Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 06, No. 02, July-Dec 2026
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.62.56.72

Research Paper

Do saliency maps find the defect? A ground-truth evaluation of explainable AI for safety-critical magnetic tile inspection

Heri Nurdiyanto[1*], Nuno Alexandre Soares Domingues[2]
[1*]Department of Industrial Engineering, Faculty of Engineering, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia.
[2]Adjunct Professor, Department of Mechanical Engineering, Instituto Superior de Engenharia de Lisboa (ISEL), Instituto Politécnico de Lisboa, Lisbon, Portugal.

Article Info

Article History:
Received: 14 June 2026
Revised: 14 August 2026
Accepted: 08 September 2026
Published: 22 October 2026

Keywords:
Explainable Artificial Intelligence
Saliency Maps
Surface Defect Detection
Industrial Visual Inspection
Safety-Critical Systems

ABSTRACT

Saliency maps are widely used to justify the decisions of deep-learning inspection systems, on the assumption that the highlighted region is the defect that caused the decision. This study tests that assumption on real industrial images. A convolutional neural network was trained to separate defective from defect-free tiles in the public Magnetic Tile Surface Defect dataset (1,344 images, five defect types, pixel-level defect masks), using five stratified train–test splits. Six saliency methods (Gradient, Gradient × Input, SmoothGrad, Integrated Gradients, Grad-CAM and Occlusion), a model-free Sobel edge map and a random map were evaluated on 382 held-out defective images for localisation against the masks, deletion-based faithfulness, cascading weight randomisation and agreement across independently trained models. The detector reached a test AUROC of 0.870 ± 0.032 and a balanced accuracy of 0.769 ± 0.061. Integrated Gradients, Gradient and Grad-CAM placed their maximum on the defect in about 60% of images, against 31% for the edge map and 15% for the random map, and Grad-CAM gave the best pixel AUROC (0.786 vs. 0.729 for edges, $p = 7 \times 10{-}9$). Localisation was poor for large, diffuse defects (fray, uneven surface: AUROC 0.47–0.72). Deleting the 5% most salient pixels reversed 18–35% of correct detections, compared with 1–2% for random deletion of the same number of pixels. However, removing the annotated defect itself by inpainting left 72% of detections unchanged, the same edit made only 10% of correctly classified defect-free tiles look defective, and blurring a whole defect-free image made 95% of them look defective. The maps also pointed at the defect about as often when the model missed it as when it detected it. Gradient-based maps lost their structure after weight randomisation, whereas Integrated Gradients retained a rank correlation of 0.44 with the original map even when every layer was randomised, and agreement between three identically trained models ranged from 0.21 (Gradient) to 0.71 (Grad-CAM). Saliency maps that look plausible on a defect therefore do not show that the decision was based on that defect.

Check for updates

Journal of Artificial Intelligence, Machine Learning and Neural

For safety-critical inspection, saliency should be validated with mask-based localisation, defect-removal tests, sanity checks and stability analysis before it is used as evidence.

Corresponding Author:
Heri Nurdiyanto
Department of Industrial Engineering, Faculty of Engineering, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia.
Email: herinurdiyanto@uny.ac.id

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Automated visual inspection is one of the most mature industrial applications of deep learning. Convolutional neural networks (CNNs) now detect cracks, pores, scratches and other surface defects on steel, ceramics, textiles and electronic components with an accuracy that matches or exceeds that of human inspectors [1], [2], [3]. Many of these components are safety-critical. Magnetic tiles, for example, are the permanent-magnet segments of electric motors used in vehicles, pumps and industrial drives; a tile that cracks or fragments in service can jam or destroy the motor. Inspection decisions for such parts are therefore increasingly expected to be not only accurate but also explainable and auditable, a requirement reinforced by recent regulation of high-risk AI systems [4] and by a wider debate on the use of black-box models in high-stakes settings [5], [6], [7].

The most common way to explain an image classifier is a saliency map, which assigns an importance score to every pixel [8], [9]. In an inspection setting, a saliency map is typically shown to an operator or quality engineer as a heat map overlaid on the part, with the implicit claim that the highlighted region is the defect that triggered the decision. If this claim is false, the explanation is worse than useless: it can lead an inspector to accept a rejected part because the highlighted area looks clean, or to trust a model that relies on background texture, lighting or image borders rather than on the defect itself, a failure known as “Clever Hans” behaviour [10].

Whether saliency maps can carry this claim has been questioned repeatedly. Sanity checks have shown that several popular methods produce almost the same maps when the network weights are randomised, which means that they reflect the image rather than the model [11]. Saliency maps can also change under transformations of the input that do not change the prediction [12], and different methods disagree with one another and with human-annotated regions [13], [14]. Most of this evidence, however, comes from natural-image or medical-image benchmarks. Industrial surface inspection has particular characteristics that make the question both easier to answer and more pressing: defects are small and often occupy less than 2% of the image, textures are highly repetitive, image borders and fixture edges produce strong gradients, and, importantly, public datasets provide pixel-level ground-truth masks of the defects, so that localisation can be measured directly rather than through proxy tasks [15].

This study uses that opportunity. A CNN was trained to separate defective from defect-free magnetic tiles on a public dataset of 1,344 real industrial images with pixel-level defect annotations [16]. Six widely used saliency methods and two reference baselines were then evaluated on held-out defective tiles along four dimensions that matter for safety-critical inspection: (i) localisation, i.e. whether the map points at and concentrates on the annotated defect; (ii) faithfulness, i.e. whether removing the highlighted pixels actually changes the model’s decision; (iii) model sensitivity, assessed with the cascading weight-randomisation sanity check [11]; and (iv) stability across independently trained models with identical data.

The contributions are as follows: A reproducible, mask-based benchmark of saliency explanations for industrial defect detection on real images, with five independent train–test splits and per-defect-class analysis. Evidence on how often each method points at the true defect, how much of its saliency mass falls on the defect, and how this compares with a model-free edge detector and a random map. A joint assessment of localisation, deletion-based faithfulness, sanity checks and cross-model stability, showing that these properties do not coincide and that no single metric is sufficient. Practical recommendations for using saliency maps as evidence in safety-critical visual inspection. The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the data, the detector, the explanation methods and the evaluation protocol. Section 4 presents the results, Section 5 discusses their implications and limitations, and Section 6 concludes.

