Full Set of References

[1] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence,* vol. 267, pp. 1-38, 2019.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)," *IEEE Access,* vol. 6, pp. 52138-52160, 2018.

[3] L. Antwarg, B. Shapira and L. Rokach, "Explaining anomalies detected by autoencoders using SHAP," *arXiv preprint arXiv:1903.02407,* 2019.

[4] D. Gunning and D. W. Aha, "DARPA's explainable artificial intelligence program," *AI Magazine,* vol. 40, pp. 44-58, 2019.

[5] A. B. Arrieta, N. Díaz- Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil- López, D. Molina, R. Benjamins and others, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion,* vol. 58, pp. 82-115, 2020.

[6] D. Linsley, D. Scheibler, S. Eberhardt and T. Serre, "Global-and-local attention networks for visual recognition," *arXiv preprint arXiv:1805.08819,* 2018.

[7] S. Seo, J. Huang, H. Yang and Y. Liu, "Interpretable convolutional neural networks with dual local and global attention for review rating prediction," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017.

[8] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608,* 2017.

[9] F. K. Došilović, M. Brčić and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, 2018.

[10] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018.

[11] D. Wang, Q. Yang, A. Abdul and B. Y. Lim, "Designing theory-driven user-centric explainable AI," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019.

[12] S. Chari, D. M. Gruen, O. Seneviratne and D. L. McGuinness, "Directions for Explainable Knowledge-Enabled Systems," *arXiv preprint arXiv:2003.07523,* 2020.

[13] C. Rudin and B. Ustun, "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice," *Interfaces,* vol. 48, pp. 449-466, 2018.

[14] P. Madumal, T. Miller, L. Sonenberg and F. Vetere, "A grounded interaction protocol for explainable artificial intelligence," *arXiv preprint arXiv:1903.02409,* 2019.

[15] S. Gregor, "The nature of theory in information systems," *MIS quarterly,* pp. 611-642, 2006.

[16] H. J. P. Weerts, W. Ipenburg and M. Pechenizkiy, "A Human-Grounded Evaluation of SHAP for Alert Processing," *arXiv preprint arXiv:1907.03324,* 2019.

[17] H. J. P. Weerts, W. Ipenburg and M. Pechenizkiy, "Case-Based Reasoning for Assisting Domain Experts in Processing Fraud Alerts of Black-Box Machine Learning Models," *arXiv preprint arXiv:1907.03334,* 2019.

[18] B. Laughlin, K. Sankaranarayanan and K. El-Khatib, "A Service Architecture Using Machine Learning to Contextualize Anomaly Detection," *Journal of Database Management (JDM),* vol. 31, pp. 64-84, 2020.

[19] A. Shrikumar, P. Greenside and A. Kundaje, "Learning important features through propagating activation differences," *arXiv preprint arXiv:1704.02685,* 2017.

[20] D. Collaris, L. M. Vink and J. J. Wijk, "Instance-level explanations for fraud detection: A case study," *arXiv preprint arXiv:1806.07129,* 2018.

[21] M. Moalosi, H. Hlomani and O. S. D. Phefo, "Combating credit card fraud with online behavioural targeting and device fingerprinting," *International Journal of Electronic Security and Digital Forensics,* vol. 11, pp. 46-69, 2019.

[22] A. J. Barda, "Design and Evaluation of User-Centered Explanations for Machine Learning Model Predictions in Healthcare," 2020.

[23] S. Mohseni, N. Zarei and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," *arXiv,* pp. arXiv--1811, 2018.

[24] A. R. Akula, C. Liu, S. Saba-Sadiya, H. Lu, S. Todorovic, J. Y. Chai and S.-C. Zhu, "X-tom: Explaining with theory-of-mind for gaining justified human trust," *arXiv preprint arXiv:1909.06907,* 2019.

[25] K. Sokol and P. Flach, "Explainability fact sheets: a framework for systematic assessment of explainable approaches," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[26] M. T. Ribeiro, S. Singh and 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.

[27] J. M. Logg, "When do people rely on algorithms?," 2016.

[28] R. A. Leite, T. Gschwandtner, S. Miksch, S. Kriglstein, M. Pohl, E. Gstrein and J. Kuntner, "Eva: Visual analytics to identify fraudulent events," *IEEE transactions on visualization and computer graphics,* vol. 24, pp. 330-339, 2017.

