Evaluating Explainable AI. A Comparative Study of SENN, IG, and LIME

by Alessandra Gandini and Gaudenzia Genoni
GitHub repository: https://github.com/Ggenoni/SENN.git

Abstract
This study compares the intelligibility and faithfulness of explanations from a self-explainable neural network (SENN) and two post-hoc methods—Integrated Gradients (IG) and LIME—on MNIST and on a confounded MNIST dataset. Through a primarily qualitative analysis, supported by quantitative measures, we show that SENN fails to provide meaningful explanations, while IG and LIME offer more faithful and interpretable attributions. IG and LIME evaluations on the Confounded MNIST reveal the typical Clever Hans effect.

1 Introduction

As AI advances, machine learning algorithms are increasingly deployed in high-stakes domains such as healthcare and finance: since the opacity of deep learning models—often termed "black boxes"—raises critical concerns about fairness, bias, safety, and accountability, Explainable AI (XAI) has emerged as a research field to provide insights into model behavior and the reasoning behind its predictions.

In this study, we assess the intelligibility and faithfulness of explanations from a self-explainable neural network (SENN), proposed in 2018 by Alvarez-Melis and Jaakkola[1], and compare them to attributions obtained from the same model using Integrated Gradients (IG) and Local Interpretable Model-agnostic Explanations (LIME). While IG, introduced by Sundararajan et al.[6], and LIME, presented by Ribeiro et al.[5], are post-hoc methods that explain individual predictions by attributing importance to input features—IG computes feature attributions by integrating the gradients of the model's output with respect to the input along a straight-line path from a baseline to the given input, whereas LIME perturbs the input, evaluates the model's responses, and trains a simple interpretable surrogate model to approximate local feature importance—, SENN provides ante-hoc, concept-based explanations. Here, in particular, we use the SENN implementation designed by Elbaghdadi et al.[2]: the model comprises a Parameterizer (a neural network) that assigns relevance scores to input features, a Conceptizer (an autoencoder) that maps inputs to a small set of interpretable basis concepts, and an Aggregator that combines these concepts and their relevance scores via a summation operator to produce the final prediction.

2 Experimental setup

2.1 Dataset

For our experiments, we use the MNIST dataset[4], chosen for its interpretability. It includes 60,000 training images (6,000 used for validation) and 10,000 test images, all grayscale 28×28 handwritten digits (0-9). Furthermore, to evaluate the model's explanations under the Clever Hans phenomenon[3], we use a confounded MNIST dataset, where a small cross is added at a fixed, class-specific margin in the training and validation images but placed at a position corresponding to a different class in the test set (Figure 1).

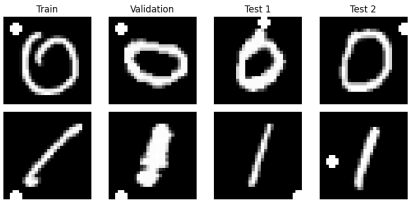


Figure 1: Samples of digits 0 and 1 from the Confounded MNIST dataset. In training and validation images, the confounder is placed at a fixed, class-specific position; in test set instances, it is randomly placed at one of the locations assigned to other classes.

2.2 Training

We train our model on both datasets (batch size: 200) for 40 epochs using the Adam optimizer with a learning rate of 2×10^{-4}: while MNIST typically converges within 10–20 epochs, our SENN model requires longer training than conventional classifiers to refine meaningful feature representations in the Conceptizer. To prevent overfitting, we apply regularization techniques such as dropout (0.5) and sparsity regularization (2×10^{-5}). We set the robustness loss regularization parameter (λ) to 1×10^{-4}, as it yields the best test accuracy (see Elbaghdadi et al. [2]). Finally, we select 5 as the number of concepts, to balance interpretability and expressiveness: future experiments could assess whether training with more concepts improves disentanglement or instead introduces redundancy in representation.

