A New Paradigm for the Empirical Evaluation of the Understandability of LIME Visualizations

Marc Halbrügge

IU International University of Applied Sciences

Abstract

LIME is a method for creating explanations for classifiers that has accumulated thousands of academic references since its publication in 2016. The authors claim that by highlighting the most important features of an image, correct classifications can be distinguished from erroneous ones without any training in AI or LIME specifically. This study challenges that view based on new empirical data and a highly reproducible paradigm. Participants (n=108) were shown LIME visualizations comparing outputs from a well-trained (F_1=.944) and a bogus (F_1=.085) classifier, using randomly selected MNIST digits. By carefully balancing other influencing factors such as presentation order, visualization style, and correctness of the classification, it can be shown that for the given stimuli the general positive impact of a LIME visualization is negligible.

Keywords

Explainable AI, Understandability, Classification, LIME

1. Introduction and Related Work

Given the omnipresence of artificial intelligence (AI) and its impact on academia [1], work environments [2], and news media [3], understanding how an AI system reached its "conclusion" is more important than ever. Explainable AI (XAI) methods aim to provide this understanding [4, 5]. While many XAI techniques have been introduced in recent years, their overall suitability and the evaluation thereof remains an open problem [6]. This paper introduces a new paradigm for the empirical evaluation of image classification explainers and applies it to the frequently-used LIME method. The results highlight weaknesses in LIME that should be addressed in future research.

1.1. Local Interpretable Model-agnostic Explanations (LIME)

Ribeiro et al. [7] proposed LIME as a method to create explanations for any classifier, which are readable and understandable by non-experts. To achieve this goal, local perturbations are added to the input data, and changes in the output are observed. These changes are then used to judge how important the currently perturbed part of the input is for the classification. This procedure can be roughly compared to leverage statistics in linear regression, e.g., Cook's distance [8].

Previous evaluations of LIME have produced mixed results. Dieber & Kirrane performed a usability study using LIME explanations of tabular data with a small sample size (n=6) that yielded overall favorable results [9]. On the other hand, Knab et al. [10] state issues with LIME in five areas, namely locality, fidelity, interpretability, stability, and efficiency. At the same time, the authors highlight the lack of reproducible evaluation paradigms for LIME applications (c.f. [11]).

This paper provides such an evaluation method for the interpretability of LIME visualizations. The main research goal is to quantify how much information can be gained using LIME. A new experimental paradigm is presented for the evaluation of XAI techniques with the hope that this will improve the reproducibility in this new and important field of research.

1st Workshop on Explainability, Transparency, and Safety (ExTraSafe)

marc.halbruegge@acm.org (M. Halbrügge)

0009-0007-3208-1297 (M. Halbrügge)

2. Methods

Trust in an AI system depends in part on its perceived usefulness. In image classification, correctness therefore needs to be treated as an independent variable that influences trust. To measure and control this influence, we designed an experimental paradigm that compares participants' evaluations of actual LIME results with completely random yet believable distractors. The correctness of the underlying classifications was fixed at 50% for both the target classifier and the distractor. This design disentangles LIME's additional explanatory value from classification correctness.

The selection of stimuli can easily lead to implicit biases unknowingly applied by the experimenter. To avoid such biases, we scripted the full process and randomly selected suitable images from the well-known MNIST data set [12]. The source code was based on previous research [13]. The complete script together with the web application and data analysis is available for download at [14] and https://github.com/mahal-tu/xai-lime-demo.

2.1. Stimuli Generation

Verifying whether something counts as an *explanation* is difficult. We therefore used a cognitively simple task: participants saw two LIME visualizations of the same image, one from a well-trained target classifier and one from a bogus distractor classifier, and had to decide which classifier was better and more trustworthy.

The target classifier was a support vector machine (SVM) from Scikit-learn [15] using a radial basis function kernel trained on 3000 randomly selected MNIST digits. On the MNIST test set, it achieved an F_1 score of .944.