[29] T. Spinner, U. Schlegel, H. Schäfer and M. El-Assady, "explAIner: A visual analytics framework for interactive and explainable machine learning," *IEEE transactions on visualization and computer graphics,* vol. 26, pp. 1064-1074, 2019.

[30] A. Chatzimparmpas, R. M. Martins, I. Jusufi and A. Kerren, "A survey of surveys on the use of visualization for interpreting machine learning models," *Information Visualization,* p. 1473871620904671, 2020.

[31] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi and A. Kerren, "The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations," in *Computer graphics forum (Print)*, 2020.

[32] G. J. Browne and M. B. Rogich, "An empirical investigation of user requirements elicitation: Comparing the effectiveness of prompting techniques," *Journal of Management Information Systems,* vol. 17, pp. 223-249, 2001.

[33] R. Galliers, Information analysis: selected readings, Addison-Wesley Longman Publishing Co., Inc., 1987.

[34] F. Creedon, "The framework for REVIEWS: an exploration into design principles for an electronic medical early warning system observation chart," 2016.

[35] N. A. Stanton, A. Hedge, K. Brookhuis, E. Salas and H. W. Hendrick, Handbook of human factors and ergonomics methods, CRC press, 2004.

[36] M. G. Helander, Handbook of human-computer interaction, Elsevier, 2014.

[37] A. Preece, D. Harborne, D. Braines, R. Tomsett and S. Chakraborty, "Stakeholders in explainable AI," *arXiv preprint arXiv:1810.00184,* 2018.

[38] M. Ribera and A. Lapedriza, "Can we do better explanations? A proposal of user-centered explainable AI.," in *IUI Workshops*, 2019.

[39] M. A. Ahmad, C. Eckert and A. Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 2018.

[40] R. Srinivasan, A. Chander and P. Pezeshkpour, "Generating user-friendly explanations for loan denials using GANs," *arXiv preprint arXiv:1906.10244,* 2019.

[41] L. Zheng, G. Liu, C. Yan and C. Jiang, "Transaction fraud detection based on total order relation and behavior diversity," *IEEE Transactions on Computational Social Systems,* vol. 5, pp. 796-806, 2018.

[42] S. B. E. Raj and A. A. Portia, "Analysis on credit card fraud detection methods," in *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)*, 2011.

[43] C. M. Bishop, Pattern recognition and machine learning, New, York: Springer, 2006.

[44] W. N. Dilla and R. L. Raschke, "Data visualization for fraud detection: Practice implications and a call for future research," *International Journal of Accounting Information Systems,* vol. 16, pp. 1-22, 2015.

[45] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein and J. Kuntner, "Visual analytics for event detection: Focusing on fraud," *Visual Informatics,* vol. 2, pp. 198-212, 2018.

[46] T. Munzner, "A nested model for visualization design and validation," *IEEE transactions on visualization and computer graphics,* vol. 15, pp. 921-928, 2009.

[47] L. Franklin, M. Pirrung, L. Blaha, M. Dowling and M. Feng, "Toward a visualization-supported workflow for cyber alert management using threat models and human-centered design," in *2017 IEEE Symposium on Visualization for Cyber Security (VizSec)*, 2017.

[48] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017.

[49] R. Klopper, S. Lubbe and H. Rugbeer, "The matrix method of literature review," *Alternation,* vol. 14, pp. 262-276, 2007.

[50] E. Novikova, I. Kotenko and E. Fedotov, "Interactive Multi-View Visualization for Fraud Detection in Mobile Money Transfer Services," *International Journal of Mobile Computing and Multimedia Communications (IJMCMC),* vol. 6, pp. 73-97, 2014.

[51] W. Didimo, G. Liotta and F. Montecchiani, "Network visualization for financial crime detection," *Journal of Visual Languages \& Computing,* vol. 25, pp. 433-451, 2014.

[52] C. R. Kothari, Research methodology: Methods and techniques, New Age International, 2004.

[53] A. R. Hevner, S. T. March, J. Park and S. Ram, "Design science in information systems research," *MIS quarterly,* pp. 75-105, 2004.

[54] A. Hevner and S. Chatterjee, "Design science research in information systems," in *Design research in information systems*, Springer, 2010, pp. 9-22.

[55] H. Takeda, P. Veerkamp and H. Yoshikawa, "Modeling design process," *AI magazine,* vol. 11, pp. 37-37, 1990.

[56] P. Offermann, O. Levina, M. Schönherr and U. Bub, "Outline of a design science research process," in *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, 2009.