2. RELATED WORK

2.1 Deep Learning for Surface-Defect Inspection

Surveys of visual defect detection document a shift from hand-crafted texture and filter features to CNNs for classification, detection and segmentation of defects [1]. Segmentation-based networks can be trained with relatively few defective samples and achieve high detection accuracy on industrial benchmarks [2], while unsupervised anomaly detection has been encouraged by datasets such as MVTec AD [3]. The magnetic tile dataset used in this study was released. together with a saliency-based defect segmentation network and contains five defect types with pixel-level masks [16]. These works focus on detection performance; explanations, when shown, are usually qualitative.

2.2 Saliency Methods

Gradient-based saliency computes the derivative of the class score with respect to the input pixels [8]. Gradient × Input multiplies this derivative by the input and is closely related to several propagation rules [17], [18]. SmoothGrad averages gradients over noisy copies of the input to reduce visual noise [19]. Integrated Gradients accumulates gradients along a path from a baseline to the input and satisfies completeness and sensitivity axioms [20]. Grad-CAM weights the feature maps of the last convolutional layer by their pooled gradients, producing a coarse class-specific map [9]. Occlusion, a perturbation-based approach, measures the drop in the class score when patches of the image are masked [21].

2.3 Evaluating Explanations

Explanations are commonly evaluated by localisation against annotated regions, for example with the pointing game [22] or the fraction of relevance inside a bounding box or mask [15], and by perturbation, where the most salient pixels are progressively removed and the decay of the class score is measured [23], [24]. Perturbation scores can be confounded by the distribution shift that the removal itself introduces [25]. Model-sensitivity tests randomise network weights or labels and check whether the explanation changes [11]; input-invariance tests check whether it changes when it should not [12]. Toolkits such as Quantus collect many of these metrics [26]. In medical imaging, systematic benchmarks have found that saliency maps often localise pathology worse than the models classify it, and that some methods fail sanity checks [13], [14]. To the author's knowledge, a comparable multi-criterion evaluation on real industrial defect images with pixel-level ground truth has not been reported, which motivates the present study.

3. METHODOLOGY

Figure 1 summarises the study design.

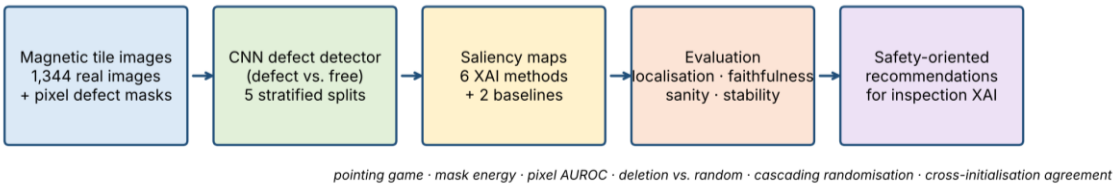

Figure 1. Study Pipeline for CNN-Based Magnetic Tile Defect Detection and Explainability Evaluation

3.1 Dataset

The study used the Magnetic Tile Surface Defect dataset released. [16] and distributed through a public repository [27]. It contains 1,344 grey-scale images of the curved surface of magnetic tiles cropped from production-line photographs, each with a binary mask that marks the defect pixels. Images are of variable size (103–632 pixels wide and 222–403 pixels high). The dataset has six categories: defect-free tiles and five defect types, namely blowholes, breaks, cracks, fray and uneven surfaces Table 1. Figure 2 shows real examples of each category with the annotated defect outlined. The defects differ strongly in size: blowholes and cracks typically cover less than 0.3% of the image, whereas fray and uneven-surface masks cover about 30%. Six uneven-surface images have empty masks and were excluded from the localisation analysis but retained for training.

Table 1. Composition of the Magnetic Tile Surface Defect Dataset and the Share of the Image Covered by the Defect Mask (After Resizing to 128 × 128 Pixels)

Category	Images	Share of Dataset (%)	Median Mask Area (% of Image)	Defective Test Images Evaluated*
Free (no defect)	952	70.8	—	—
Blowhole	115	8.6	0.12	115
Break	85	6.3	1.26	85
Crack	57	4.2	0.27	55
Fray	32	2.4	30.4	30
Uneven	103	7.7	28.4	97
Total	1,344	100	—	382

*Image–split instances with a non-empty mask over five test splits; an image can appear in the test set of more than one split (263 distinct images).

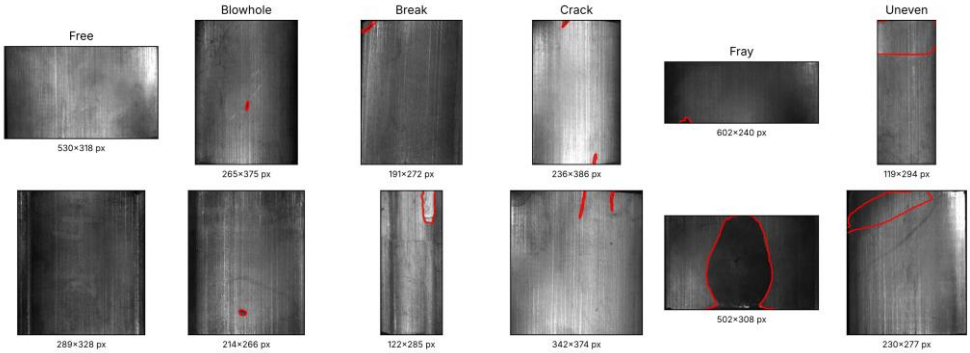

Figure 2. Representative Magnetic Tile Surface Defect Images with Annotated Defect Masks

All images were converted to grey scale, resized to 128 × 128 pixels with bilinear interpolation and scaled to [0, 1]; masks were resized with nearest-neighbour interpolation. Resizing to a square distorts the

aspect ratio, which differs between categories (median width-to-height ratio 0.86 for defect-free tiles and 2.33 for fray). This choice was kept deliberately because it is common in practice, and its consequences are examined in Section 4.5. The task was binary, defective versus defect-free, which is the decision that matters for accepting or rejecting a part.

3.2 Defect Detector

The detector was a compact CNN with four convolutional blocks [28], max pooling after the first three blocks, global average pooling of the final 128 × 16 × 16 feature map and a linear output layer with two classes. It was implemented in PyTorch [29] and trained with Adam [30] (learning rate 10−3, weight decay 10−4, batch size 32) for 30 epochs, with a class-weighted cross-entropy loss to compensate for the 70:30 class imbalance and random horizontal and vertical flips. The checkpoint with the highest balanced accuracy on a validation set was kept. Five independent experiments were run. In each, 20% of the images were held out as a stratified test set (269 images) and 15% of the remainder was used for validation (162 images), leaving 913 for training; the split and the network initialisation were controlled by the seed (0–4). No test image was used for training or model selection.

