A User Interface for Explaining Machine Learning Model Explanations

Md Abdul Kadir
Abdulrahman Mohamed Selim
abdul.kadir@dfki.de
abdulrahman.mohamed@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany

Michael Barz
Daniel Sonntag
michael.barz@dfki.de
daniel.sonntag@dfki.de
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
University of Oldenburg
Oldenburg, Germany

ABSTRACT

Explainable Artificial Intelligence (XAI) is an emerging subdiscipline of Machine Learning (ML) and human-computer interaction. Discriminative models need to be understood. An explanation of such ML models is vital when an AI system makes decisions that have significant consequences, such as in healthcare or finance. By providing an input-specific explanation, users can gain confidence in an AI system's decisions and be more willing to trust and rely on it. One problem is that interpreting example-based explanations for discriminative models, such as saliency maps, can be difficult because it is not always clear how the highlighted features contribute to the model's overall prediction or decisions. Moreover, saliency maps, which are state-of-the-art visual explanation methods, do not provide concrete information on the influence of particular features. We propose an interactive visualisation tool called EMILE-UI that allows users to evaluate the provided explanations of an image-based classification task, specifically those provided by saliency maps. This tool allows users to evaluate the accuracy of a saliency map by reflecting the true attention or focus of the corresponding model. It visualises the relationship between the ML model and its explanation of input images, making it easier to interpret saliency maps and understand how the ML model actually predicts. Our tool supports a wide range of deep learning image classification models and image data as inputs.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

KEYWORDS

Explainability, Interpretability, Transparency, Trustworthiness, AI, ML

ACM Reference Format:
Md Abdul Kadir, Abdulrahman Mohamed Selim, Michael Barz, and Daniel Sonntag. 2023. A User Interface for Explaining Machine Learning Model Explanations. In *28th International Conference on Intelligent User Interfaces (IUI '23 Companion), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3581754.3584131

IUI '23 Companion, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0107-8/23/03.
https://doi.org/10.1145/3581754.3584131

1 INTRODUCTION

Deep Neural Networks (DNNs) have proven to be useful in multiple domains; however, because of their complexity, they are often regarded as black-box solutions because it can be difficult to understand how they arrive at their predictions [4]. Providing explanations for such models' predictions is a solution for understanding their working mechanisms. Several saliency map-based explanation methods are popular in Machine Learning (ML) research and are used to provide information on why an ML model makes a certain prediction [5]. Examples include layer-wise relevance propagation [17], Grad-CAM [25], integrated gradient [27], guided back-propagation [26], pixel-wise decomposition [3], and contrastive explanations [16].

However, a question remains on how to assess the validity of these explanation methods. Evaluation methods have emerged to measure the quality of these explanations. Some researchers used ground truth data, such as object-localisation data or ground truth masks [8, 18]. There are also methods for evaluating how an explanation reflects the true attention or focus of the corresponding model, also known as the faithfulness of an explanation [22]. Faithfulness is a measure of the accuracy of the explanation and the reasoning behind the class prediction [15]. For example, Yeh et al. [31] used a square-kernel approach in their concept called infidelity, whose premise is that appropriately reducing the sensitivity can lower the infidelity by using a simple kernel smoothing-based algorithm, and the models that optimise infidelity offer better explanations. Rieger and Hansen [20] presented their solution called IROF (Iterative Removal Of Features) which used image segmentation to divide the image into coherent segments, and a good explanation would attribute high relevance to segments important for classification. Bach et al. [3] proposed a simple approach of flipping pixels to their opposites and seeing how the evaluation metric was affected. Samek et al. [22] took another pixel-related evaluation approach to evaluate the quality of a heatmap produced by the explanation algorithm. A step further would be to remove groups of relevant pixels. Hooker et al. [14] presented ROAR (Remove And Retrain) where they replace a fraction of the pixels, which were estimated to be most relevant, with a fixed uninformative value, but their approach is computationally expensive because it requires retraining