[57] A. R. Hevner, "A three cycle view of design science research," *Scandinavian journal of information systems,* vol. 19, p. 4, 2007.

[58] K. Peffers, T. Tuunanen, M. A. Rothenberger and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems,* vol. 24, pp. 45-77, 2007.

[59] Ł. Ostrowski, M. Helfert and F. Hossain, "A conceptual framework for design science research," in *International Conference on Business Informatics Research*, 2011.

[60] K. Pfeffers, T. Tuunanen, C. E. Gengler, M. Rossi, W. Hui, V. Virtanen and J. Bragge, "The design science research process: A model for producing and presenting information systems research," in *Proceedings of the First International Conference on Design Science Research in Information Systems and Technology (DESRIST 2006), Claremont, CA, USA*, 2006.

[61] M. Helfert, B. Donnellan and L. Ostrowski, "The case for design science utility and quality-Evaluation of design science artifact within the sustainable ICT capability maturity framework," *Systems, Signs and Actions: An International Journal on Information Technology, Action, Communication and Workpractices,* vol. 6, pp. 46-66, 2012.

[62] M. Hall, D. Harborne, R. Tomsett, V. Galetic, S. Quintana-Amate, A. Nottle and A. Preece, "A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems," in *Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019), Macau, China*, 2019.

[63] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith and D. Bohlender, "Explainability as a non-functional requirement," in *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 2019.

[64] C. T. Wolf, "Explainability scenarios: towards scenario-based XAI design," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019.

[65] Q. V. Liao, D. Gruen and S. Miller, "Questioning the AI: Informing Design Practices for Explainable AI User Experiences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[66] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug and H. Hussmann, "Bringing transparency design into practice," in *23rd international conference on intelligent user interfaces*, 2018.

[67] M. Hertzum, "Making use of scenarios: a field study of conceptual design," *International Journal of Human-Computer Studies,* vol. 58, pp. 215-239, 2003.

[68] J. M. Carroll, "Becoming social: expanding scenario-based approaches in HCI," *Behaviour \& Information Technology,* vol. 15, pp. 266-275, 1996.

[69] M. B. Rosson and J. M. Carroll, "Scenario based design," *Human-computer interaction. boca raton, FL,* pp. 145-162, 2009.

[70] A. Witzel and H. Reiter, The problem-centred interview, Sage, 2012.

[71] J. Dick, E. Hull and K. Jackson, Requirements engineering, Springer, 2017.

[72] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS quarterly,* pp. xiii--xxiii, 2002.

[73] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey and G. Klein, "Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI," *arXiv preprint arXiv:1902.01876,* 2019.

[74] R. Chang, A. Lee, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern and A. Sudjianto, "Scalable and interactive visual analysis of financial wire transactions for fraud detection," *Information visualization,* vol. 7, pp. 63-76, 2008.

[75] Y. Shi, Y. Liu, H. Tong, J. He, G. Yan and N. Cao, "Visual Analytics of Anomalous User Behaviors: A Survey," *arXiv preprint arXiv:1905.06720,* 2019.

[76] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein and J. Kuntner, "Visual analytics for fraud detection and monitoring," in *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2015.

[77] J. Sun, Q. Zhu, Z. Liu, X. Liu, J. Lee, Z. Su, L. Shi, L. Huang and W. Xu, "FraudVis: understanding unsupervised fraud detection algorithms," in *2018 IEEE Pacific Visualization Symposium (PacificVis)*, 2018.

[78] E. N. Argyriou, A. Symvonis and V. Vassiliou, "A fraud detection visualization system utilizing radial drawings and heat-maps," in *2014 International Conference on Information Visualization Theory and Applications (IVAPP)*, 2014.

[79] M. Ahmed, A. N. Mahmood and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Generation Computer Systems,* vol. 55, pp. 278-288, 2016.

[80] C. Phua, V. Lee, K. Smith and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *arXiv preprint arXiv:1009.6119,* 2010.

[81] R. J. Bolton, "Statistical fraud detection: A review," in *JSTOR*, 2002.

[82] J. Zerilli, A. Knott, J. Maclaurin and C. Gavaghan, "Transparency in algorithmic and human decision-making: is there a double standard?," *Philosophy \& Technology,* vol. 32, pp. 661-683, 2019.

[83] J. Schneider and J. Handali, "Personalized explanation in machine learning: A conceptualization," *arXiv preprint arXiv:1901.00770,* 2019.