The distractor classifier was a linear SVM trained on 500 random MNIST images. Its labels were determined based on random linear combinations of a subset of the pixels. The linear combinations were sampled from a larger set such that they were mutually uncorrelated and also uncorrelated with the correct labels ($F_1 = .085$ on the MNIST test set). The rationale was to create a distractor that uses visual image features but remains completely uncorrelated with the ground truth. While it is semantically unconnected to the handwritten digits in the images, it should still produce meaningful LIME visualizations.

Both classifiers were applied to all MNIST digits, and images were selected where the two top predicted classes were the same for both classifiers[1] (in any order). There are four possible combinations of correct/incorrect classifications for the two classifiers, and 6 images were randomly selected for each of these combinations, resulting in 24 images in total. This procedure minimized selection bias and balanced classification correctness across trials.

As shown in [16, 17], LIME visualizations are highly dependent on the image segmentation algorithm used, especially for very small images like MNIST digits. To balance segmentation effects, we generated three segmentations based on the SLIC algorithm [18] using different parameters[2] and three regular grid segmentations.[3] The six resulting LIME weights per pixel were then averaged to create the final LIME visualization for each image and classifier.[4]

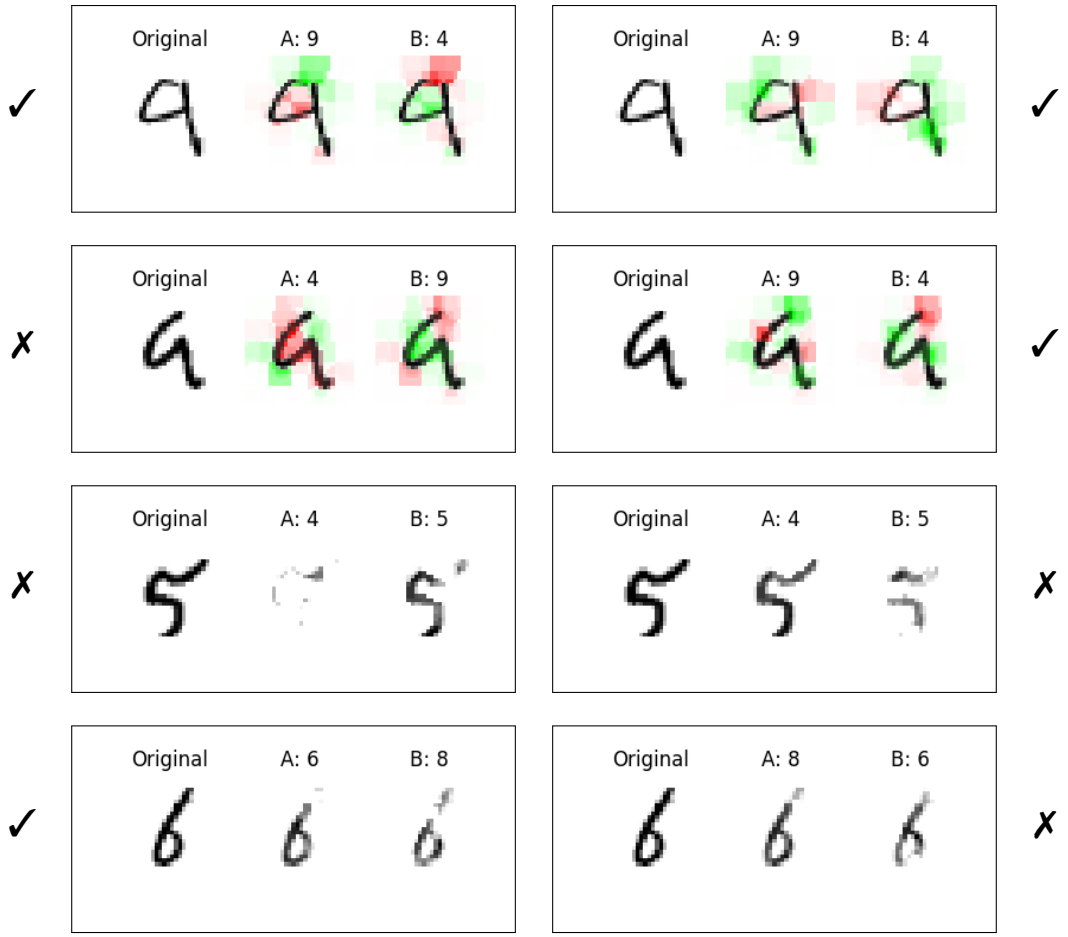
Two different visualization styles were used: (a) a heatmap style, where the LIME weights were shown as a heatmap overlay on the original image, and (b) a masked image style, where the pixels with the highest LIME weights were shown and the rest of the pixels were masked. The heatmap overlays showed the most important areas with a positive impact on the classification using a green overlay, while areas with a negative impact were shown with a red overlay.