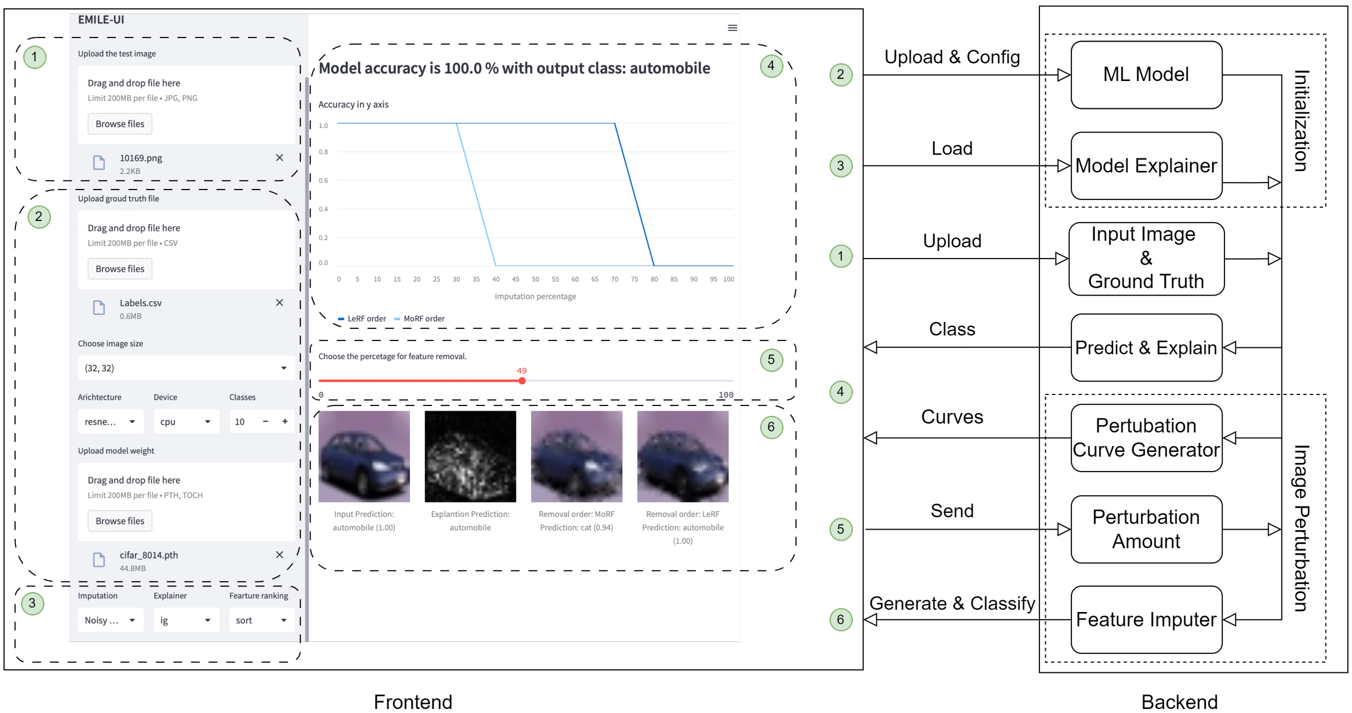

Figure 1: System architecture of our tool with an accompanying UI screenshot. In the frontend, the user uploads the model and its configuration (2). The user then select the explanation method that he or she wants to evaluate (3) and uploads the test image and ground truth (1). Afterwards, the model generates the saliency map (6) and perturbation curves for the saliency map (4). Patterns (Indicated in Fig. 2) of the perturbation curves can indicate the explanation map's faithfulness. Using the percentage slider (5), the user selects the amount of feature to remove from the input image. The two bottom-right pictures (6) are the resulting images after the feature removal. Change in the model's classification from 'automobile' to 'cat' due to the evident feature removal in the number plate region suggests a potential false correlation between the model and the explanation map.

the evaluated models. As an improvement on ROAR, Rong et al. [21] presented ROAD (Remove And Debias) which measures global fidelity among attribution explanations. Moreover, there has not yet been a single evaluation technique that considers the wrong model creation (Right for wrong reason). Interactively showing users the decision-making pattern of the model can help them understand the false correlations [12]. Most explanation evaluation techniques, such as those of Ghorbani et al. [13] and Dabkowski and Gal [9], do not consider user interactions, which are essential because they allow users to explore and interact intuitively and flexibly with the results of an ML model [19, 29].

