Year	Author	Title	Publication Title	Url	Date
2018	Alvarez Melis, David; Jaakkola, Tommi	Towards Robust Interpretability with Self-Explaining Neural Networks	Advances in Neural Information Processing Systems 31	http://papers.nips.cc/paper/8003-towards-robust-interpretability-with-self-explaining-neural-networks.pdf	2018
2017	Samek, Wojciech; Binder, Alexander; Montavon, Grégoire; Lapuschkin, Sebastian; Müller, Klaus-Robert	Evaluating the Visualization of What a Deep Neural Network Has Learned	IEEE Transactions on Neural Networks and Learning Systems		2017-11
2018	Gilpin, Leilani H.; Bau, David; Yuan, Ben Z.; Bajwa, Ayesha; Specter, Michael; Kagal, Lalana	Explaining Explanations: An Overview of Interpretability of Machine Learning	2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)		2018-10
2018	Robnik-Šikonja, Marko; Bohanec, Marko	Perturbation-Based Explanations of Prediction Models	Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent	https://doi.org/10.1007/978-3-319-90403-0_9	2018
2007	Tintarev, Nava; Masthoff, Judith	A Survey of Explanations in Recommender Systems	2007 IEEE 23rd International Conference on Data Engineering Workshop		2007-04
2019	Miller, Tim	Explanation in artificial intelligence: Insights from the social sciences	Artificial Intelligence	http://www.sciencedirect.com/science/article/pii/S0004370218305988	2/1/19
2009	Chang, Jonathan; Gerrish, Sean; Wang, Chong; Boyd-graber, Jordan L.; Blei, David M.	Reading Tea Leaves: How Humans Interpret Topic Models	Advances in Neural Information Processing Systems 22	http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf	2009
2019	Lage, Isaac; Chen, Emily; He, Jeffrey; Narayanan, Menaka; Kim, Been; Gershman, Sam; Doshi-Velez, Finale	An Evaluation of the Human-Interpretability of Explanation	Workshop on Correcting and Critiquing Trends in Machine Learning,		2019
2018	Narayanan, Menaka; Chen, Emily; He, Jeffrey; Kim, Been; Gershman, Sam; Doshi-Velez, Finale	How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation	arXiv:1802.00682 [cs]	http://arxiv.org/abs/1802.00682	2/2/18
2019	Schmidt, Philipp; Biessmann, Felix	Quantifying Interpretability and Trust in Machine Learning Systems	Workshop on Network Interpretability for Deep Learning,		2019
2020	Bhatt, Umang; Moura, José M. F.; Weller, Adrian	Evaluating and Aggregating Feature-based Model Explanations		https://www.ijcai.org/proceedings/2020/417	7/9/20
2019	Hooker, Sara; Erhan, Dumitru; Kindermans, Pieter-Jan; Kim, Been	A Benchmark for Interpretability Methods in Deep Neural Networks	Advances in Neural Information Processing Systems 32	http://papers.nips.cc/paper/9167-a-benchmark-for-interpretability-methods-in-deep-neural-networks.pdf	2019
2020	Rieger, Laura; Hansen, Lars Kai	IROF: a low resource evaluation metric for explanation methods	Workshop,		2020
2018	Ancona, Marco; Ceolini, Enea; Öztireli, Cengiz; Gross, Markus	Towards better understanding of gradient-based attribution methods for Deep Neural Networks		https://openreview.net/forum?id=Sy21R9JAW	2/15/18
2019	Yeh, Chih-Kuan; Hsieh, Cheng-Yu; Suggala, Arun; Inouye, David I; Ravikumar, Pradeep K	On the (In)fidelity and Sensitivity of Explanations	Advances in Neural Information Processing Systems 32	http://papers.nips.cc/paper/9278-on-the-infidelity-and-sensitivity-of-explanations.pdf	2019