[1] To achieve a minimum level of ambiguity, cases where the target classifier's top class probability was more than .8 higher than the follow-up probability were discarded.

[2] compactness in $\{0.01, 0.1, 1.0\}$

[3] 5×5, 7×7, and 11×11 grids

[4] The official LIME Python package (v0.2.0.1) was used to create the visualizations.



(a) Real Classifier (b) Bogus Classifier

Figure 1: LIME Visualizations of the four training MNIST images with the real and the bogus classifier. Color-coded heatmap style for image 1 (top) and 2, masked images style for image 3 and 4 (bottom). The "real" target classifier is correct on image 1 and 4 and wrong on image 2 and 3. The bogus distractor classifier is "correct" on image 1 and 2 and "wrong" on image 3 and 4.

2.2. Participants

Participants were recruited in two different ways: (1) flyers at the Bielefeld University mensa and during the KogWis 2025 cognitive science conference at Bochum, and (2) via e-mail to business students at the University of Applied Science Osnabrück/Lingen. The first group was expected to have a general academic background, with self-selection effects attracting participants with higher interest in machine learning and AI, while the second group was expected to represent a more general student population without hands-on experience with machine learning. Recruiting started in August 2025 and ended in September 2025.

2.3. Procedure

The experiment was implemented as a web application using PHP and JavaScript. Participants were asked to complete the experiment online on their personal devices. Depending on the browser language settings of the participants, the experiment was presented in either German or English.

The participants were informed that they were about to see "images of *handwritten digits* that two AI systems are trying to recognize" and that "one of the AIs is much better than the other." LIME was explained as showing "which areas of the images the AIs paid special attention to." The final instruction

was to "select *the better, more trustworthy AI*," which was inspired by the title of original LIME paper, "Why should I trust you?" [7].

At the beginning of the online session, participants were randomly assigned to either heatmaps or masked images for the entire session. Each trial consisted of a pair of LIME visualizations created for the same MNIST digit. For both the target and the distractor classifier, the top prediction was presented right beside the original image, followed by the second-best prediction. The top prediction was labeled "A" and the second-best was labeled "B". The images were presented in a random order, and the presentation order (above/below or left/right depending on screen size) of target and distractor was balanced across trials. Four training trials were shown at the beginning, where the participants received feedback on whether their choice was correct or not, followed by 20 trials without feedback. The training stimuli are given in Fig. 1 and an example of the complete user interface is given in Fig 2.