3 Analysis

In our analysis, which is primarily *qualitative* in nature, we compare the **intelligibility** of explanations from SENN, IG, and LIME on a set of ten images (one per class) randomly sampled from the MNIST test set, using fixed seeds for reproducibility. An identical, separate evaluation is carried out for Confounded MNIST. The observations are always complemented by *quantitative* metrics. For SENN, we assess **concept activations** $h(x)_i$, which represent interpretable features extracted from the input, and **relevance scores** $\theta(x)_i$, which quantify their contribution to the final prediction: together, the set $\{(h(x)_i, \theta(x)_i)\}_{i=1}^{k}$ defines the explanation of the model output, where k is the number of basis concepts. For IG, we evaluate the **completeness gap**, which measures how well attributions A_i approximate the difference in model predictions between a fully black baseline input x_0 and the actual input x: $(\sum_{i=1}^{d} A_i) - (f(x) - f(x_0))$, where d represents the dimensionality of the input space. For LIME, we measure the **explanation score** as the sum of the positive weights assigned to superpixels in the local linear regression surrogate model: $S = \sum_{i|w_i>0} w_i$, where w_i represents the importance weight of superpixel i.
Further experiments are also conducted on the ten MNIST samples to determine the **faithfulness** of the explanations, relying on a proxy notion of importance: to evaluate the accuracy of estimated feature relevances, we analyze the impact of removing features identified as most important by the explanation methods. For SENN, in particular, we perform an ablation study, setting multiple concept activations to zero to test whether the model relies on them for classification. For post-hoc methods, we progressively mask the top-ranked pixels for IG and the most influential superpixels for LIME, observing the impact on model's confidence and predictions.

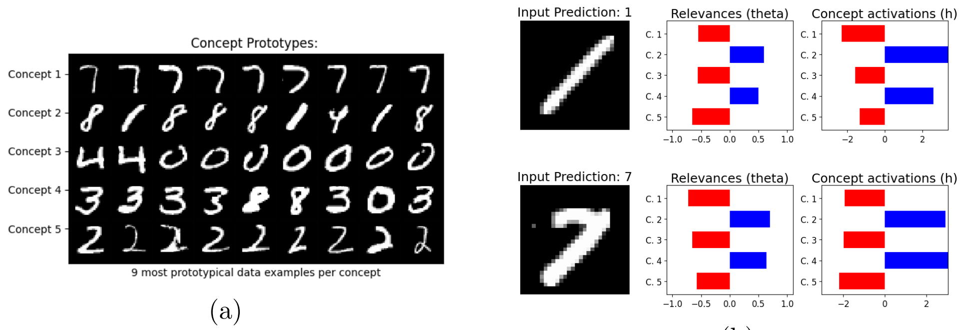
4 Results

Before evaluating explanations, we report the **model's performance**. On MNIST, it achieves a test accuracy of 98.9%. On Confounded MNIST, the test accuracy drops to 34%, while the validation accuracy is 100%: clearly, SENN relies on confounders for its predictions.

4.1 SENN explanations

4.1.1 Intelligibility

To qualitatively examine the interpretability of the explanations generated by the SENN model, we show the top nine prototypical test examples that most strongly activate a certain concept (Figure 2a). Although the prototypes for a given concept are not always of the same digit class (except for concepts 1 and 5, which align with digits 7 and 2, respectively), some patterns emerge: concept 2 appears to capture diagonal strokes, while concept 4 highlights loops, as seen in digits 3 and 8; concept 3 is harder to interpret, because it includes both squared and rounded shapes. Overall, however, the concepts do not appear to encode a single, consistent semantic meaning and remain only partially human-interpretable.
A further issue arises when analyzing relevance scores and concept activations for our ten samples: each one consistently presents positive values for concepts 2 and 4, while all the other concepts receive negative scores (Figure 2b). This behavior is clearly implausible and suggests that the model does not dynamically adjust relevance scores nor concept activations based on the input, thereby failing to provide intelligible and diverse concept-based explanations. No better results were obtained on the Confounded MNIST dataset.



(a) (b)

Figure 2: (a) Top nine most prototypical data examples per concept. (b) Concept-based explanations, showing results for digits 1 (top) and 7 (bottom): while in this case concept 2's high activation is understandable (given the diagonal orientation of the two digits), the positive score assigned to concept 4 lacks a clear interpretation. Additionally, concept 1 shows a negative activation for digit 7, which is counterintuitive.

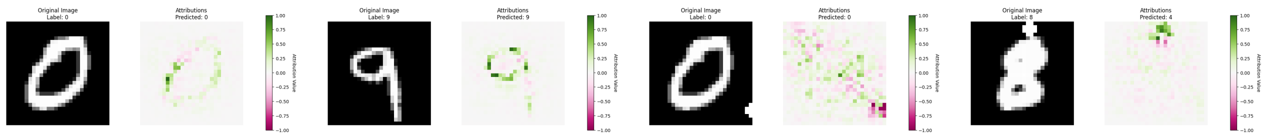
4.1.2 Faithfulness

The results of the ablation study are equally concerning: removing four out of five concepts has an almost negligible effect on classification outcomes, with the percentage of altered predictions remaining below 1%. In the absence of a performance drop, the model's decisions appear to be largely independent of the learned concepts, which do not fully capture the discriminative features actually used for classification.