3.3 Saliency Methods

Six saliency methods were applied to the defect logit f(x) of the trained network. All maps were computed at the input resolution of 128 × 128 pixels and, where signed, their absolute value was used. Gradient [8]: $|\partial f/\partial x|$. Gradient × Input [17]: $|x \odot \partial f/\partial x|$. SmoothGrad [19]: the mean of $|\partial f/\partial x|$ over 25 copies of the image with Gaussian noise ($\sigma = 0.15$). Integrated Gradients [20]: path integral from a constant baseline x' equal to the mean image intensity, approximated with 32 steps (Equation 1). Grad-CAM [9]: the gradient-weighted combination of the last convolutional feature maps Ak, followed by ReLU and bilinear upsampling (Equation 2).

Occlusion [21]: the drop in P(defect) when a 16 × 16 patch, moved with a stride of 8 pixels, is replaced by the mean intensity, averaged over overlapping patches and clipped at zero. [1], [2] Two reference maps were included. A Sobel edge map (gradient magnitude of the image) uses no information from the model at all; any method that does not beat it cannot be said to reveal what the model has learned, because a defect is usually a local change in intensity and therefore an edge. A uniform random map gives the chance level of each metric.

3.4 Localisation Metrics

For every defective test image with a non-empty mask G, each saliency map s was rescaled to [0, 1] and compared with the mask using three complementary metrics: Pointing game [22]: a hit if the pixel with the highest saliency falls inside the mask, dilated by 7 pixels to tolerate annotation and resizing error. Energy inside the mask [15]: the fraction of total saliency mass that falls on defect pixels (Equation 3). A random map scores, on average, the relative mask area, so the ratio of energy to area was also reported. Pixel AUROC: the area under the ROC curve obtained when the saliency value is used to separate defect from non-defect pixels; 0.5 is chance and 1.0 perfect ranking [3].

3.5 Faithfulness and Defect-Removal Tests

Localisation measures agreement with a human annotation, not whether the model actually used the highlighted pixels. Faithfulness was therefore assessed by deletion [23], [24]: for each correctly detected defective image, the 1, 2, 5, 10, 20 and 50% most salient pixels were replaced by the corresponding pixels of a Gaussian-blurred copy of the image ($\sigma = 5$), and P(defect) was recomputed. The same number of randomly chosen pixels was replaced as a control. Outcomes were the average drop in probability over the deletion levels up to 20% [23], the drop at 5%, and the proportion of detections that were reversed (P(defect) < 0.5).

Because the ground-truth mask is available, a stronger test of what the model relies on is also possible. The annotated defect, dilated by 3 pixels, was removed from each correctly detected image by inpainting with the Telea algorithm (OpenCV) and, separately, by blurring. If the model relied on the defect,

removing it should reverse the detection. As a control for artefacts introduced by the edit itself, the same operation was applied to correctly classified defect-free images, using defect masks drawn at random from the test set, and the whole defect-free image was also blurred.

3.6 Sanity Check and Stability

Model sensitivity was tested with the cascading randomisation sanity check of [11]. Starting from the trained seed-0 model, the weights of the output layer and then of convolutional blocks 4, 3, 2 and 1 were re-initialised cumulatively, and saliency maps were recomputed after each step for 57 correctly detected defective test images. Similarity to the original map was measured by Spearman's rank correlation ρ; a method that is sensitive to the model should give ρ close to zero once the learned weights have been destroyed. Localisation metrics were also recomputed for the randomised models. Stability across models was assessed by training two further networks on exactly the same seed-0 training and validation data with different initialisations (seeds 101 and 102) and computing ρ between the maps of the three models for the same images.

3.7 Statistics

Detection metrics are reported as mean ± standard deviation over the five splits, and localisation metrics as the mean over all image–split instances with the standard deviation of the five split means. For paired comparisons between methods, metrics were averaged per distinct image and compared with the two-sided Wilcoxon signed-rank test [31]. Given the number of comparisons, p-values below 0.001 are treated as strong evidence, and effect sizes are emphasised over significance. All code was written in Python with PyTorch, scikit-learn, SciPy and OpenCV, and all experiments ran on a two-core CPU.

4. RESULTS AND DISCUSSION

4.1 Detection Performance

Table 2 summarises the detector. Over the five splits, it achieved a test AUROC of 0.870 ± 0.032, an accuracy of 0.804 ± 0.033 and a balanced accuracy of 0.769 ± 0.061. Recall of defective tiles was 0.685 ± 0.178 at a false-positive rate of 0.147 ± 0.079. Performance varied considerably between splits: the model trained on split 3 converged to a conservative operating point with a recall of only 0.385, whereas the others recalled 0.70–0.86. Uneven surfaces were the hardest defect type (mean recall 0.543) and blowholes the easiest (0.791). This level of accuracy is below that of dedicated segmentation networks on the same data [16], but it is typical of a compact classifier and, importantly, the explanation analyses below are conditioned on whether each image was detected.

Table 2. Defect-Detection Performance on the Held-Out Test Sets (269 Images per Split)

Split	Accuracy	Balanced Accuracy	AUROC	Recall (Defect)	False-Positive Rate
0	0.833	0.806	0.892	0.744	0.131
1	0.833	0.803	0.879	0.731	0.126
2	0.762	0.791	0.875	0.859	0.277
3	0.777	0.661	0.815	0.385	0.063
4	0.818	0.785	0.891	0.705	0.136
Mean ± SD	0.804 ± 0.033	0.769 ± 0.061	0.870 ± 0.032	0.685 ± 0.178	0.147 ± 0.079

4.2 Localisation against the Annotated Defect

Table 3 and Figure 3 compare the methods. Every model-based method beat the random map by a wide margin on the pointing game, and Integrated Gradients (0.621), Gradient (0.597) and Grad-CAM (0.596) placed their maximum on or near the defect in about 60% of images, roughly twice as often as the

model-free edge map (0.309; p < 10−10 for all three). The pixel AUROC, which considers the whole map rather than its peak, gave a less flattering picture. Grad-CAM ranked defect pixels best (0.786), clearly ahead of the edge map (0.729, p = 7 × 10−9) and of Integrated Gradients (0.751, p = 4 × 10−8). Gradient and Integrated Gradients were only marginally better than edges (p = 0.013 and 0.009), Gradient × Input was not distinguishable from them (p = 0.22), and SmoothGrad and Occlusion were significantly worse than a detector that knows nothing about the model (0.626 and 0.629). The energy metric showed that even the best maps put only 12–15% of their mass on the defect, compared with a mean mask area of 9.5%; for a typical small defect, however, the median energy-to-area ratio of Integrated Gradients and Grad-CAM was about 4.5, meaning that the defect received four to five times its share of saliency.