[84] J. Krause, A. Perer and E. Bertini, "Using visual analytics to interpret predictive machine learning models," *arXiv preprint arXiv:1606.05685,* 2016.

[85] J. Krause, A. Perer and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[86] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan and W.-K. Wong, "Too much, too little, or just right? Ways explanations impact end users' mental models," in *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 2013.

[87] R. C. Nickerson, U. Varshney and J. Muntermann, "A method for taxonomy development and its application in information systems," *European Journal of Information Systems,* vol. 22, pp. 336-359, 2013.

[88] C. Molnar, Interpretable Machine Learning, Lulu. com, 2020.

[89] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovi{\'c} and others, "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," *arXiv preprint arXiv:1909.03012,* 2019.

[90] M. Du, N. Liu and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM,* vol. 63, pp. 68-77, 2019.

[91] C. Henin and D. Le Métayer, "Towards a generic framework for black-box explanations of algorithmic decision systems," in *IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)*, 2019.

[92] I. Nunes and D. Jannach, "A systematic review and taxonomy of explanations in decision support and recommender systems," *User Modeling and User-Adapted Interaction,* vol. 27, pp. 393-444, 2017.

[93] P. M. Singh, M. Van Sinderen and R. Wieringa, "Reference architecture for integration platforms," in *2017 IEEE 21st International Enterprise Distributed Object Computing Conference (EDOC)*, 2017.

[94] Y. Wu, Y. Xu and J. Li, "Feature construction for fraudulent credit card cash-out detection," *Decision Support Systems,* vol. 127, p. 113155, 2019.

[95] X. Zhou, S. Cheng, M. Zhu, C. Guo, S. Zhou, P. Xu, Z. Xue and W. Zhang, "A state of the art survey of data mining-based fraud detection and credit scoring," in *MATEC Web of Conferences*, 2018.

[96] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, pp. 273-297, 1995.

[97] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, 1995.

[98] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016.

[99] R. Jääskeläinen, "Think-aloud protocol," *Handbook of translation studies,* vol. 1, pp. 371-374, 2010.

[100] A. Bussone, S. Stumpf and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *2015 International Conference on Healthcare Informatics*, 2015.

[101] R. K. Yin, Case study research and applications: Design and methods, Sage publications, 2017.

[102] M. Madsen and S. Gregor, "Measuring human-computer trust," in *11th australasian conference on information systems*, 2000.

[103] S. Sousa, D. Lamas and P. Dias, "A model for Human-computer Trust," in *International Conference on Learning and Collaboration Technologies*, 2014.

[104] D. Holliday, S. Wilson and S. Stumpf, "User trust in intelligent systems: A journey over time," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 2016.

[105] D. S. Johnson, "Achieving customer value from electronic channels through identity commitment, calculative commitment, and trust in technology," *Journal of interactive marketing,* vol. 21, pp. 2-22, 2007.

[106] R. Larasati, A. D. Liddo and E. Motta, "The effect of explanation styles on user's trust," in *2020 Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies*.

[107] K. Schaefer, "The perception and measurement of human-robot trust.(2013)," 2013.

[108] L. Wang, G. A. Jamieson and J. G. Hollands, "Trust and reliance on an automated combat identification system," *Human factors,* vol. 51, pp. 281-291, 2009.

[109] P. Pu and L. Chen, "Trust building with explanation interfaces," in *Proceedings of the 11th international conference on Intelligent user interfaces*, 2006.

[110] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce and H. P. Beck, "The role of trust in automation reliance," *International journal of human-computer studies,* vol. 58, pp. 697-718, 2003.

[111] B. Cahour and J.-F. Forzy, "Does projection into use improve trust and exploration? An example with a cruise control system," *Safety science,* vol. 47, pp. 1260-1270, 2009.

[112] I. L. Singh, R. Molloy and R. Parasuraman, "Automation-induced" complacency": Development of the complacency-potential rating scale," *The International Journal of Aviation Psychology,* vol. 3, pp. 111-122, 1993.

[113] J.-Y. Jian, A. M. Bisantz and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International journal of cognitive ergonomics,* vol. 4, pp. 53-71, 2000.

[114] S. M. Merritt, H. Heimbaugh, J. LaChapell and D. Lee, "I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system," *Human factors,* vol. 55, pp. 520-534, 2013.

[115] R. R. Hoffman, S. T. Mueller, G. Klein and J. Litman, "Metrics for explainable AI: Challenges and prospects," *arXiv preprint arXiv:1812.04608,* 2018.