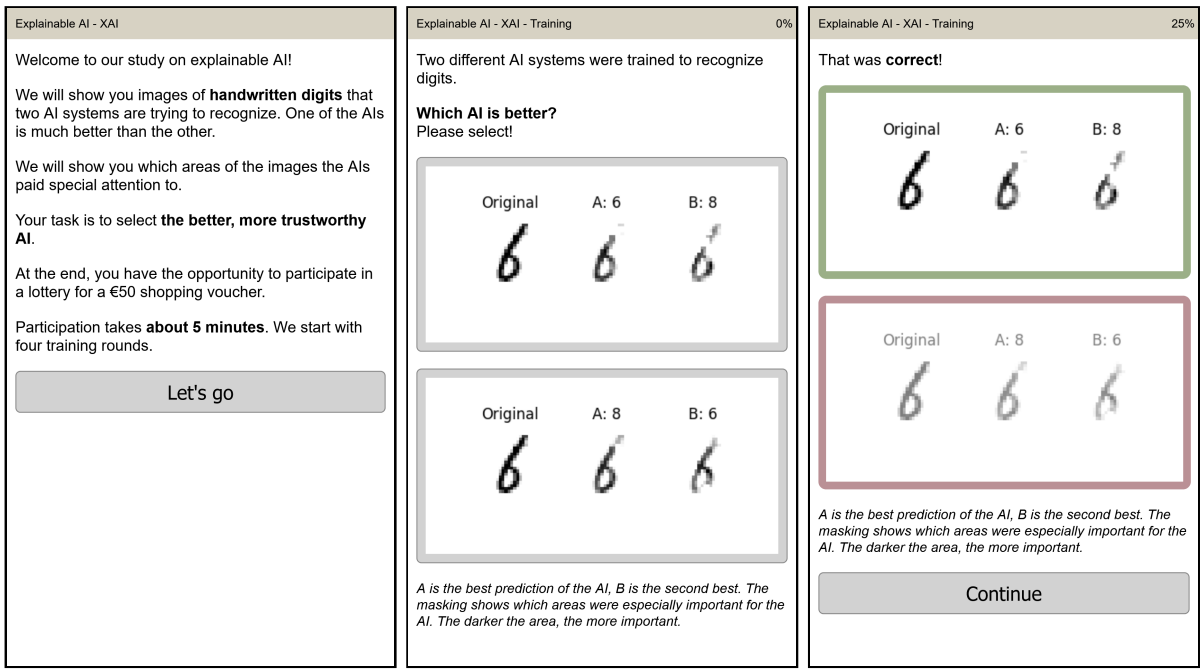


Figure 2: Example flow of the first three screens of the online study: Introduction, first training trial and feedback screen (only shown during the four training trials). On larger devices, e.g. desktop computers, the LIME visualizations were shown side-by-side, instead.

After the LIME section, participants completed a questionnaire about machine learning knowledge and some optional demographic questions. Machine learning knowledge was measured using seven multiple-choice questions about the meaning of acronyms like GPT, DNN, and PCA. Each question had four answer options, the correct answer and three distractors.

At the end, participants could enter a raffle for a gift voucher, which served as an incentive to complete the experiment. Participants could drop out at any time. A typical session lasted between 3.5 and 8.5 minutes, with a median time of 5.5 minutes.

2.4. Data Cleaning

Participants who showed a high statistical dependency between left and right choice from one trial to the next (e.g., always choosing the same side, or always switching sides) were excluded from the analysis, as were participants who used less than 2 seconds to make a decision ($n = 16$). Of the remaining participants ($n = 145$), 37 completed less than half of the 24 LIME trials and were also excluded from the analysis.

3. Results

After cleaning, 108 participants remained in the sample. 50 (46%) participants were business bachelor students and 58 (54%) were from the general academic audience that was addressed in the mensa and at a cognitive science conference. Descriptive demographic statistics about the two groups are provided in Tab. 1. Based on the web user agent reported by their browsers, most participants completed the experiment using a smartphone (57% iOS, 29% Android, 14% Desktop OSes).

Table 1
Demographics of Participants

Group	N	Mean Age	% Female	Median Education
Business Students	50	22.0 ($SD = 2.0$)	82.1%	High School
General Academics	58	28.5 ($SD = 11.6$)	40.5%	Bachelor

3.1. Detection of the Target Classifier

All 24 LIME image pairs went into the analysis. Generalized linear mixed models [19] with a logit link function were used to analyze the influence of the variables on the correctness of the participants' choices. Participant and trial numbers were included as random effects. Odds ratios (OR) are reported for the fixed effects.

