** [50] and **ImageNet-V2** [51] datasets are studied. As in [52], a pre-trained ResNet-50 [53] model was finetuned for 10 categories of the caltech 256 dataset[1]

[1] The employed categories are "bear", "chimp", "giraffe", "gorilla", "llama", "ostrich", "porcupine", "skateboard", "triceraptors", and "zebra".

For the analysis of the 1000 categories of ImageNet-V2 a pre-trained VGG-16 from [54] was used.

In Appendix D we also study the effect of calibrating the considered networks via temperature scaling [55], which yields comparable results to the ones in this section.

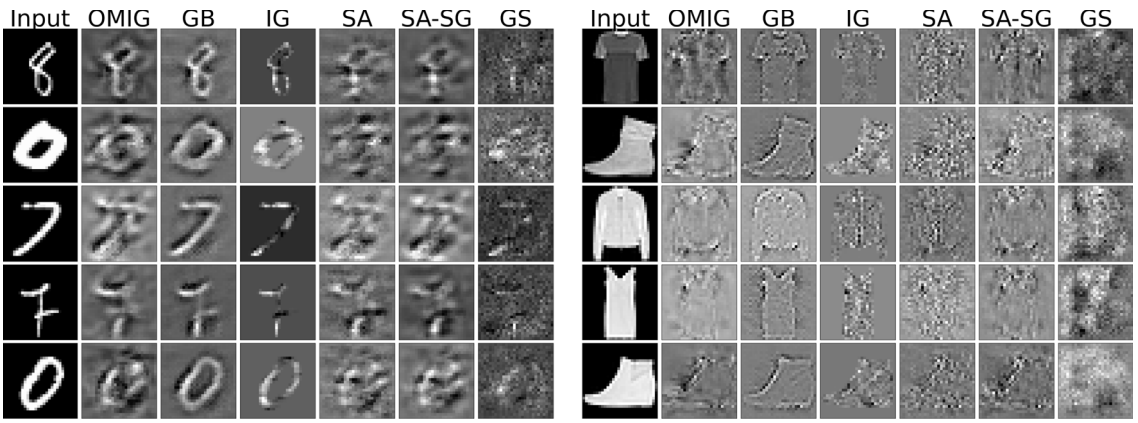
### 3.2 Qualitative analysis

In this section we evaluate the heatmaps generated by OMIG and competing methods from a qualitative point of view. More precisely, we evaluate whether the heatmaps are interpretable from a human perspective, that is whether the highlighted features are clearly perceptible and whether they are connected to actual structures in the input images. As the clear visibility of such features in the heatmaps is not necessarily linked to the actual impact of these features on the predicted entropy and the behavior of the network we evaluate the heatmaps quantitatively in the subsequent section.

The heatmaps of five randomly selected test images from MNIST and fashion-MNIST are shown in Fig. 3(a) and (b) respectively. Each row represents the selected image, and each column shows an explanation of the corresponding heatmaps made by OMIG, GB, IG, SA, SA-SG, and GS (in order from left to right), with abbreviations as in Section 2.2. Explainability results for further explainability techniques such as occlusion sensitivity (OS) and Guided Grad-CAM (GG) method are shown in Fig. 12 in Appendix E.

As can be seen from the analysis of MNIST in (a), the relevant features of the figures are highlighted by OMIG, GB, and IG. While the images of OMIG are slightly less clearcut than GB and IG, all three methods are of comparable visual quality and show the relevant areas of the image. The clearest images are those of IG. This can be explained by the nature of the IG estimation, where the gradients are "integrated" along a straight line between the baseline $x'$ (a zero-intensity image) and the input $x$. Since the MNIST dataset has a completely black background, gradients for background pixels are not computed because of the zero path length. We observed that the IG performance strongly depends on the nature of the image background: the more the background differs from zero, the worse the IG's performance becomes, as shown in Fig. 8 in Appendix A. Thus, the IG explanations can be compromised by the background. The results of SA, SA-SG, and GS are clearly unsatisfactory, highlighting no relevant features on the heatmaps.

In Fig. 3(b), the heatmaps for five examples from the fashion-MNIST dataset are presented. Here, we observe a similar performance of the analyzed explainability methods on the five randomly selected images as on the MNIST dataset. Object silhouettes are barely distinguishable in the




(a) Five randomly selected test images from the MNIST dataset

(b) Five randomly selected test images from the fashion-MNIST dataset

**Fig. 3** Heatmaps estimated using various explainability algorithms. From left to right: original **input** image, explainability obtained with **OMIG**, **G**uided **B**ackpropagation, **I**ntegrated **G**radients, **S**ensitivity **A**nalysis, **S**ensitivity **A**nalysis with applied **S**mooth**G**rad, and **G**radients **S**HAP values. The rows represent individual inputs. The test images are randomly chosen and only the classes, that were correctly predicted by the trained models, are analyzed

heatmaps obtained with SA, SA-GS, and GS in the last three columns of Fig. 3(b), showing the impracticability of using these methods to explain predicted classes. On the IG heatmaps, the objects are again separated from the background due to the black intensity of the background. The objects are correctly highlighted by OMIG and GB, just as for MNIST.