Table 3. Localisation Performance of Different Methods on Defective Test Images

Method	Pointing Game	P (vs. Edges)	Energy in Mask	Energy / Area (Median)	Pixel AUROC	P (vs. Edges)
Gradient	0.597 (0.116)	2 × 10−11	0.123 (0.017)	3.83	0.744 (0.020)	0.013
Gradient × Input	0.550 (0.117)	5 × 10−9	0.121 (0.016)	2.67	0.715 (0.023)	0.22
SmoothGrad	0.510 (0.030)	1 × 10−5	0.096 (0.010)	1.08	0.626 (0.014)	1 × 10−10 (worse)
Integrated Gradients	0.621 (0.059)	6 × 10−15	0.131 (0.011)	4.55	0.751 (0.020)	0.009
Grad-CAM	0.596 (0.044)	1 × 10−11	0.152 (0.033)	4.28	0.786 (0.012)	7 × 10−9
Occlusion	0.417 (0.056)	0.011	0.143 (0.020)	0.86	0.629 (0.028)	2 × 10−10 (worse)
Sobel edges (model-free)	0.309 (0.072)	—	0.109 (0.009)	1.96	0.729 (0.020)	—
Random	0.149 (0.041)	2 × 10−6 (worse)	0.095 (0.010)	1.00	0.499 (0.003)	5 × 10−38 (worse)

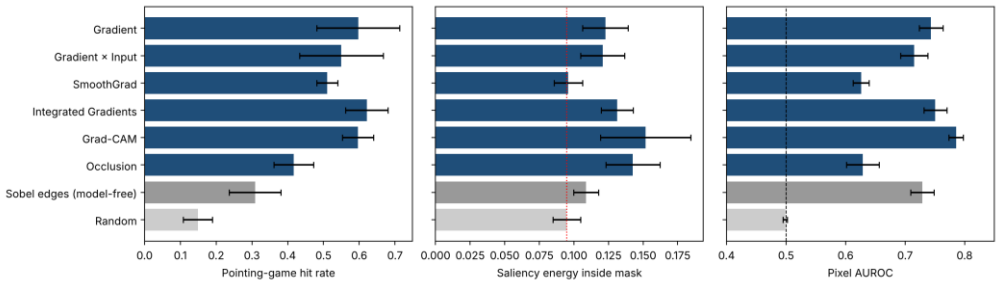

Figure 3. Comparison of Localisation Metrics across Different Methods

Error! Reference source not found. shows real test images for each defect type, chosen as the detected image of each class with the median pixel AUROC averaged over all methods, rather than as the most convincing example. For the small blowhole, Grad-CAM and Occlusion highlight the defect, and Gradient and Integrated Gradients concentrate near it. For the break, crack and uneven-surface examples, Grad-CAM places most of its mass along the image borders, and the gradient-based maps are diffuse; for the fray example, Grad-CAM highlights a printed number inside the defective region rather than the frayed edge. SmoothGrad produces nearly uniform maps across the tile, which explains its low AUROC.

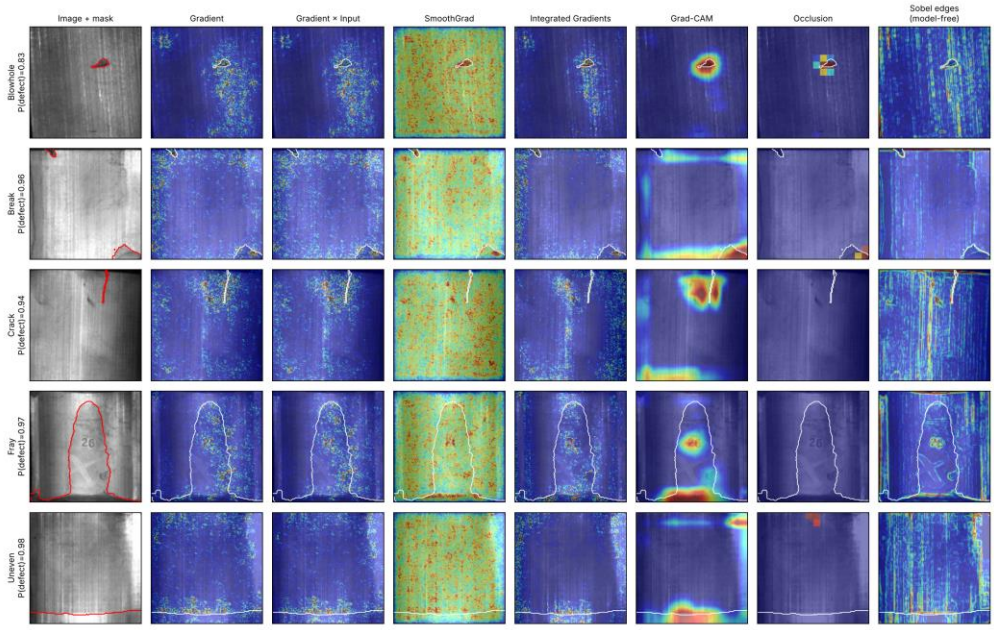

Figure 4. Saliency Map Visualization of Defect Detection on Held-Out Test Images

4.3 Localisation Depends on the Defect Type

Localisation was strongly dependent on the type of defect Figure 4. For small, compact blowholes, all model-based methods except Occlusion reached a pixel AUROC of 0.88–0.97, and Grad-CAM and Integrated Gradients pointed at the defect in 72–74% of cases. For cracks, the pointing-game hit rate dropped to 0.36–0.51, although the AUROC of most methods remained between 0.76 and 0.83, because thin cracks are surrounded by many equally bright texture lines. For the large, diffuse defects, localisation was close to chance: on uneven surfaces the AUROC of every method lay between 0.47 and 0.57, and on fray between 0.51 and 0.72. On these classes, the pointing game is also inflated by the size of the mask: a random map hit an uneven-surface mask in 41% of cases. Defect types that are hardest to detect were therefore also those whose explanations were least informative.