Due to the absence of such user interfaces and interactions, there could be a lack of user trust in the understanding of ML models and their explanations [6, 10, 23, 32]. However, presenting explanations in the form of interactive systems can help build trust and confidence in ML models [30, 33, 34]. In addition, using such interfaces can help users better understand an ML model's behaviour [7] and identify biases or errors in the decision-making process [2].

This paper demonstrates how saliency-based explanations can be used as building blocks of interactive systems to interpret ML image classifications and evaluate their faithfulness. Our tool, EMILE-UI, shown in Figure 1, is a user interface for the recent state-of-the-art approach of Rong et al. [21] and extends it. It allows users to interactively and visually test the faithfulness of ML explanations. EMILE allows ML practitioners, even those without any experience in explainable ML, to actively evaluate a variety of explanation techniques, and helps ML experts debug ML models similar to the works of Suresh et al. [28]. We argue that integrating interaction with saliency map-based explanations offers good interpretability, because users can interact with explanations visually and observe the behaviour of the models. We hope that our work will encourage further approaches to attempt to build interactive user interfaces for explainable ML evaluation systems. Our demo is publicly available at https://iml.dfki.de/demos/emile.

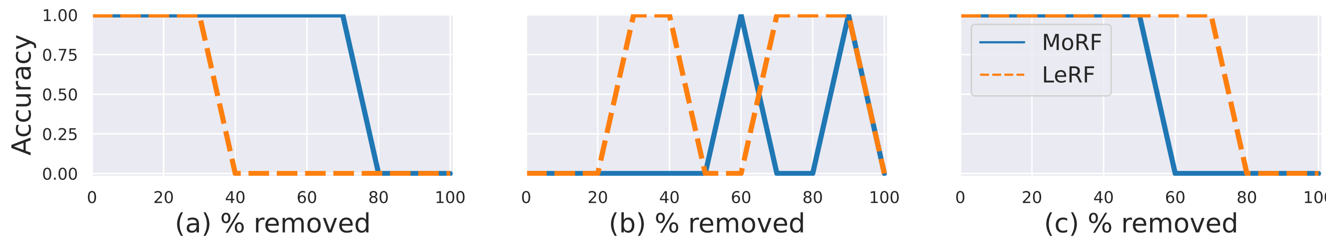

Figure 2: We present a set of perturbation curves for an ML model on a single input image. Plot (c) represents the consistency of the saliency map, and Plot (a) represents the inconsistency of the saliency map. Plot (b) also represents the inconsistency and random feature importance in the saliency map.

2 TECHNICAL BACKGROUND

Attribution map, also known as a saliency map or heatmap, is a visual representation of which parts of the input have the most influence on the output of an ML model [25, 27]. **Faithfulness** testing involves applying perturbations to the input, based on the attribute map, and checking whether the resulting changes in the output match the changes indicated by the saliency map. Faithfulness testing requires consistency between the changes in the model output when the features are removed using the Most Relevant Feature (MoRF) and the Least Relevant Feature (LeRF) orders [21]. **Linear Imputation** is one of the methods for removing a feature from an image [14, 20–22]. According to the rigorous evaluation of Rong et al. [21] and Hooker et al. [14], noisy linear imputation reduces information leakage through the shape of the imputation mask and provides consistent results for different evaluation orders. **Region perturbation** is a technique that can be used to evaluate decisions made by DNNs by perturbing specific regions of the input data and observing how the prediction of the model changes [21, 22]. Rong et al. [21] and Samek et al. [22] showed how input perturbation based on the attribution map influences the accuracy of an ML model. Figure 2 shows how the accuracy of the model changed during perturbation for a single test image.