So far, only black and white datasets have been analyzed. The results of explaining the model predictions for the caltech 256 dataset are demonstrated in Fig. 4. We present the results in grayscale for the purpose of presentation, since we are not discussing heatmap colors, but structures highlighted by a particular method. The heatmaps evaluated by IG, SA, and SA-GS highlight no meaningful input features. GS highlights some structures that can be attributed to the target classes. In total, we observe that OMIG and GB are the only explainability methods highlighting relevant features for this dataset. The features of the target classes "gorilla", "zebra", "ostrich", and "bear" are clearly recognizable. The prediction for the "llama" class in the second row is explained by OMIG more

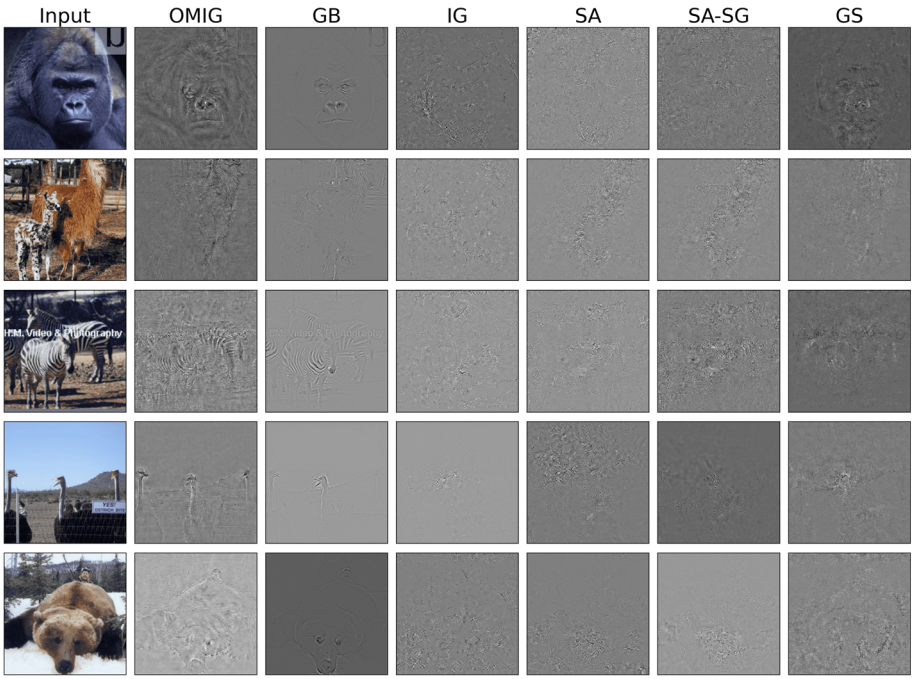



**Fig. 4** Heatmaps estimated using various explainability algorithms for five randomly selected test images from the caltech 256 dataset. From left to right: original **input** image, explainability obtained with **OMIG**, **G**uided **B**ackpropagation, **I**ntegrated **G**radients, **S**ensitivity **A**nalysis, **S**ensitivity **A**nalysis with applied **S**mooth**G**rad, and **G**radients **S**HAP values. The rows represent individual inputs. The test images are randomly chosen and only the classes, that were correctly predicted by the trained models, are analyzed

adequately. It is notable that in the fourth row, where GB highlights mostly a single ostrich (in center) as the most relevant object for the "ostrich" classification, whereas OMIG highlights all three ostriches that are present on the image.

Some heatmaps for the predictions made for the ImageNet-V2 dataset are presented in Fig. 5. The SA, SA-GS, IG, and GS heatmaps in the last three columns are more meaningful than those for the previous datasets. But the features of the objects are still not particularly recognizable - we again observe a better performance by OMIG and GB.

The "wine bottle" category in the first row is explained by OMIG and GB by correctly highlighting the target object. The features that belong to the target categories "brambling" and "chimpanzee" are also clearly highlighted by OMIG and GB on the two heatmaps in the second and third rows. The predictions for the fourth and the fifth images from the category "hot dog" and "football helmet" are explained by OMIG and GB by marking the areas that can be associated with the target objects, but these explanations are hardly comprehensible if the ground truth classes are beforehand unknown.

In the Appendix F we included for the images from Figs. 4 and 5 heatmaps showing only the 20% highest values. These plots exemplify even more that different features are deemed most relevant by the studied methods.

Summarizing the visual analysis of the heatmaps in this section, we can conclude that only OMIG and GB methods can explain model predictions well by highlighting meaningful input features for all datasets, while other methods, like IG, GS, SA, or SA-GS, did not yield meaningful results for some datasets. However, there are some cases where both OMIG and GB "fail" to explain the prediction, presenting a heatmap that is difficult to interpret. While GB yields more clearcut heatmaps, especially concerning structures such as