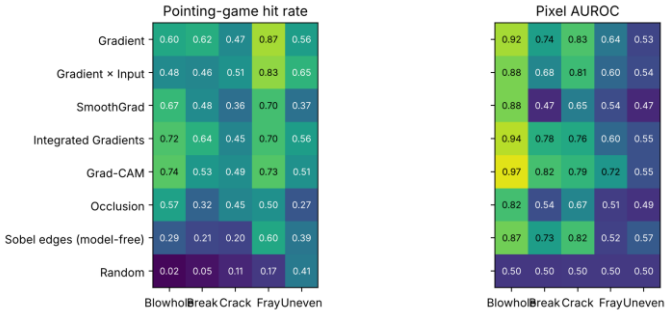

Figure 4. Pointing-Game Hit Rate (Left) and Pixel AUROC (Right) by Method and Defect Type, Pooled over Five Splits

4.4 Faithfulness: Deleting Salient Pixels

Deleting the most salient pixels lowered the detector's confidence much faster than deleting random pixels Table 4, Figure 5. Replacing just 2% of the image according to Gradient or Gradient × Input reversed 27–30% of correct detections, and replacing 5% reversed 34–35%, whereas random deletion of the same number of pixels reversed only 1–2%. Every method, including the model-free edge map, outperformed random deletion ($p < 10–19$). The ranking differed from that for localisation: the simple Gradient was the most faithful method (AOPC 0.174), Integrated Gradients and Grad-CAM were intermediate (0.150), and Occlusion, whose patches were computed with a different perturbation, was the

least faithful model-based method (0.093), only slightly ahead of edges (0.072). The curves in Figure 5 do not decrease monotonically: beyond 20% deletion, the mean P(defect) increased again for all methods, which points to an artefact of the blurring perturbation that is examined next.

Table 4. Deletion-Based Faithfulness Evaluation of Defect Detection Methods

Method	AOPC	Drop at 5%	Reversed at 2% (%)	Reversed at 5% (%)	Reversed at 10% (%)
Gradient	0.174	0.186	30.2	35.1	35.8
Gradient × Input	0.167	0.180	27.2	34.3	36.2
SmoothGrad	0.127	0.122	11.7	17.7	27.5
Integrated Gradients	0.150	0.152	23.8	26.4	31.7
Grad-CAM	0.150	0.159	21.5	29.1	28.7
Occlusion	0.093	0.104	12.8	18.1	20.4
Sobel edges (model-free)	0.072	0.074	11.7	14.3	17.7
Random deletion	≈0	≈0	0.4–1.1	1.1–2.3	4.2–5.7

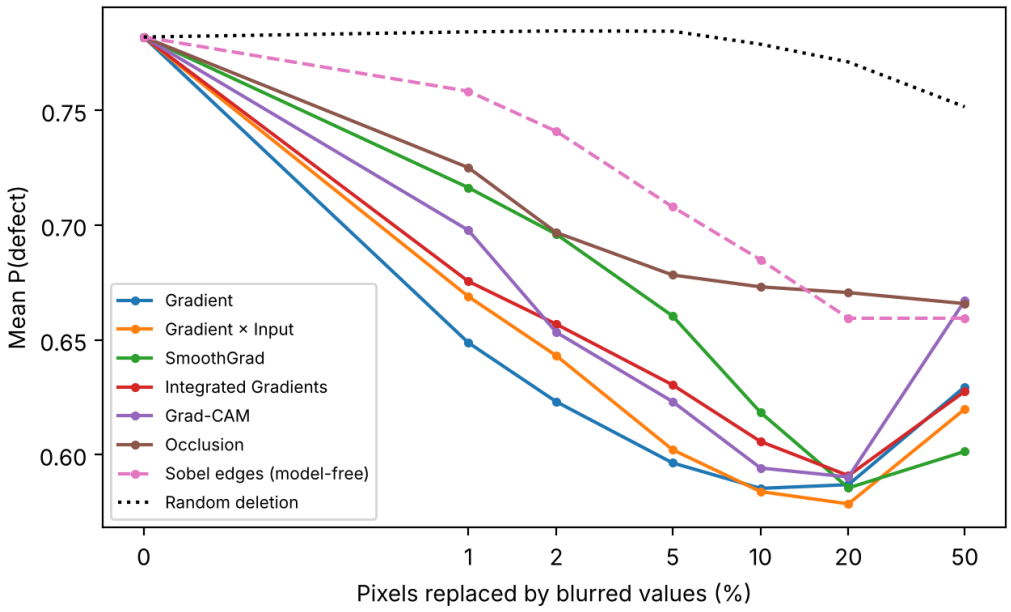

Figure 5. Deletion-Based Faithfulness Analysis of Defect Detection Methods

4.5 Removing the Defect Does Not Remove the Decision

The defect-removal test Table 5 produced the most important result of the study. When the annotated defect was removed from correctly detected images by inpainting, 71.7% of them were still classified as defective, and the median P(defect) fell only from 0.80 to 0.59. Blurring the defect gave almost the same result (72.5%). The same inpainting edit applied to correctly classified defect-free tiles at a random defect-shaped location produced a false detection in only 10.4% of cases, so the persistence of the decision cannot be explained by the edit alone. The effect varied by defect type: removing fray reversed 61% of detections, but removing an uneven-surface region reversed only 5%, and removing a crack only

21%. For most detections, the model therefore relied substantially on evidence outside the annotated defect.