3 EMILE-UI

Explaining MachIne Learning Explanations User Interface or EMILE-UI is a simple yet powerful tool for understanding the behaviour of an ML image classification model with respect to its explanation. The main goal of EMILE-UI is to allow an ML user to evaluate the faithfulness of a generated saliency map. EMILE-UI operates in three steps. In the first step, the user selects the deep learning architecture, uploads the weights, selects the attribution method (Explainer), and sets the hyperparameters. In the second step, the user uploads the test image and ground truth. In the third step, the model predicts the class of the image and generates a saliency map. EMILE-UI first draws the perturbation curves for the MoRF and LeRF orders and shows the user whether the attribution map is consistent for both evaluation techniques. Afterwards, the user can slide the percentage bar, which is the orange bar in Step (5) of Figure 1, to select the amount of features to be removed from the input image. The tool then shows the images after the removal of the most and least important features in an orderly manner.

Explanation evaluation is a valuable tool for ML users to evaluate the faithfulness of saliency maps produced by an explanation method and to gain insights into the factors that influence the model's predictions. During the evaluation, the user can look at the saliency map and perturbation curves. If the perturbation curves are inconsistent, the user can be sure that the saliency map is not faithful [21]. The user can move the percentage slider to a value where the perturbation curves make transitions. The tool then removes the corresponding amount of relevant or irrelevant features from the input image, which are shown in the two bottom-right images of Figure 1. Consequently, the user can observe the changes that occur in the input image when the model changes its output. The graph in the middle of Figure 1 shows the changes in the model's prediction with probability values. This helps the user understand whether a model explanation is faithful and provides deeper insights into the models' prediction behaviour and intuition on the information that the saliency map carries. It can also be used to debug image classification models [24]. For example, it can detect whether a model is biased towards certain features of an image. In addition, we want to leverage EMILE-UI to conduct a user study to evaluate users' trust in ML models and their explanations by allowing them to interact with our tool and obtain feedback on how their trust in the ML model changes. We believe that this can be very helpful rather than simply showing users saliency maps, which have been found not to provide users with sufficient information [1].

To ensure hardware independence for end users, we built EMILE-UI with a browser-based framework, Streamlit[1]. EMILE-UI can be deployed on any Linux server with or without a GPU. We used the PyTorch framework[2] for deep learning tasks, and our implementation can also be extended with other deep learning frameworks.

4 CONCLUSION

We presented EMILE-UI, a tool for evaluating and understanding image-based deep learning classifiers' explanations. EMILE-UI is a user interface for the method proposed by Rong et al. [21], which may facilitate a better understanding of explanations through interaction and visual feedback. We wanted to make the system more appealing to a broader range of potential users by not requiring explicit programming, as suggested by Dudley and Kristensson [11]. Our future goal is to conduct a user study to evaluate how our tool affects user trust in ML classifications and explanations.

[1] https://docs.streamlit.io
[2] https://pytorch.org

ACKNOWLEDGMENTS

This work was funded by the German Federal Ministry of Health (BMG) under grant number 2520DAT0P2 (pAItient).

REFERENCES

[1] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3377325.3377519

[2] Alessa Angerschmid, Kevin Theuermann, Andreas Holzinger, Fang Chen, and Jianlong Zhou. 2022. Effects of Fairness and Explanation on Trust in Ethical AI. In *Machine Learning and Knowledge Extraction: 6th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2022, Vienna, Austria, August 23–26, 2022, Proceedings* (Vienna, Austria). Springer-Verlag, Berlin, Heidelberg, 51–67. https://doi.org/10.1007/978-3-031-14463-9_4

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* 10, 7 (July 2015), e0130140. https://doi.org/10.1371/journal.pone.0130140 Publisher: Public Library of Science.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[5] Nadia Burkart and Marco F. Huber. 2021. A Survey on the Explainability of Supervised Machine Learning. *J. Artif. Int. Res.* 70 (may 2021), 245–317. https://doi.org/10.1613/jair.1.12228

[6] Margaret Burnett. 2020. Explaining AI: Fairly? Well?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/3377325.3380623

[7] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The Effects of Example-Based Explanations in a Machine Learning Interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 258–262. https://doi.org/10.1145/3301275.3302289

[8] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Lake Tahoe, Nevada, 839–847. https://doi.org/10.1109/WACV.2018.00097

[9] Piotr Dabkowski and Yarin Gal. 2017. Real Time Image Saliency for Black Box Classifiers. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., Long Beach. https://proceedings.neurips.cc/paper/2017/file/0060ef47b12160b9198302ebdb144dcf-Paper.pdf