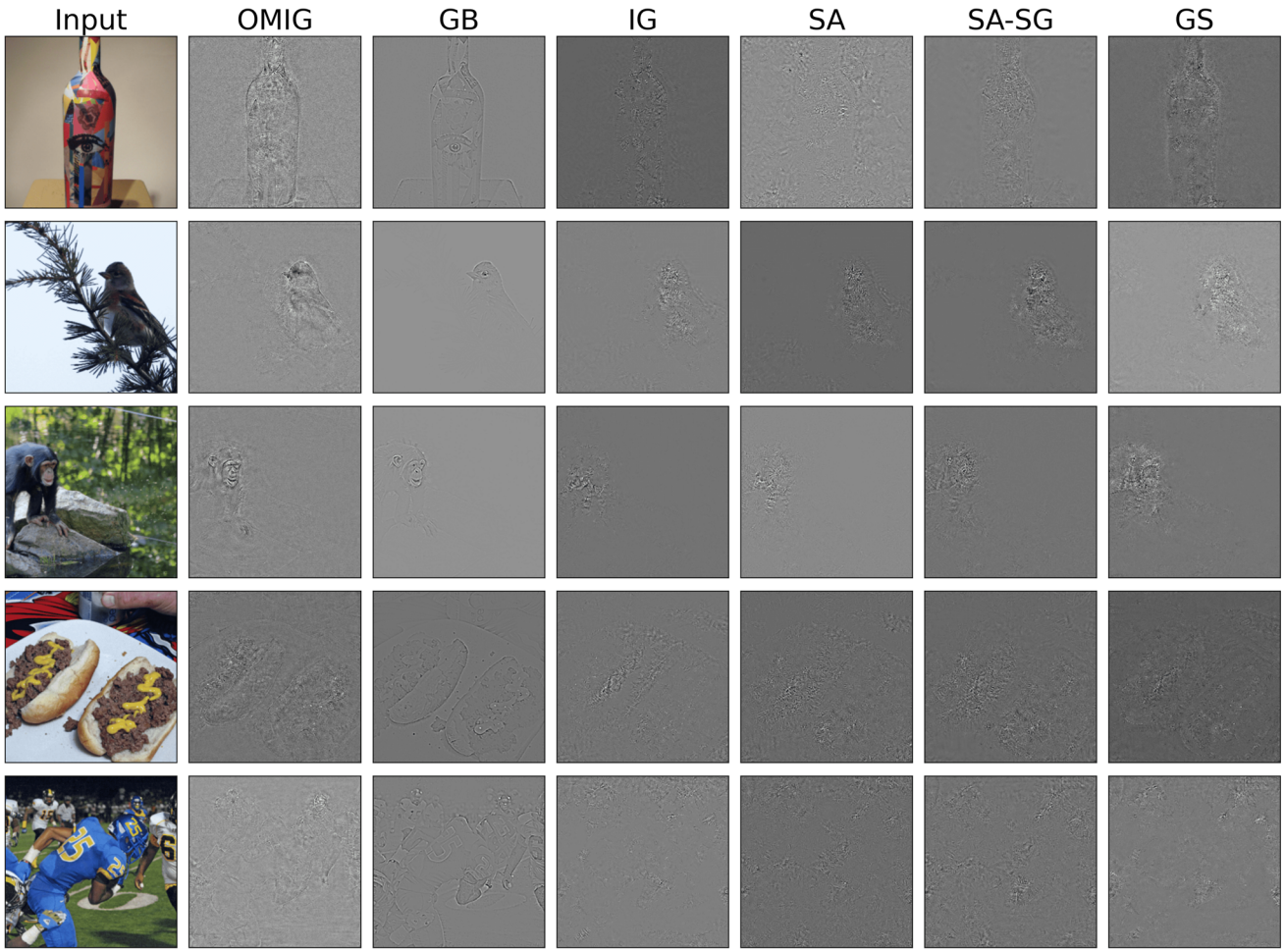



**Fig. 5** Heatmaps estimated using various explainability algorithms for five randomly selected test images from the ImageNet-V2 dataset. From left to right: original **input** image, explainability obtained with **OMIG**, **G**uided **B**ackpropagation, **I**ntegrated **G**radients, **S**ensitivity **A**nalysis, **S**ensitivity **A**nalysis with applied **S**mooth**G**rad, and **G**radients **SHAP** values. The rows represent individual inputs. The test images are randomly chosen and only the classes, that were correctly predicted by the trained models, are analyzed

edges, OMIG often highlights a larger part of the "relevant object". The different highlighting of different explainability methods is natural, as explainability methods are, by design, answering "different questions". While GB, for instance, highlights those features that are associated with some target class, we designed OMIG to highlight features that are most relevant for a change in the entropy, as this is the aspect we are most interested in this work. In following Section 3.3, we will study the performance of the explainability methods above under this aspect.

### 3.3 Quantitative analysis

One of the goals of this paper is to highlight those features that have the most impact on the entropy of the predicted class probabilities. To this end, we now analyze the performance of the explainability methods from Section 3.2 under the metric (6). The results are presented in Fig. 6.

In Fig. 6, the dashed lines show the median value of the relation (6) evaluated for a set of test images for different values of $\delta$ and for the explainability methods from Section 3.2. The shaded area depicts the interquartile range. The test images are randomly chosen and only the classes, that were correctly predicted by the trained models, are analyzed. For the chosen batch size of 500 images in the case of MNIST in (a), target classes were correctly predicted for 496 images, while for fashion-MNIST in (b), this number is represented by 465 images. For color datasets, the target classes were correctly predicted for 369 images for caltech 256. For ImageNet-V2 the batch size was reduced to 300, resulting in 175 correctly predicted classes.

In the top row of each image from Fig. 6, we observe that the features highlighted by OMIG lead to a more consistent entropy reduction when the input image is perturbed accordingly. The effect of the input features highlighted by the other methods on the entropy depends on the dataset and is not even present for some datasets. Interestingly, although IG gave the clearest heatmaps for MNIST in Fig. 3(a) and MNIST-fashion in (b), its entropy reduction effect for this dataset ranks as one of the smallest. Similarly, while GB showed a very clear representation of object edges for ImageNet-V2 and caltech 256, it does not seem to represent features whose perturbation reduces the entropy at any scale for these two datasets. For ImageNet-V2 and caltech 256, SA has an entropy reduction effect comparable to that of OMIG (and even stronger for caltech 256) at small perturbations $\delta$. Recall, however, that for these two datasets, SA gave only hard-to-interpret heatmaps, while OMIG provided heatmaps highlighting the structure of the corresponding objects much more clearly.

In addition to the evaluation via (6), we also evaluated the considered explainability methods under the Selectivity cri-