The participants successfully detected the target classifier in 51.8% of the trials, which is indistinguishable from chance ($OR = 1.077, z = 1.042, p = .297$). The presentation order (left vs. right, $OR = 1.077, z = 0.779, p = .436$) did not have a significant influence. There was no significant main effect of the visualization style (masks vs. heatmaps, $OR = 1.007, z = 0.052, p = .959$), but interactions with faithfulness that will by analyzed in a follow-up paper.

Whether the target classifier was correct had a significant positive influence on the correctness of the participants' choices ($OR = 1.553, z = 3.29, p = .001$), while the correctness of the distractor classifier had a significant negative influence ($OR = 0.714, z = -2.52, p = .012$). The interaction between the two was not significant ($OR = 1.107, z = 0.537, p = .591$). Bootstrapped confidence intervals [20] of the correct choice for the four combinations of correct/incorrect classifications of the two classifiers are shown in Fig. 3.

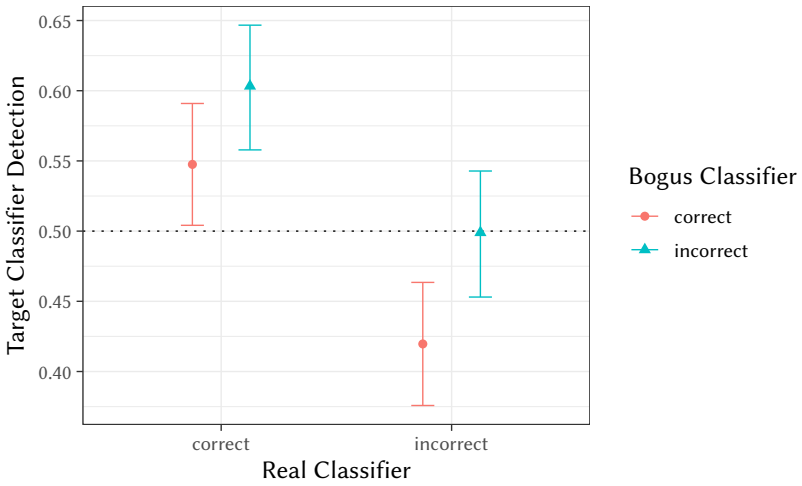


Figure 3: Correctness of Participants' Choices. Error bars are 95% confidence intervals. The dotted line at .5 denotes guessing level.

We estimated relationships between target detection and other covariates using each participant's relative frequency of correct choices. None of the covariates showed significant relationships with target detection; statistics are reported in Tab. 2.

3.2. Prior Machine Learning Knowledge Questionnaire

74 participants completed the machine learning knowledge questionnaire, with a mean score of 2.2 out of 7 ($SD = 2.4$). This is only slightly above the guessing level of 1.75, which is the expected score when randomly selecting one of the four answer options for each question. The knowledge questionnaire showed a sufficiently high internal consistency (Cronbach's $\alpha = .78$). There was no significant difference in machine learning knowledge between business students ($M = 2.13$) and general academics ($M = 2.27$), $t_{51.3} = -.23, p = .816$, but a small insignificant correlation between machine learning knowledge and highest education level (Spearman correlation, $r_s = .19, t_{72} = 1.64, p = .105$).

Table 2
Covariates of Target Detection

Variable	Correlation	Test Statistic	p	Partial η^2
Machine Learning Knowledge	$r = .114$	$t_{72} = 0.977$	.332	.013
Age	$r = -.033$	$t_{63} = -0.259$	.796	.001
Highest Education	$r_s = .126$	$t_{80} = 1.138$	.259	.016
Gender		$F_{2,67} = 0.731$	.485	.021
Group		$F_{1,80} = 1.974$	.164	.024
Device OS		$F_{3,78} = 0.302$	.824	.011
Browser Language		$F_{1,80} = 0.063$	.803	< .001

4. Discussion

LIME was first presented in 2016 and gained high popularity in the XAI community, with more than 32,000 citations on Google Scholar[5] at the time of writing. However, in this study, LIME-based explanations could not help distinguish a well-trained classifier from a completely bogus one. Participants' trust was driven more by prediction correctness than by the visual representation of which image regions were used for the prediction.