Table 5. Effect of Removing the Annotated Defect (Dilated by 3 Pixels) from Correctly Detected Images, and Control Edits on Correctly Classified Defect-Free Images (Five Splits)

Images	N	Edit	Still Classified Defective (%)	Median Drop in P(defect)
Blowhole	91	Defect inpainted	65	0.23
Break	54	Defect inpainted	69	0.09
Crack	42	Defect inpainted	79	0.10
Fray	23	Defect inpainted	39	0.35
Uneven	55	Defect inpainted	95	0.02
All defective	265	Defect inpainted	71.7	0.12
All defective	265	Defect blurred	72.5	—
Defect-free (control)	815	Random defect-shaped region inpainted	10.4 (false alarms)	—
Defect-free (control)	815	Random defect-shaped region blurred	12.0 (false alarms)	—
Defect-free (all test)	955	Whole image blurred	94.6 (false alarms)	—

Two further observations help explain this behaviour. First, blurring an entire defect-free image caused 94.6% of them to be classified as defective, which shows that the detector treats loss of the fine vertical grinding texture as evidence of a defect. This also explains the rebound of the deletion curves in Figure 5: deleting a large fraction of pixels with blurred values creates exactly that cue. Second, image size alone carried information about the label: a random forest trained only on the original width and height of each image separated defective from defect-free tiles with a cross-validated AUROC of 0.727. Because all images were resized to a square, the aspect ratio becomes visible to the network as a stretching of the texture, which is a plausible shortcut [10]. Strikingly, the saliency maps did not reveal this reliance. They pointed at the defect about as often when the model missed it as when it detected it (pointing game for missed vs. detected defects: Gradient 0.68 vs. 0.56, Integrated Gradients 0.63 vs. 0.62, Grad-CAM 0.55 vs. 0.62), indicating that a defect-centred map reflects the conspicuousness of the defect in the image at least as much as the model's reasons for its decision.

4.6 Sanity Check: Do the Maps Depend on the Model?

Figure 6 and Table 6 show the cascading randomisation results. Gradient, Gradient × Input and SmoothGrad passed the test: once the last convolutional block was randomised, their rank correlation with the original maps fell to 0.20–0.29 and approached zero (0.01–0.18) when all layers had been randomised, and their localisation dropped to near chance. Integrated Gradients, in contrast, retained a correlation of 0.44 with the original map after every layer had been randomised, and the fully random network still achieved a pixel AUROC of 0.57; its maps are partly determined by the image term $(x - x')$ in Equation 1 rather than by the model [11], [18]. Figure 7 illustrates these patterns for one image. Grad-CAM and Occlusion responded strongly to randomisation of the output layer ($\rho = -0.25$ and -0.20; with two output classes, a random head can reverse the sign of the defect evidence). After the fourth block had also been randomised, however, Grad-CAM recovered a pixel AUROC of 0.76 and a pointing-game hit rate of 0.56

using random high-level features on top of trained lower layers, which shows that good localisation can be obtained from a network that has no trained decision layers. The edge map is, by construction, unaffected by the model (ρ = 1).

Table 6. Cascading Randomisation Sanity Check: Map Similarity and Pixel AUROC

Method	P: + Random fc	P: + conv4	P: + conv3	P: + conv2	P: + conv1 (All)	AUROC: + conv4	AUROC: All Random
Gradient	0.63	0.20	0.12	0.07	0.01	0.56	0.50
Gradient × Input	0.66	0.23	0.18	0.15	0.10	0.50	0.44
SmoothGrad	0.94	0.29	0.35	0.23	0.18	0.58	0.44
Integrated Gradients	0.74	0.53	0.48	0.46	0.44	0.63	0.57
Grad-CAM	−0.25	0.35	0.15	0.11	0.06	0.76	0.55
Occlusion	−0.20	0.22	0.21	0.20	−0.01	0.65	0.53
Sobel edges (model-free)	1.00	1.00	1.00	1.00	1.00	0.74	0.74

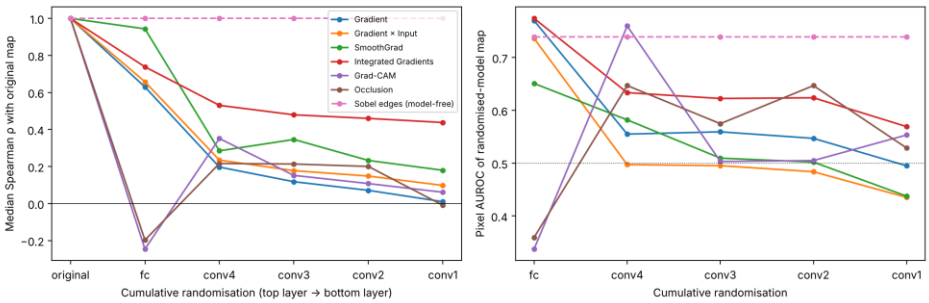

Figure 6. Cascading Randomization Sanity Check for Saliency Map Reliability and Defect Detection Performance

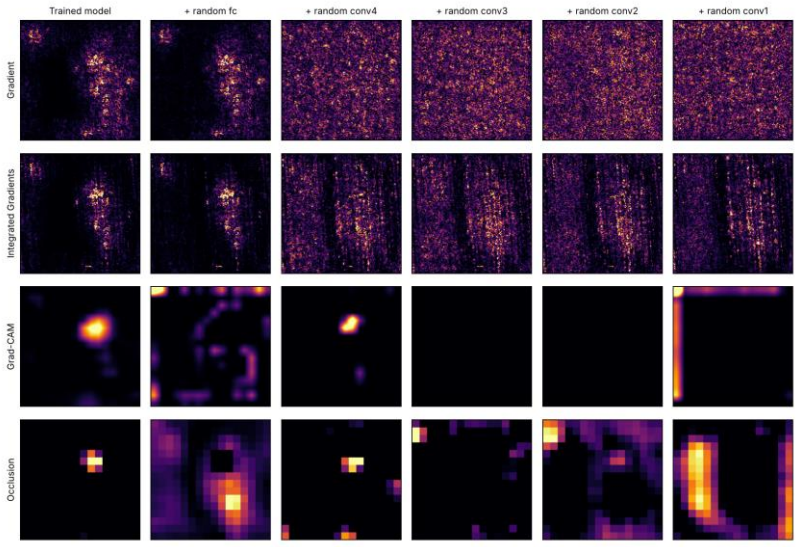

Figure 7. Saliency Map Visualisation during Progressive Model Randomisation

4.7 Stability across Identically Trained Models

Three networks trained on identical data with different random initialisations reached similar accuracy (test AUROC 0.892, 0.887 and 0.870 on split 0), but their explanations agreed only partly Table 7. Pixel-wise rank agreement between the models was low for Gradient (median ρ = 0.21) and Gradient × Input (0.24), moderate for SmoothGrad (0.42), Integrated Gradients (0.54) and Occlusion (0.59), and highest for Grad-CAM (0.71). Part of the apparent stability of Integrated Gradients is image-driven, since its maps from a fully randomised network still correlated at 0.44 with those of the trained one (Section 4.6). An inspector who receives a gradient map from one of these equally accurate models would therefore see a substantially different explanation from one produced by another.