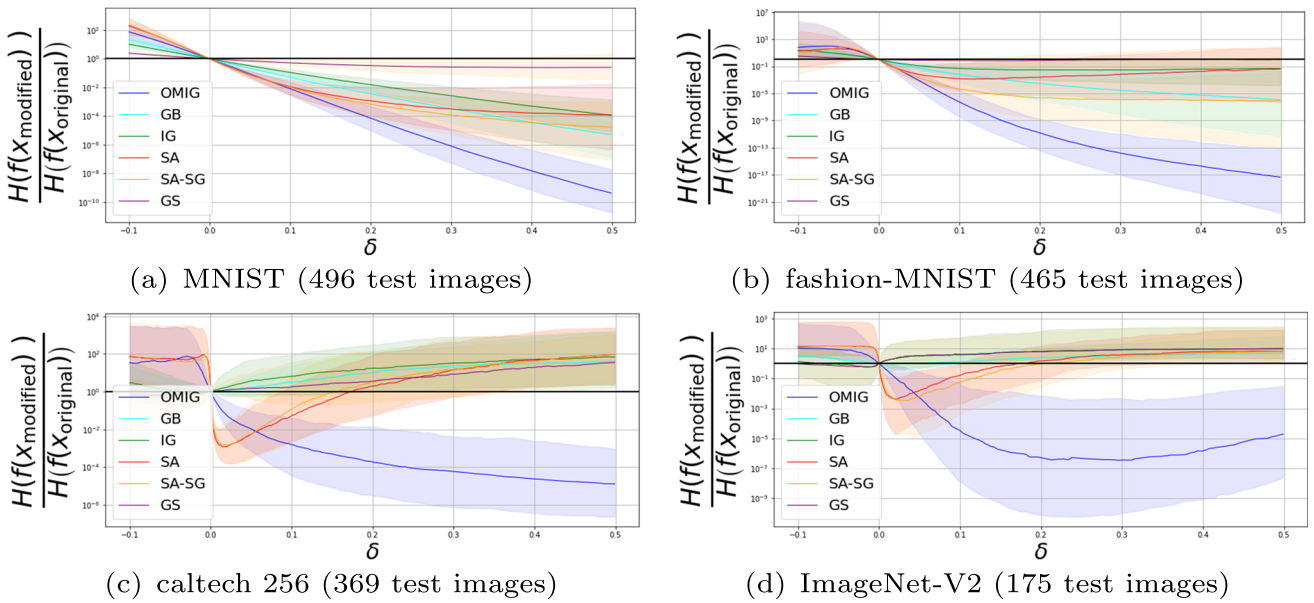



**Fig. 6** The relation between the prediction entropy before and after the application of the estimated heatmaps for different datasets estimated by (6). The colored dashed lines represent the median value over the range of analyzed input images, and the shaded areas show the interquartile range. The heatmaps are estimated with **OMIG**, **G**uided **B**ackpropagation, **I**ntegrated **G**radients, **S**ensitivity **A**nalysis, **S**ensitivity **A**nalysis with applied **S**mooth**G**rad, and **G**radients **SHAP** values. The black dashed line shows the relation (6) where $H\left(f\left(x_{\text{modified}}\right)\right)$ and $H\left(f\left(x_{\text{original}}\right)\right)$ are identical

terion from [42] on the ImageNet-V2 dataset with the same test images as above. For this criterion, patches of $8 \times 8$ pixels are "removed" (colored black) according to their relevance attributed by the explainability heatmap. Then, the predicted probability of the network for the class of the original image, which coincides with the true class for our test images, is computed. We limited our analysis to the removal of 200 patches. Figure 7 shows the according results for ImageNet-V2 after averaging over the considered test images. The lower the area under a curve the better it performs under the Selectivity criterion. We observe that OMIG yields in Fig. 7 the lowest curve with an AUC of 21.7, whereas the AUC for the competing methods are 25.7 (GB), 29.3 (IG), 32.6 (SA), 26.4 (SA-SG), and 23.1 (GS). OMIG yields the smallest standard error of the average predicted score for the ImageNet-V2 dataset with a value of 0.013. The values for other methods are 0.014 (GB), 0.016 (IG), 0.017 (SA), 0.014 (SA-SG), and 0.014 (GS). Hence, we see that OMIG does not only lead to the best entropy loss but also to the best Selectivity under the considered methods.

Let us summarize the observations from this section:

- The calculated heatmaps show an advantage of the OMIG and GB methods in explaining the trained models. We additionally observe that the quality of the heatmaps produced by IG strongly depends on the background of the input. While GB focuses on the edges of the object and shows clearer heatmaps, OMIG highlights most of the object.
- The use of SmoothGrad and the gradients normalization improves the quality of the heatmaps produced by OMIG and significantly improves the interpretability of the preliminary OMIG results.
- The quantitative results show that OMIG achieves the most consistent entropy loss and performs best under the Selectivity criterion from [42].

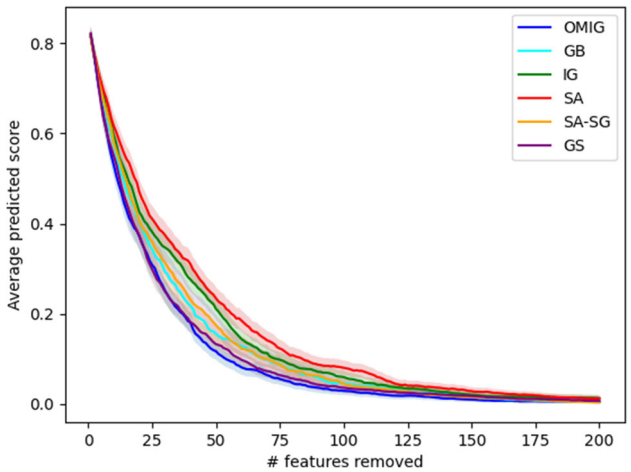



**Fig. 7** Evaluation of the Selectivity criterion from [42] for the explainability methods studied in this work on the ImageNet-V2 dataset. The x-axis shows the number of patches ($8 \times 8$) that were "removed" (that is colored in black) while the y-axis shows the predicted probability for the class that was predicted for the original image (which is identical with the true class for the considered images). The corresponding results were computed for the same images used for Fig. 6 and then averaged. The standard error of the mean is indicated by the shaded areas. The lower the area under a curve the better an explainability method performs under this criterion

## 4 Conclusion and outlook

We introduced an explainability method that highlights those features in the input which reduces the entropy of a neural network prediction. For this purpose, we used the OMIG explainability method, initially developed for mammography image quality assessment and adapted it for classification tasks. The resulting method produces heatmaps of a visual quality comparable to or better than other explainability methods from the literature. It shows features that lead to a more consistent reduction in entropy under the input perturbation.

The OMIG method relies on various hyperparameters, which in this article were finetuned for each dataset separately, see Table 1 in Appendix B. As mentioned in Section 2.1, the finetuning of the OMIG parameters can substantially change the visual performance of the estimated heatmaps. However, it is to be noted that this aspect influences the "meaningfulness" of a heatmap only in a human understanding. Determining best practices for finding optimal parameters given a dataset, or even identifying natural choices that work for each dataset, could be an interesting perspective for future work. Moreover, while entropy is a natural measure of uncertainty for classification problems, it only measures the aleatoric uncertainty of a prediction. Since OMIG only requires a natural direction, one could also use it to highlight features that affect other quantities, such as the epistemic uncertainty of Bayesian neural networks, which might[2] be another interesting topic of future work.