[10] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. 2021. Who Needs to Know What, When?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1591–1602. https://doi.org/10.1145/3461778.3462131

[11] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (June 2018), 8:1–8:37. https://doi.org/10.1145/3185517

[12] Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. 2022. Aligning Eyes between Humans and Deep Neural Network through Interactive Attention Alignment. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 489 (nov 2022), 28 pages. https://doi.org/10.1145/3555590

[13] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., Vancouver. https://proceedings.neurips.cc/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf

[14] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., Vancouver, Canada, 9. https://proceedings.neurips.cc/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf

[15] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386

[16] Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Yunfeng Zhang, Karthikeyan Shanmugam, and Chun-Chen Tu. 2021. Leveraging Latent Features for Local Explanations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*. Association for Computing Machinery, New York, NY, USA, 1139–1149. https://doi.org/10.1145/3447548.3467265

[17] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73 (Feb. 2018), 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

[18] Fabrizio Nunnari, Md Abdul Kadir, and Daniel Sonntag. 2021. On the Overlap Between Grad-CAM Saliency Maps and Explainable Visual Features in Skin Cancer Images. In *Machine Learning and Knowledge Extraction*, Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar Weippl (Eds.). Springer International Publishing, Cham, 241–253.

[19] Fabrizio Nunnari and Daniel Sonntag. 2021. A Software Toolbox for Deploying Deep Learning Decision Support Systems with XAI Capabilities. In *Companion of the 2021 ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (Virtual Event, Netherlands) *(EICS '21)*. Association for Computing Machinery, New York, NY, USA, 44–49. https://doi.org/10.1145/3459926.3464753

[20] Laura Rieger and Lars Kai Hansen. 2020. IROF: a low resource evaluation metric for explanation methods. https://doi.org/10.48550/arXiv.2003.08747 arXiv:2003.08747 [cs].

[21] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. 2022. A Consistent and Efficient Evaluation Strategy for Attribution Methods. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, Baltimore, Maryland, 18770–18795. https://proceedings.mlr.press/v162/rong22a.html

[22] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11 (Nov. 2017), 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820 Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[23] Téo Sanchez, Baptiste Caramiaux, Pierre Thiel, and Wendy E. Mackay. 2022. Deep Learning Uncertainty in Machine Teaching. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 173–190. https://doi.org/10.1145/3490099.3511117

[24] Udo Schlegel, Eren Cakmak, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke, and Daniel A. Keim. 2020. Towards Visual Debugging for Multi-Target Time Series Classification. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 202–206. https://doi.org/10.1145/3377325.3377528

[25] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 618–626. https://doi.org/10.1109/ICCV.2017.74

[26] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. 2015. Striving for Simplicity: The All Convolutional Net. In *ICLR (workshop track)*. ICLR, San Diego, California, 10. http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, Sydney, Australia, 3319–3328.

[28] Harini Suresh, Kathleen M Lewis, John Guttag, and Arvind Satyanarayan. 2022. Intuitively Assessing ML Model Reliability through Example-Based Explanations and Editing Model Inputs. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 767–781. https://doi.org/10.1145/3490099.3511160

[29] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (2022), 364–381. https://doi.org/10.1016/j.future.2022.05.014

[30] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 189–201. https://doi.org/10.1145/3377325.3377480

[31] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. 2019. On the (In)fidelity and Sensitivity of Explanations. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., Vancouver, Canada, 9 pages. https://proceedings.neurips.cc/paper/2019/hash/

a7471fdc77b3435276507cc8f2dc2569-Abstract.html

[32] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate? An Investigation from Perception to Decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277

[33] Jan Zacharias, Michael Barz, and Daniel Sonntag. 2018. A Survey on Deep Learning Toolkits and Libraries for Intelligent User Interfaces. *CoRR* abs/1803.04818 (2018), 6. arXiv:1803.04818 http://arxiv.org/abs/1803.04818

[34] Jianlong Zhou, Zhidong Li, Huaiwen Hu, Kun Yu, Fang Chen, Zelin Li, and Yang Wang. 2019. Effects of Influence on User Trust in Predictive Decision Making. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312962