4.1. Why Could Ribeiro et al. not be Replicated?

Ribeiro et al. [7] presented a validation study based on a husky vs. wolf classifier, where the classifier was in fact using the presence of snow in the background as the most prominent feature. After being shown masked images like the ones in the lower part of Fig. 1, 27 graduate students with some background in machine learning were able to question the validity of the classifier, i.e., they changed their judgments compared to before seeing the LIME visualizations.

Why could this not be replicated with MNIST digits in the given study? Several possible reasons come to mind.

Choice of Stimuli Maybe MNIST digits are just too hard and therefore the participants were just guessing all the time? Or the application of image segmentation algorithms (as part of LIME) led to more believable LIME visualizations of the bogus classifier than technically justified?

Although the MNIST dataset is often used in comparable studies, e.g. [21], the empirical results may be specific to this kind of data. Additional studies with more complex stimuli are needed. Because the source code of the paradigm is available,[6] such studies should be straightforward .

[5] https://scholar.google.com/scholar?cluster=16724302923321127943
[6] https://github.com/mahal-tu/xai-lime-demo

Sampling Bias The Ribeiro study used hand-picked images as stimuli. This may have introduced implicit experimenter biases that influenced the observed results in [7]. In the given study, the balancing of correct/incorrect classifications has led to a relative emphasis of misclassified images, which may have reduced the participants' trust in the classifiers and thus reduced the influence of the LIME visualizations.

Lack of Control Group in the Ribeiro Study No control conditions have been used in the Ribeiro study. This is a severe design flaw that allows many alternative explanations for their empirical results. The observed attitude changes in students could even have occurred without using LIME at all. Imagine that during class, an instructor asks a question, records the answers, and then asks the same question a second time. Most students will assume that the instructor was not satisfied with the initial responses and might contemplate changing their previous choice.

4.2. What Made the Distractor Visualizations Believable?

The bogus classifier was trained on random linear combinations of randomly selected pixels. One might expect that the resulting LIME visualizations would appear arbitrary or meaningless to the viewer. Looking closely at the column (b) of Fig. 1, one should be able to find examples of such arbitrary highlights, confirming this expectation. Still, the participants were not able to distinguish the classifiers based on the noisiness of the images.

One reason could be that the noise was spread evenly across the image space. As the digits were visually centered during MNIST preprocessing [12], the probabiliy of a random highlight being colocated with a sematically important feature might be high enough. The situation might change if the position of defining features change from image to image. Future studies should adress this question.

A second reason may be the aggregation of superpixels that is part of the LIME method. The choice of image segmentation method during this processing step can have strong impact on the resulting visualizations [17]. Superpixels following color boundaries result in visualizations that may look more convincing than they should. In the present study, this effect has been mitigated during stimuli generation by aggregating across six different segmentations and adding several regular grid segmentations that do not take the pixel color and brightness into account at all.

A third possible reason why the noisiness of the bogus classifier could not be spotted by the participants could be the general noise that is introduced by the LIME estimation process. This adds additional ambiguity to the visualizations and thereby increases the probability of misjudgements.

4.3. The Psychology of Post-diction

Humans are meaning makers [22] who strive for consistent representations of the world and tend to recognize familiar patterns in ambiguous environments, e.g., when looking at clouds in the sky. When interpreting LIME visualizations, similar processes might come into play. Especially when observers know whether an image belongs to the well-trained classifier, they may focus on parts of the LIME visualization that align with this knowledge and downplay contradictory parts.

The examples in Fig. 1 may serve as a demonstration of this phenomenon. The third row shows a MNIST digit with ground truth 5. While the target classifier wrongly predicts a 4, the second best prediction is correct and the corresponding LIME visualization looks like a 5. The distractor classifier is also "wrongly" predicting a 4, but the LIME visualization of it rather looks like a 5, which should point the viewer to the fact that the classifier is of low quality. Still, empirically, only 44% of the participants were able to use this information to correctly identify the bogus classifier for this image pair.