## Appendix A: Influence of the background color on integrated gradients

Figure 8 shows the analysis of the influence of the image background on the explainability results generated by the IG method. The results demonstrate that the image background can challenge the IG explanations. The more the background differs from zero, the worse the IG's performance becomes.

[2] narbota.amanova@gmail.com

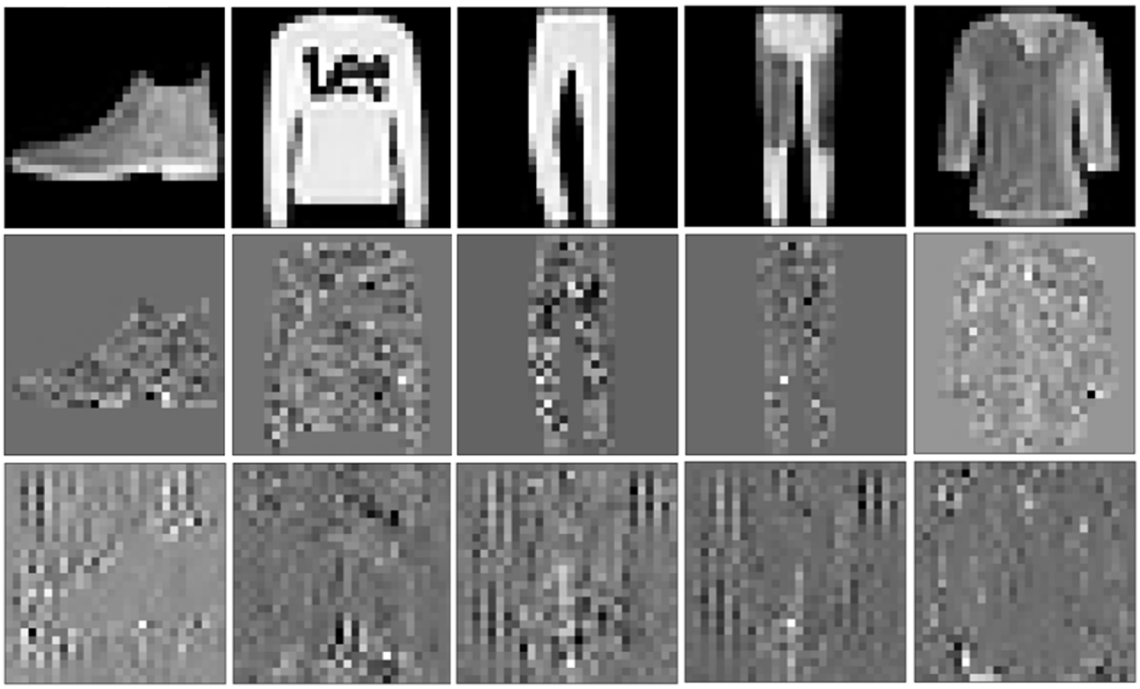


**Fig. 8** Dependence of IG performance on the input background. The top line shows five randomly selected images from fashion-MNIST, and the two bottom lines show the heatmaps estimated with IG on images with different normalization. The center line shows the explainability results on images where the background was black, and the bottom line shows the case where the images were normalized between -1 and 1. The employed networks were trained on the data with the corresponding normalization. The object on the images are clearly separated from the background on the heatmaps obtained from the images with the black background

## Appendix B: OMIG parameters

Table 1 shows the OMIG hyperparameters used for the analyzed image classification datasets were finetuned for each dataset separately. The recommendations on the finetuning procedure are given in Section 1.

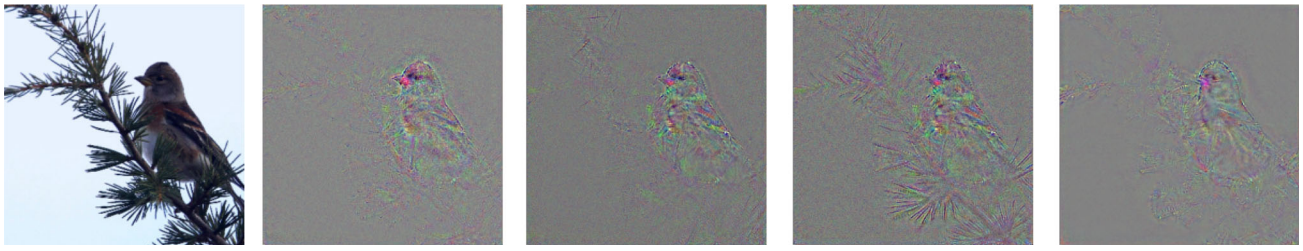
## Appendix C: Hyperparameters finetuning

The finetuning procedure for the OMIG hyperparameters should be adjusted for the particular classification problem considered. We recommend starting the finetuning with $n_{\text{samples}} = 50$ and $s = 0.1$ for initial SmoothGrad values. The OMIG hyperparameters are set to $c = 5$ and $\varepsilon = 0.1$. Depending on the visual quality of the achieved explainability results of the initial parameters, the step size $\varepsilon$ must be reduced. Increasing the number of samples controls the sharpness of the heatmaps – using a step size increment of twice the current number of samples is recommended. The $c$ value needs to be adjusted to control the strength of feature highlighting. The explainability results for various OMIG hyperparameters are presented in Fig. 9.