Table 7. Agreement of Saliency Maps Between Three Models Trained on the Same Data with Different Initialisations (57 Images × 3 Model Pairs).

Method	Median ρ	Interquartile Range
Gradient	0.21	0.16–0.26
Gradient × Input	0.24	0.21–0.29
SmoothGrad	0.42	0.40–0.50
Integrated Gradients	0.54	0.50–0.58
Grad-CAM	0.71	0.56–0.80
Occlusion	0.59	0.36–0.71
Sobel edges (model-free)	1.00	1.00–1.00

4.8 Discussion

4.8.1 Principal Findings

This study evaluated six saliency methods for a defect detector on real industrial images with pixel-level ground truth, and found that no method was satisfactory on all criteria Table 8. Grad-CAM localised defects best and was the most stable across models, but it failed for large and diffuse defects, it was drawn to image borders, and it could localise defects with randomised high-level layers. Gradient was the most faithful in deletion and passed the sanity check, but it was the least stable between equally accurate models. Integrated Gradients pointed at defects most often but partly failed the sanity check. SmoothGrad and Occlusion, as configured here, did not beat a model-free edge detector on pixel AUROC. These results are consistent with earlier findings on natural and medical images [11], [12], [13], [14] and extend them to industrial inspection.

Table 8. Qualitative Summary of the Methods across the Four Evaluation Criteria (✓ Satisfactory, ~ Partly, ✗ Unsatisfactory, Relative to the Edge and Random Baselines)

Method	Localisation	Faithfulness (Deletion)	Sanity Check	Cross-Model Stability
Gradient	✓	✓	✓	✗
Gradient × Input	~	✓	✓	✗
SmoothGrad	✗	~	✓	~
Integrated Gradients	✓	~	✗	~
Grad-CAM	✓	~	~	✓
Occlusion	✗	~	~	~

The most consequential finding, however, concerns the model rather than the explanations. Removing the annotated defect left about 72% of detections unchanged, the detector classified blurred defect-free tiles as defective, and image size alone predicted the label with an AUROC of 0.73. The model had therefore learned cues that are correlated with defects in this dataset, such as loss of texture, image geometry and borders, rather than only the defects themselves. The saliency maps were not able to reveal this: on the median image they looked plausible, they pointed at the defect in about 60% of images, and they did so as often for missed defects as for detected ones. This is a textbook example of a "Clever Hans" predictor [10] that standard explanations make look trustworthy.

4.8.2 Why Plausible Maps Can Be Misleading

Three mechanisms explain the gap between what the maps show and what the model uses. First, small defects such as blowholes and cracks are high-contrast local structures, and gradient-based maps are partly driven by the input itself, so they tend to light up on such structures whether or not the decision depends on them; the edge map shows how much of this can be achieved with no model at all. Second, metrics that reward the peak of the map, such as the pointing game, can be satisfied by a small part of the map while the rest of the evidence lies elsewhere; for large masks they are also inflated by chance. Third, deletion with a fill value that the model itself regards as abnormal, here blur, conflates removing evidence with adding new evidence [25], so faithfulness scores must be interpreted with a control for the perturbation, as done here with defect-free images.

4.8.3 Implications for Safety-Critical Inspection

Several recommendations follow for engineers and auditors who intend to use saliency maps to support inspection decisions: Do not accept a saliency map as evidence that a decision was based on the defect. A plausible map on a typical image says little about the model's reasons. Where defect masks are available, report localisation against them with several metrics and always against a model-free edge baseline and a random baseline. Test the model, not only the explanation: remove the annotated defect (and apply the same edit to defect-free parts) to check whether the decision depends on the defect. This simple test exposed a shortcut that none of the maps revealed. Run cascading randomisation sanity checks and report cross-model stability; methods whose maps survive weight randomisation or change between equally accurate models should not be used for auditing. Control for dataset artefacts such as image size, aspect ratio, borders and acquisition settings before training, for example by cropping fixed-size patches rather than resizing whole images. Evaluate per defect type: explanations for large, diffuse defects such as uneven surfaces were close to chance, and these were also the hardest defects to detect. These recommendations are consistent with the expectation, expressed in recent AI regulation, that high-risk systems are transparent and that their outputs can be interpreted by the people who oversee them [4], transparency cannot be achieved with explanations that have not themselves been validated [5], [6].

4.8.4 Limitations

The study has several limitations. Only one dataset and one compact CNN were used, and the detector was considerably less accurate than state-of-the-art segmentation networks for this data; a more accurate model might rely more on the defects, and the defect-removal test should be repeated with stronger architectures and with fixed-size crops that remove the aspect-ratio cue. The test sets of the five splits overlap, so the split-level standard deviations underestimate the variability of fully independent experiments. Each method was used with one set of hyperparameters (for example, the noise level of SmoothGrad, the baseline of Integrated Gradients and the patch size of Occlusion), which may not be optimal. The sanity check and stability analysis were limited to one split and 57 images. Inpainting removes the defect only approximately, especially for large masks, and the 3-pixel dilation may not capture the full extent of thin cracks. Finally, the study did not include human inspectors, so it cannot say how the observed failure modes would affect real decisions.

5. CONCLUSION

This study asked whether saliency maps of a CNN defect detector point at the defect, using real images of magnetic tiles with pixel-level annotations. Integrated Gradients, Gradient and Grad-CAM pointed at the defect in about 60% of held-out defective images, twice as often as a model-free edge detector, and Grad-CAM gave the best pixel-level ranking of defect pixels; SmoothGrad and Occlusion did not beat the edge detector. Deletion confirmed that the salient pixels mattered for the model's confidence, and Gradient-based maps passed the randomisation sanity check while Integrated Gradients partly failed it. Yet removing the defect itself left most detections unchanged, the model treated texture loss and image geometry as defects, and the maps looked equally convincing when the model was right and when it was wrong. In safety-critical inspection, a saliency map should therefore be treated as a hypothesis about the model, not as evidence for a decision. Future work will repeat the analysis with stronger detectors, patch-based training without geometric shortcuts, other public defect datasets and a user study with quality inspectors.

Acknowledgments

The authors thank Y. Huang, C. Qiu and K. Yuan for making the Magnetic Tile Surface Defect dataset publicly available, the developers of PyTorch, scikit-learn, SciPy and OpenCV, and the researchers whose work is cited in this article.