[116] D. H. Mcknight, M. Carter, J. B. Thatcher and P. F. Clay, "Trust in a specific technology: An investigation of its components and measures," *ACM Transactions on management information systems (TMIS),* vol. 2, pp. 1-25, 2011.

[117] F. Yang, Z. Huang, J. Scholtz and D. L. Arendt, "How do visual explanations foster end users' appropriate trust in machine learning?," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 2020.

[118] D. J. McAllister, "Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations," *Academy of management journal,* vol. 38, pp. 24-59, 1995.

[119] M. Yin, J. Wortman Vaughan and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019.

[120] S. Antifakos, N. Kern, B. Schiele and A. Schwaninger, "Towards improving trust in context-aware systems by displaying system confidence," in *Proceedings of the 7th international conference on Human computer interaction with mobile devices \& services*, 2005.

[121] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician,* vol. 46, pp. 175-185, 1992.

[122] E. Schubert and P. J. Rousseeuw, "Faster k-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms," in *International Conference on Similarity Search and Applications*, 2019.

[123] S. Lundberg and S.-I. Lee, "An unexpected unity among methods for interpreting model predictions," *arXiv preprint arXiv:1611.07478,* 2016.

[124] M. T. Ribeiro, S. Singh and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[125] J. Zhu, A. Liapis, S. Risi, R. Bidarra and G. M. Youngblood, "Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation," in *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, 2018.

[126] J. Zhou, H. Hu, Z. Li, K. Yu and F. Chen, "Physiological Indicators for User Trust in Machine Learning with Influence Enhanced Fact-Checking," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 2019.

[127] J. Zhou and F. Chen, "2D Transparency Space—Bring Domain Users and Machine Learning Experts Together," in *Human and Machine Learning*, Springer, 2018, pp. 3-19.

[128] J. Zhou and F. Chen, "Towards Trustworthy Human-AI Teaming under Uncertainty," in *IJCAI 2019 Workshop on Explainable AI (XAI)*, 2019.

[129] X. Zhao, Y. Wu, D. L. Lee and W. Cui, "iforest: Interpreting random forests via visual analytics," *IEEE transactions on visualization and computer graphics,* vol. 25, pp. 407-416, 2018.

[130] J. Zhang, Y. Wang, P. Molino, L. Li and D. S. Ebert, "Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models," *IEEE transactions on visualization and computer graphics,* vol. 25, pp. 364-373, 2018.

[131] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, 2014.

[132] C. Zednik, "Solving the black box problem: A normative framework for explainable artificial intelligence," *Philosophy \& Technology,* pp. 1-24, 2019.

[133] C. Xie, W. Chen, X. Huang, Y. Hu, S. Barlowe and J. Yang, "VAET: A visual analytics approach for e-transactions time-series," *IEEE transactions on visualization and computer graphics,* vol. 20, pp. 1743-1752, 2014.

[134] L. S. Whitmore, A. George and C. M. Hudson, *Explicating feature contribution using Random Forest proximity distances,* 2018.

[135] F. Westphal, N. Lavesson and H. Grahn, "A Case for Guided Machine Learning," 2019, pp. 353-361.

[136] D. K. I. Weidele, J. D. Weisz, E. Oduor, M. Muller, J. Andres, A. Gray and D. Wang, "AutoAIViz: opening the blackbox of automated artificial intelligence with conditional parallel coordinates," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 2020.

[137] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl and P. MacNeille, "A bayesian framework for learning rule sets for interpretable classification," *The Journal of Machine Learning Research,* vol. 18, pp. 2357-2393, 2017.

[138] D. Wang, J. Lin, P. Cui, Q. Jia, Z. Wang, Y. Fang, Q. Yu, J. Zhou, S. Yang and Y. Qi, "A Semi-supervised Graph Attentive Network for Financial Fraud Detection," in *2019 IEEE International Conference on Data Mining (ICDM)*, 2019.

[139] S. Wachter, B. Mittelstadt and C. Russell, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR,* 2017.

[140] J. Waa, J. Diggelen and M. Neerincx, "The design and validation of an intuitive confidence measure," *memory,* vol. 2, p. 1, 2018.

[141] J. Waa, M. Robeer, J. Diggelen, M. Brinkhuis and M. Neerincx, "Contrastive explanations with local foil trees," *arXiv preprint arXiv:1806.07470,* 2018.

[142] J. Voogd, P. Heer, K. Veltman, P. Hanckmann and J. Lith, "Using Relational Concept Networks for Explainable Decision Support," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 2019.