**Table 1** The OMIG parameters used for the analysis of four datasets

| OMIG parameters | MNIST | Fashion-MNIST | Caltech 256 | ImageNet-V2 |
|---|---|---|---|---|
| $c$ | 10 | 5 | 5 | 50 |
| $\varepsilon$ | 0.05 | 0.01 | 0.3 | 0.1 |
| $L$ | 200 | 500 | 16 | 500 |
| SmoothGrad, $n_{\text{samples}}$ | 100 | 100 | 100 | 100 |
| SmoothGrad, $\sigma$ | 0.15 | 0.05 | 0.08 | 0.08 |

The parameters were chosen empirically



(a) Input image (b) c = 50, $L = 500$ (c) c = 500, $L = 50$ (d) c = 50, $L = 10$ (e) c = 1, $L = 1000$

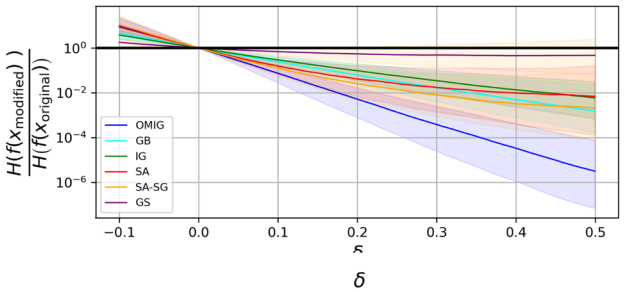
**Fig. 9** Explainability results with varied OMIG hyperparameters $\varepsilon$ for the input image in (a) from the ImageNet-V2 dataset

## Appendix D: Influence of the calibration of the trained classifiers on the OMIG results

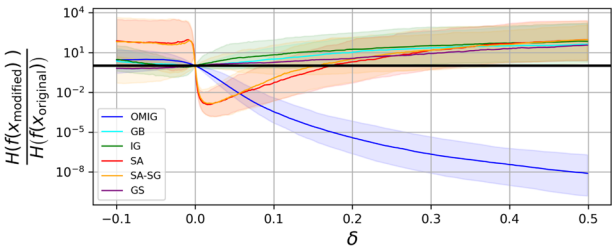
Figures 10 and 11 show the effect of calibrating the trained classifiers with the temperature scaling technique from [55] on the explainability results achieved using the OMIG method.


(a) MNIST (b) Caltech 256

**Fig. 10** Explainability results after applying the temperature scaling calibration method [55]




(a) MNIST




(b) Caltech 256

**Fig. 11** The relation between the prediction entropy before and after the application of the estimated heatmaps using calibrated trained classifiers for different scalings $\delta$

## Appendix E: Results of further explainability techniques

The following figures present the explainability results achieved using techniques such as occlusion sensitivity (OS) and Guided Grad-CAM (GG).

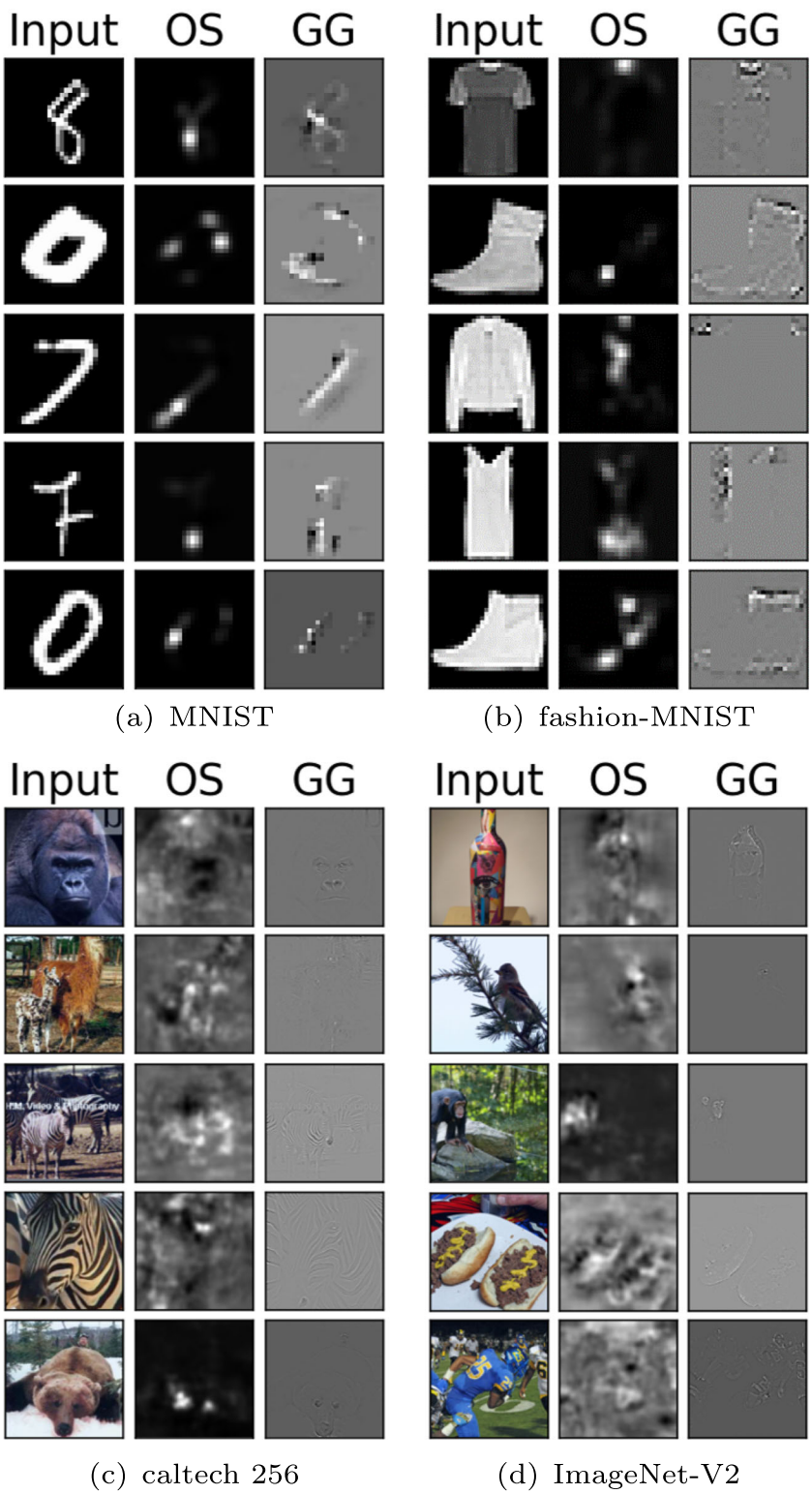



**Fig. 12** Explainability results achieved with occlusion sensitivity [56] and Guided Grad-CAM [57] analyzing five randomly drawn images from the corresponding datasets

## Appendix F: Discarding lower relevances

The images from Figs. 4 and 5 heatmaps are presented in the following, showing (Fig. 13) only the 20% highest values. The results illustrate that various methods prioritize different features.