4.2 IG explanations

4.2.1 Intelligibility

For all ten MNIST samples, IG attributions highlight important regions of the digit, which align with human-understandable features and support classification interpretability (Figure 3a-b). The completeness gap remains consistently low across all cases ($< 10^{-3}$), indicating that the attributions accurately reflect the model's decision-making process. In the case of Confounded MNIST, an interesting pattern can be observed: when predictions are incorrect, IG attributions assign positive importance to the confounder, whereas when predictions are correct the confounder receives negative attribution values (Figure 3c-d).

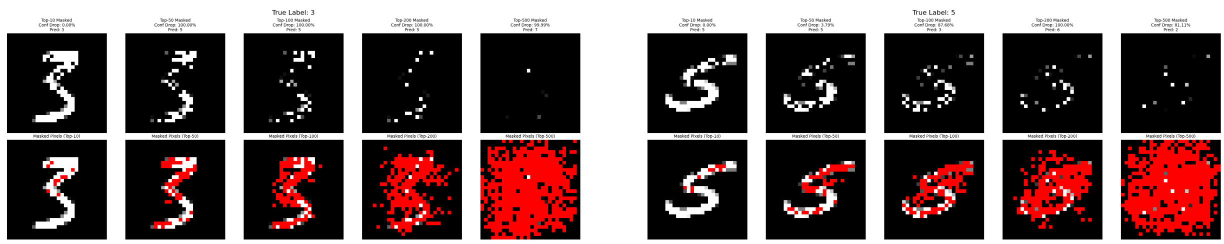


(a) (b) (c) (d)

Figure 3: IG attributions for MNIST samples: (a) digit 0 and (b) digit 9 highlight interpretable features. IG attributions for Confounded MNIST samples: (c) digit 0 is correctly classified, showing negative values on the confounder (though the attribution map lacks a distinct pattern), while (d) digit 8 is misclassified as 4 clearly due to the confounder effect.

4.2.2 Faithfulness

Masking the most important pixels identified by IG attributions generally has a significant impact the model's confidence and predictions. Even though masking only 10 pixels can leave the model's confidence unchanged, removing between 50 and 100 pixels causes a clear drop in confidence, often leading to misclassification; beyond 200 masked pixels, confidence typically falls to zero (Figure 4). This behavior suggests that, although the model does not rely solely on a small set of pixels, collectively the pixels identified by IG play a crucial role in classification, providing empirical support for the faithfulness of IG explanations.



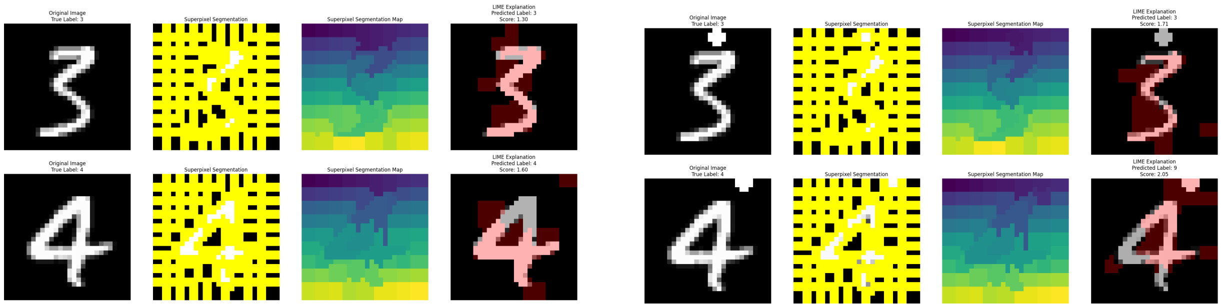
(a) (b)

Figure 4: IG faithfulness analysis on two MNIST samples: the first row shows the masked inputs, with the corresponding masked pixels highlighted in red below. (a) For digit 3, confidence drops completely when the 50 most relevant pixels identified by IG are masked, leading to a misclassification as 5. (b) For digit 5, confidence declines more gradually, possibly due to less accurate attributions.