Post-diction, i.e., finding explanations in hindsight, is a natural psychological process that can be observed in many situations [23]. Importantly, once you know the correct answer, you cannot unknow it, and LIME visualizations may therefore appear more useful than they actually are.

4.4. Are Hand-Written Digits Too Abstract?

While the current results should be replicated with image data from different domains, some aspects of hand-written digits can serve as a good testbed for XAI visualizations.

One drawback of visualization methods like LIME is their sometimes reduced fidelity [24]. LIME estimates a linear surrogate model on superpixel-level [5]. This means that only *main effects* of superpixels can be represented well. For image classification problems that combine identification with localization of targets, this approach should be sufficient.

In the case of single digits, a linear approximation might not be sufficient. Two examples shall illustrate this. If the classification problem is zero vs. any other digit, then the most telling feature is that the center of the zero is empty, while all other digits cross the center (see Fig. 4.4). This can be well represented as a *main effect* of a central superpixel.

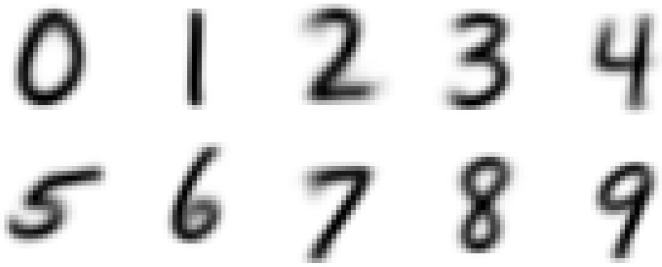


Figure 4: Centroids of the top 10% most similar MNIST images per digit. Each centroid represents the average pixel values of roughly 600 images (class sizes slightly vary within MNIST). Based on pixel-level clustering with complete-linkage.

The problem three vs. eight (vs. everything else) is different though. Three and eight share two loops on the right, which are open for a three and complete for an eight. Therefore from a classification point-of-view, a three must be represented as *interaction effect* between superpixels on the right (loops) and superpixels on the left (empty). Mathematically, this is a multiplicative relationship. But LIME can only approximate this as addition of positive weights on the right and negative weights on the left.

Visualizing negative weights is also not easy. While the heatmaps used in the given study show negative weights in red, they are completely omitted in masked images. This makes superpixels with negative weights indistinguishable from image areas with near-zero weights in masked images.

One tradeoff to consider here is that omitting negative weights can hide important information (lower fidelity), while having to reason about negative weights creates additional cognitive load for the viewer (lower comprehensibility). The advantages and disadvantages of heatmaps vs. masked images should be analyzed in future studies.

5. Conclusion and Outlook

The participants in this study were not better than chance at distinguishing a real from a completely bogus classifier based on LIME visualizations. Whether a classifier's predictions were consistent with the ground truth had a significant influence on the participants' choices, even if the classifier was trained on random functions on the pixels. These results demonstrate weaknesses in LIME as a XAI technique that should be addressed and maybe alternatives should be considered, e.g. example-based methods [25].

Future work will explore the generalization of the findings and applicability of the paradigm with other participant groups and other stimuli.

Acknowledgments

The author would like to thank three anonymous reviewers for their valuable feedback.

Declaration on Generative AI

AI tools (Microsoft Copilot) were used to implement parts of the PHP code of the online experiment, suggest distractors for the machine learning knowledge questionnaire, and improve writing style and spelling/grammar. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] N. Kosmyna, E. Hauptmann, Y. T. Yuan, J. Situ, X.-H. Liao, A. V. Beresnitzky, I. Braunstein, P. Maes, Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task, 2025. doi:10.48550/arXiv.2506.08872.

[2] Z. K. Cui, M. Demirer, S. Jaffe, L. Musolff, S. Peng, T. Salz, The effects of generative AI on high skilled work: Evidence from three field experiments with software developers, 2025. doi:10.2139/ssrn.4945566.