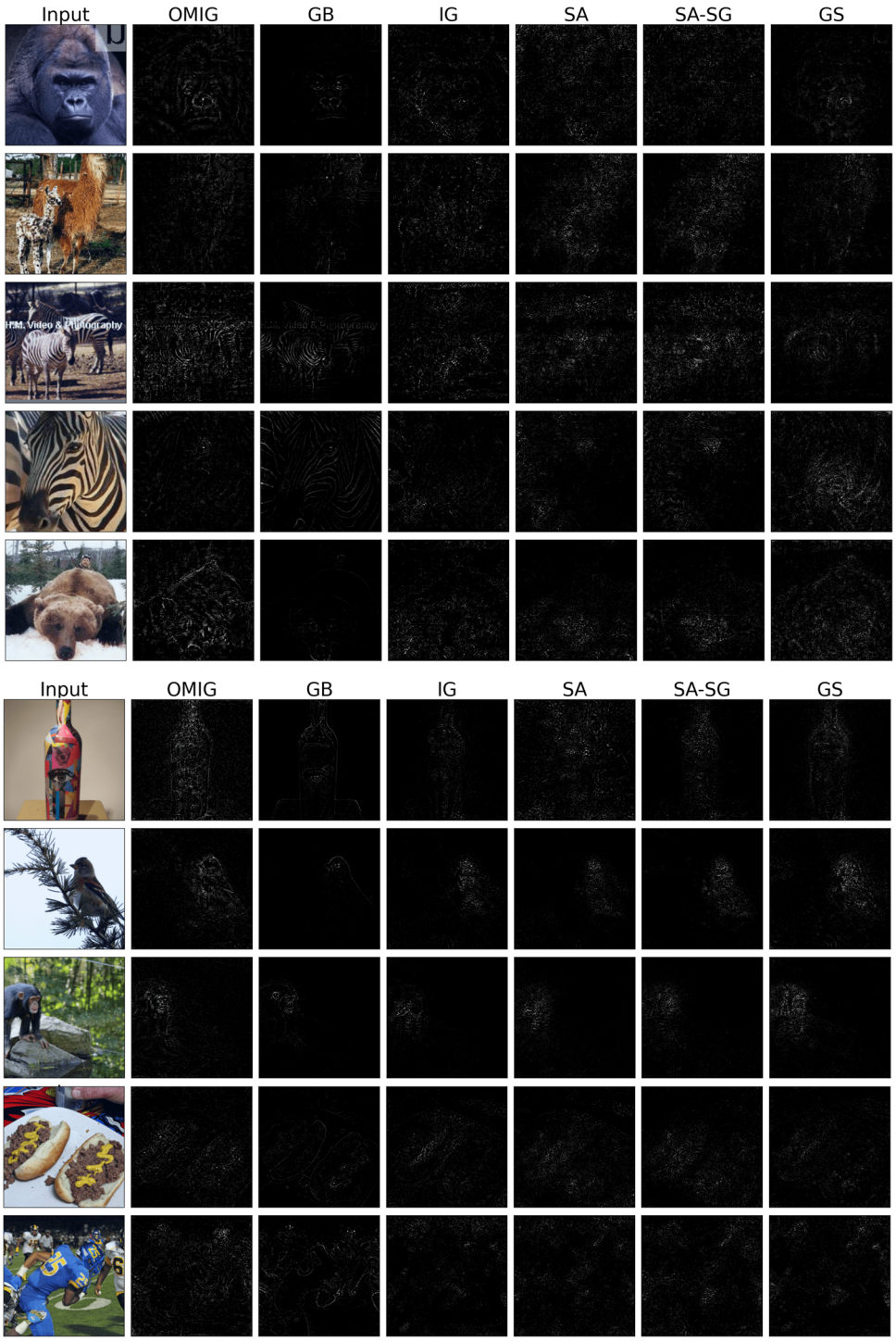



**Fig. 13** Heatmaps estimated using various explainability algorithms for five randomly selected test images from the caltech 256 and ImageNet-V2 datasets, where only the highest 20% of the heatmap values where kept and the remainder was set to zero

**Acknowledgements** This work was done within the PTB project ML4MedIm.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Data Availability** The datasets analyzed during the current study are publicly available. The code is available upon request from the corresponding author

## Declarations

**Competing interests** The authors have no financial or proprietary interests in any material discussed in this article.