4.3 LIME explanations

4.3.1 Intelligibility

Like IG, LIME explanations are highly interpretable: following segmentation into 100 regions via the SLIC algorithm, the highlighted superpixels largely correspond to crucial digit features, though some background regions are occasionally captured (Figure 5a). Furthermore, explanation scores consistently exceed 1, suggesting that LIME identifies strongly contributing superpixels: a sign that the model's predictions rely on well-defined patterns rather than noise. In Confounded MNIST, LIME frequently highlights the confounder when predictions are incorrect but ignores it when predictions are correct, mirroring IG behavior (Figure 5b).

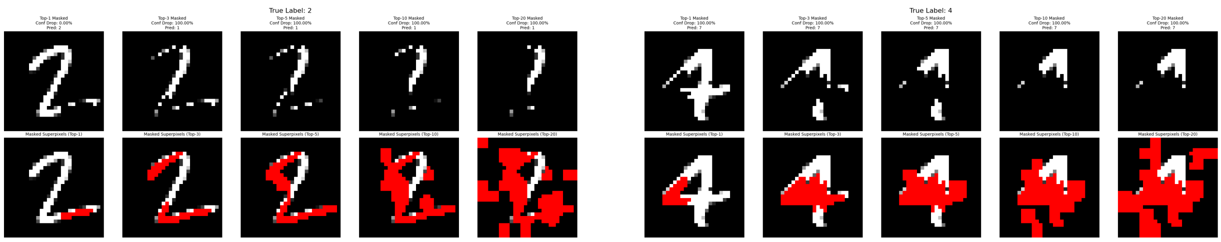


(a) (b)

Figure 5: (a) LIME attributions for MNIST: digit 3 (top) is fully covered, suggesting that almost every feature of the digit contributes to the model's recognition process; for digit 4 (bottom), on the other hand, only key regions are highlighted, leaving the top part uncovered (in both images, however, some background superpixels are also highlighted, which is suboptimal). (b) LIME attributions for Confounded MNIST: while the prediction for digit 3 (top) is strong enough to ignore the confounder effect, the model is misled by the confounder into classifying digit 4 as 9 (bottom). Superpixel segmentations are also shown (second and third columns): the input images are divided into 100 compact and visually coherent regions by the SLIC algorithm.

4.3.2 Faithfulness

The experiment on masked superpixels leads to similar conclusions as with IG: across our ten samples, removing the top 3–5 most important superpixels results in misclassification and a 100% confidence drop (Figure 6). This provides strong empirical support for the faithfulness of LIME explanations and indicates that LIME assigns importance to features that influence model decisions.



(a) (b)

Figure 6: LIME faithfulness analysis on two MNIST samples: the first row shows the masked inputs, with the corresponding masked suerpixels highlighted in red below. For both (a) digit 2 and (b) digit 4, masking the top 3 most relevant superpixels is sufficient to cause a 100% confidence drop and misclassification.

5 Conclusions

With regard to SENN, our findings suggest that although it achieves high classification accuracy on the MNIST test set, its self-produced explanations lack meaningful insights. At least for the implementation by Elbaghdadi et al.[2] and under our training specifications, the Parameterizer and Conceptizer appear to overemphasize certain concepts, leading to biases in concept activations and relevance scores. Furthermore, the ablation study confirms that the

model's decisions are largely independent of the learned concepts.
In contrast, IG and LIME provide more reliable explanations by directly analyzing the input–output relationship, and bypassing the model's internal relevance assignment. Their attributions align more intuitively with human expectations, and faithfulness evaluations confirm that the features they identify as relevant significantly impact the model's predictions.
Finally, by assigning positive attributions to the confounder in misclassified instances from the Confounded MNIST dataset, IG and LIME highlight the Clever Hans effect, where the model relies on a spurious feature rather than the actual digit structure.

References

[1] David Alvarez-Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pages 7775–7784, 2018.

[2] Omar Elbaghdadi, Aman Hussain, Christoph Hoenes, and Ivan Bardarov. Self explaining neural networks: A review. Technical report, University of Amsterdam, 2020. Available on GitHub. Last accessed on February 2, 2025.

[3] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1–8, 2019.

[4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, New York, NY, USA, 2016. ACM.

[6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint*, arXiv:1703.01365, 2017.