[3] V. Schmitt, B. P. Csomor, J. Meyer, L.-F. Villa-Areas, C. Jakob, T. Polzehl, S. Möller, Evaluating human-centered AI explanations: Introduction of an XAI evaluation framework for fact-checking, in: Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation, MAD '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 91–100. doi:10.1145/3643491.3660283.

[4] D. Minh, H. X. Wang, Y. F. Li, T. N. Nguyen, Explainable artificial intelligence: A comprehensive review, Artificial Intelligence Review 55 (2022) 3503–3568. doi:10.1007/s10462-021-10088-y.

[5] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts, Data Mining and Knowledge Discovery 38 (2024) 3043–3101. doi:10.1007/s10618-022-00867-8.

[6] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, et al., Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301. doi:10.1016/j.inffus.2024.102301.

[7] M. T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you? Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[8] R. D. Cook, Detection of influential observation in linear regression, Technometrics 19 (1977) 15–18. doi:10.1080/00401706.1977.10489493.

[9] J. Dieber, S. Kirrane, Why model why? assessing the strengths and limitations of LIME, 2020. doi:10.48550/arXiv.2012.00093.

[10] P. Knab, S. Marton, U. Schlegel, C. Bartelt, Which LIME should I trust? concepts, challenges, and solutions, in: R. Guidotti, U. Schmid, L. Longo (Eds.), Explainable Artificial Intelligence. xAI 2025, Springer, 2025, pp. 28–52. doi:10.1007/978-3-032-08324-1_2.

[11] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Computing Surveys 55 (2023) 1–42. doi:10.1145/3583558.

[12] Y. LeCun, C. Cortes, C. Burges, MNIST handwritten digit database, 2010. URL: https://huggingface.co/datasets/ylecun/mnist.

[13] M. Halbrügge, KI-Entscheidungen visualisieren und erklären mit LIME – Source Code, 2025. doi:10.5281/zenodo.15608360.

[14] M. Halbrügge, Web experiment for the evaluation of LIME understandability, 2026. doi:10.5281/zenodo.21677145, source code release.
[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[16] L. Schallner, J. Rabold, O. Scholz, U. Schmid, Effect of superpixel aggregation on explanations in LIME: A case study with biological data, in: P. Cellier, K. Driessens (Eds.), Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019, Springer, 2020, pp. 147–158. doi:10.1007/978-3-030-43823-4_13.
[17] M. Halbrügge, T. Jungeblut, Beyond heatmaps: Evaluating LIME visualizations for human understandability, in: Proceedings of the biannual conference of the German cognitive science society (KogWis 2025), 2025.
[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE transactions on pattern analysis and machine intelligence 34 (2012) 2274–2282. doi:10.1109/TPAMI.2012.120.
[19] D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, Journal of Statistical Software 67 (2015) 1–48. doi:10.18637/jss.v067.i01.
[20] Angelo Canty, B. D. Ripley, boot: Bootstrap R (S-Plus) Functions, 2024. R package version 1.3-30.
[21] E. Slany, S. Scheele, U. Schmid, Bayesian CAIPI: A probabilistic approach to explanatory and interactive machine learning, in: S. e. a. Nowaczyk (Ed.), Artificial Intelligence. ECAI 2023 International Workshops, Springer Nature Switzerland, Cham, 2024, pp. 285–301. doi:10.1007/978-3-031-50396-2_16.
[22] J. D. Novak, Human constructivism: A unification of psychological and epistemological phenomena in meaning making, International Journal of Personal Construct Psychology 6 (1993) 167–193. doi:10.1080/08936039308404338.
[23] N. J. Roese, K. D. Vohs, Hindsight bias, Perspectives on psychological science 7 (2012) 411–426. doi:10.1177/1745691612454303.
[24] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, Advances in neural information processing systems 31 (2018).
[25] T. Speith, A review of taxonomies of explainable artificial intelligence (XAI) methods, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2239–2250. doi:10.1145/3531146.3534639.