## References

1. Zhang D, Mishra S, Brynjolfsson E, Etchemendy J, Ganguli D, Grosz B, Lyons T, Manyika J, Niebles JC, Sellitto M et al (2021) The AI index 2021 annual report. arXiv:2103.06312
2. Holzinger A (2021) The next frontier: Ai we can really trust. In: Joint European conference on machine learning and knowledge discovery in databases, Springer, pp 427–440
3. Holzinger A, Dehmer M, Emmert-Streib F, Cucchiara R, Augenstein I, Del Ser J, Samek W, Jurisica I, Díaz-Rodríguez N (2022) Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Inf Fusion 79:263–278
4. Martin J, Elster C (2021) Detecting unusual input to neural networks. Appl Intell 51:2198–2209
5. Das A, Rad P (2020) Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv:2006.11371
6. Minh D, Wang HX, Li YF, Nguyen TN (2022) Explainable artificial intelligence: a comprehensive review. Artif Intell Rev 55(5):3503–3568
7. Lambert B, Forbes F, Tucholka A, Doyle S, Dehaene H, Dojat M (2022) Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. arXiv:2210.03736
8. Tambon F, Laberge G, An L, Nikanjam A, Mindom PSN, Pequignot Y, Khomh F, Antoniol G, Merlo E, Laviolette F (2022) How to certify machine learning based safety-critical systems? A systematic literature review. Automated Softw Eng 29(2):1–74
9. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Cao X, Khosravi A, Acharya UR et al (2021) A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf Fusion 76:243–297
10. Pintz M, Sicking J, Poretschkin M, Akila M (2022) A survey on uncertainty toolkits for deep learning. arXiv:2205.01040
11. Burkart N, Huber MF (2021) A survey on the explainability of supervised machine learning. J Artif Intell Res 70:245–317
12. Explainable AI. https://cloud.google.com/explainable-ai?hl=en. Accessed: 2023-01-27
13. Microsoft: Model interpretability. https://azure.microsoft.com/en-us/products/machine-learning/
14. Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv:1702.08608
15. Mohamed E, Sirlantzis K, Howells G (2022) A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation. Displays 102239
16. Lin ZQ, Shafiee MJ, Bochkarev S, Jules MS, Wang XY, Wong A (2019) Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms. arXiv:1910.07387
17. Alvarez-Melis D, Jaakkola TS (2018) Towards robust interpretability with self-explaining neural networks. abs/1806.07538. arXiv:1806.07538
18. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. Adv Neural Inf Process Syst 31
19. Ghorbani A, Abid A, Zou J (2019) Interpretation of neural networks is fragile. Proceedings of the AAAI conference on artificial intelligence 33:3681–3688
20. Zhou J, Gandomi AH, Chen F, Holzinger A (2021) Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics 10(5):593
21. Hooker S, Erhan D, Kindermans P-J, Kim B (2019) A benchmark for interpretability methods in deep neural networks. Adv Neural Inf Process Syst 32
22. Wilming R, Budding C, Müller K-R, Haufe S (2022) Scrutinizing XAI using linear ground-truth data with suppressor variables. Mach Learn 1–21
23. Schmidt P, Biessmann F (2019) Quantifying interpretability and trust in machine learning systems. arXiv:1901.08558
24. Nguyen A-p, Martínez MR (2020) On quantitative aspects of model interpretability. arXiv:2007.07584
25. Tjoa E, Cuntai G (2022) Quantifying explainability of saliency methods in deep neural networks with a synthetic dataset. IEEE Trans Artif Intell
26. Amanova N, Martin J, Elster C (2022) Explainability for deep learning in mammography image quality assessment. Mach Learn: Sci Technol 3(2):025015
27. Depeweg S, Hernández-Lobato JM, Udluft S, Runkler T (2018) Sensitivity analysis for predictive uncertainty in bayesian neural networks. In: Proceedings of European symposium on artificial neural networks, computational intelligence and machine learning
28. Antorán J, Bhatt U, Adel T, Weller A, Hernández-Lobato JM (2020) Getting a CLUE: a method for explaining uncertainty estimates. In: Machine learning in real life workshop at ICLR 2020
29. Phillips R, Chang KH, Friedler SA (2018) Interpretable active learning. In: Conference on fairness, accountability and transparency, PMLR, pp 49–61
30. Brown KE, Talbert DA (2022) Using explainable ai to measure feature contribution to uncertainty. In: The international FLAIRS conference proceedings, vol 35
31. Gal Y, Ghahramani Z (2016) Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International conference on machine learning, PMLR, pp 1050–1059
32. Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Association for Computing Machinery, New York, USA, pp 1135–1144

33. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one 10(7):0130140
34. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: International conference on machine learning, PMLR, pp 3319–3328
35. Hohenwarter M, Borcherds M, Ancsin G, Bencze B, Blossier M, Delobelle A, Denizet C, Éliás J, Fekete A, Gál L, Konečný Z, Kovács Z, Lizelfelner S, Parisse B, Sturr G (2024) GeoGebra 4.4. http://www.geogebra.org
36. Mordvintsev A, Olah C, Tyka M (2015) Inceptionism: Going deeper into neural networks
37. Crowson K, Biderman S, Kornis D, Stander D, Hallahan E, Castricato L, Raff E (2022) VQGAN-CLIP: open domain image generation and editing with natural language guidance. In: European conference on computer vision, Springer, pp 88–105
38. Liu X, Lin Z, Zhang J, Zhao H, Tran Q, Wang X, Li H (2020) Open-Edit: Open-domain image manipulation with open-vocabulary instructions. In: European conference on computer vision, Springer, pp 89–106
39. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M (2017) SmoothGrad: removing noise by adding noise. arXiv:1706.03825
40. Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, Melnikov A, Kliushkina N, Araya C, Yan S, Reblitz-Richardson O (2020) Captum: a unified and generic model interpretability library for PyTorch
41. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al. (2019) PyTorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32
42. Montavon G, Samek W, Müller K-R (2018) Methods for interpreting and understanding deep neural networks. Digital Signal Process 73:1–15
43. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller MA (2015) Striving for simplicity: the all convolutional net
44. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30
45. Image Classification using Transfer Learning in PyTorch. https://captum.ai/api/_modules/captum/attr/_core/gradient_shap.html#GradientShap. Accessed 08 Dec 2022
46. LeCun Y, Cortes C (2010) MNIST handwritten digit database
47. PyTorch Convolutional Neural Network With MNIST Dataset. https://medium.com/@nutanbhogendrasharma/pytorch-convolutional-neural-network-with-mnist-dataset-4e8a4265e118. Accessed 29 Nov 2022
48. Xiao H, Rasul K, Vollgraf R (2017) Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747
49. Fashion MNIST with PyTorch. https://www.kaggle.com/code/pankajj/fashion-mnist-with-pytorch-93-accuracy. Accessed 29 Nov 2022
50. Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset
51. Recht B, Roelofs R, Schmidt L, Shankar V (2019) Do ImageNet classifiers generalize to ImageNet? In: International conference on machine learning, PMLR, pp 5389–5400
52. Fashion MNIST with PyTorch. https://learnopencv.com/image-classification-using-transfer-learning-in-pytorch/. Accessed 29 Nov 2022
53. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
54. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
55. Guo C, Pleiss G, Sun Y, Weinberger KQ (2017) On calibration of modern neural networks. In: International conference on machine learning, PMLR, pp 1321–1330
56. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer, pp 818–833
57. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626