Funding Information

This research has received no external funding.

Author Contribution Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Heri Nurdiyanto	✓	✓	✓		✓				✓		✓			
Nuno Alexandre Soares Domingues		✓		✓		✓		✓		✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

No conflict of interest.

Informed Consent

Not Applicable.

Ethical Approval

Not Applicable.

Data Availability

This study uses the publicly available Magnetic Tile Surface Defect dataset [27]. No new datasets were created. The analysis code, trained model weights and per-image results are available from the corresponding author on reasonable request.

REFERENCES

[1] T. Czimmermann, G. Ciuti, M. Milazzo, M. Chiurazzi, S. Roccella, C. M. Oddo, and P. Dario, "Visual-based defect detection and classification approaches for industrial applications—A survey," Sensors, vol. 20, no. 5, Art. no. 1459, 2020. doi.org/10.3390/s20051459

[2] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, "Segmentation-based deep-learning approach for surface-defect detection," Journal of Intelligent Manufacturing, vol. 31, no. 3, pp. 759–776, 2020. doi.org/10.1007/s10845-019-01476-x

[3] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9584–9592. doi.org/10.1109/CVPR.2019.00982

[4] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, L series, 12 Jul. 2024. eur-lex.europa.eu/eli/reg/2024/1689/oj

[5] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019. doi.org/10.1038/s42256-019-0048-x

[6] A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020. doi.org/10.1016/j.inffus.2019.12.012

[7] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021. doi.org/10.1109/TNNLS.2020.3027314

[8] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in Proc. Int. Conf. Learning Representations (ICLR) Workshop, 2014, arXiv:1312.6034. arxiv.org/abs/1312.6034

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626. doi.org/10.1109/ICCV.2017.74

[10] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," Nature Communications, vol. 10, Art. no. 1096, 2019. doi.org/10.1038/s41467-019-08987-4

[11] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, 2018, pp. 9505–9515. arxiv.org/abs/1810.03292

[12] P.-J. Kindermans et al., "The (un)reliability of saliency methods," in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, LNCS vol. 11700, Springer, 2019, pp. 267–280. doi.org/10.1007/978-3-030-28954-6_14

[13] A. Saporta et al., "Benchmarking saliency methods for chest X-ray interpretation," Nature Machine Intelligence, vol. 4, no. 10, pp. 867–878, 2022. doi.org/10.1038/s42256-022-00536-x

[14] N. Arun et al., "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," Radiology: Artificial Intelligence, vol. 3, no. 6, Art. no. e200267, 2021. doi.org/10.1148/ryai.2021200267

[15] L. Arras, A. Osman, and W. Samek, "CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations," Information Fusion, vol. 81, pp. 14–40, 2022. doi.org/10.1016/j.inffus.2021.11.008

[16] Y. Huang, C. Qiu, and K. Yuan, "Surface defect saliency of magnetic tile," The Visual Computer, vol. 36, no. 1, pp. 85–96, 2020. doi.org/10.1007/s00371-018-1588-5

[17] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in Proc. 34th Int. Conf. Machine Learning (ICML), PMLR vol. 70, 2017, pp. 3145–3153. proceedings.mlr.press/v70/shrikumar17a.html

[18] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in Proc. Int. Conf. Learning Representations (ICLR), 2018, arXiv:1711.06104. arxiv.org/abs/1711.06104

[19] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," arXiv:1706.03825, 2017. arxiv.org/abs/1706.03825

[20] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. 34th Int. Conf. Machine Learning (ICML), PMLR vol. 70, 2017, pp. 3319–3328. proceedings.mlr.press/v70/sundararajan17a.html

[21] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Computer Vision – ECCV 2014, LNCS vol. 8689, Springer, 2014, pp. 818–833. doi.org/10.1007/978-3-319-10590-1_53

[22] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," International Journal of Computer Vision, vol. 126, no. 10, pp. 1084–1102, 2018. doi.org/10.1007/s11263-017-1059-x

[23] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 11, pp. 2660–2673, 2017. doi.org/10.1109/TNNLS.2016.2599820

[24] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in Proc. British Machine Vision Conf. (BMVC), 2018, arXiv:1806.07421. arxiv.org/abs/1806.07421

[25] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019, arXiv:1806.10758. arxiv.org/abs/1806.10758

[26] A. Hedström et al., "Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond," Journal of Machine Learning Research, vol. 24, no. 34, pp. 1–11, 2023. jmlr.org/papers/v24/22-0142.html

[27] Y. Huang, "Magnetic-tile-defect-datasets," GitHub repository. Accessed: Oct. 2026. github.com/abin24/Magnetic-tile-defect-datasets

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. 32nd Int. Conf. Machine Learning (ICML), PMLR vol. 37, 2015, pp. 448–456. proceedings.mlr.press/v37/ioffe15.html

[29] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019, pp. 8024–8035. arxiv.org/abs/1912.01703

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learning Representations (ICLR), 2015, arXiv:1412.6980. arxiv.org/abs/1412.6980

[31] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bulletin, vol. 1, no. 6, pp. 80–83, 1945. doi.org/10.2307/3001968

How to Cite: Heri Nurdiyanto, Nuno Alexandre Soares Domingues. (2026). Do saliency maps find the defect? A ground-truth evaluation of explainable AI for safety-critical magnetic tile inspection. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 6(2), 56-72. https://doi.org/10.55529/jaimlnn.62.56.72

BIOGRAPHIES OF AUTHORS

	Heri Nurdiyanto, is a member of the Department of Industrial Engineering, Faculty of Engineering, Universitas Negeri Yogyakarta and Yogyakarta, Indonesia. His work lies in industrial and manufacturing engineering, including automated inspection, quality control and the forecasting of demand and production quantities. He is interested in how machine learning and decision-support methods are put to work on the factory floor and in the practical constraints that industrial deployment imposes. Email: herinurdiyanto@uny.ac.id
	Nuno Alexandre Soares Domingues, is an Adjunct Professor in the Department of Mechanical Engineering at Instituto Superior de Engenharia de Lisboa (ISEL), Instituto Politécnico de Lisboa and Lisbon, Portugal. His research addresses mechanical and energy engineering, including turbomachinery, energy systems and the condition and performance of engineering assets. He is interested in the maintenance and reliability of such assets and in how data-driven methods inform decisions about their remaining service life.