[143] S. Vojíř, V. Zeman, J. Kuchař and T. Kliegr, "EasyMiner. eu: Web framework for interpretable machine learning based on rules and frequent itemsets," *Knowledge-Based Systems,* vol. 150, pp. 111-115, 2018.

[144] G. Vilone and L. Longo, "Explainable Artificial Intelligence: a Systematic Review," *arXiv preprint arXiv:2006.00093,* 2020.

[145] J. Venable, J. Pries-Heje and R. Baskerville, "A comprehensive framework for evaluation in design science research," in *International Conference on Design Science Research in Information Systems*, 2012.

[146] B. Vasu and C. Long, "Iterative and Adaptive Sampling with Spatial Attention for Black-Box Model Explanations," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020.

[147] A. Van Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," *arXiv preprint arXiv:1907.02584,* 2019.

[148] K. Vaculík and L. Popelínský, "Dgrminer: Anomaly detection and explanation in dynamic graphs," in *International Symposium on Intelligent Data Analysis*, 2016.

[149] R. Turner, "A model explanation system," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.

[150] G. Tolomei, F. Silvestri, A. Haines and M. Lalmas, "Interpretable Predictions of Tree-Based Ensembles via Actionable Feature Tweaking," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2017.

[151] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): towards medical XAI," *arXiv preprint arXiv:1907.07374,* 2019.

[152] S. Tan, M. Soloviev, G. Hooker and M. T. Wells, *Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable,* 2016.

[153] Y. Sun, H. Chockler, X. Huang and D. Kroening, "Explaining Deep Neural Networks Using Spectrum-Based Fault Localization," *arXiv preprint arXiv:1908.02374,* 2019.

[154] W. Souillard-Mandar, R. Davis, C. Rudin, R. Au and D. Penney, "Interpretable machine learning models for the digital clock drawing test," *arXiv preprint arXiv:1606.07163,* 2016.

[155] K. Sokol and P. Flach, "One explanation does not fit all," *KI- Künstliche Intelligenz,* pp. 1-16, 2020.

[156] A. Smith-Renner, R. Rua and M. Colony, "Towards an Explainable Threat Detection Tool.," in *IUI Workshops*, 2019.

[157] S. Singh, M. T. Ribeiro and C. Guestrin, "Programs as black-box explanations," *arXiv preprint arXiv:1611.07579,* 2016.

[158] A. Shrikumar, P. Greenside, A. Shcherbina and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713,* 2016.

[159] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[160] J. M. Schoenborn and K.-D. Althoff, "Recent Trends in XAI: A Broad Overview on current Approaches, Methodologies and Interactions.," in *ICCBR Workshops*, 2019.

[161] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke and D. A. Keim, "Towards a rigorous evaluation of XAI Methods on Time Series," *arXiv preprint arXiv:1909.07082,* 2019.

[162] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin and T. H{\"o}llerer, "I can do better than your AI: Expertise and explanations," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019.

[163] M. N. K. Saunders, P. Lewis, A. Thornhill and A. Bristow, "Understanding research philosophy and approaches to theory development," in *Research Methods for Business Students*, M. N. K. Saunders, P. Lewis and A. Thornhill, Eds., Harlow, : Pearson Education, 2015, pp. 122-161.

[164] M. Sato, B. Ahsan, K. Nagatani, T. Sonoda, Q. Zhang and T. Ohkuma, "Explaining recommendations using contexts," in *23rd International Conference on Intelligent User Interfaces*, 2018.

[165] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: interpreting, explaining and visualizing deep learning*, Springer, 2019, pp. 5-22.

[166] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen and K.-R. Müller, Explainable AI: interpreting, explaining and visualizing deep learning, vol. 11700, Springer Nature, 2019.

[167] C. Russell, "Efficient search for diverse coherent explanations," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[168] A. Rosenfeld and A. Richardson, "Explainability in human--agent systems," *Autonomous Agents and Multi-Agent Systems,* vol. 33, pp. 673-705, 2019.

[169] R. Rieke, M. Zhdanova, J. Repp, R. Giot and C. Gaber, "Fraud detection in mobile payments utilizing process behavior analysis," in *2013 International Conference on Availability, Reliability and Security*, 2013.

[170] A. Richardson and A. Rosenfeld, "A survey of interpretability and explainability in human-agent systems," in *XAI Workshop on Explainable Artificial Intelligence*, 2018.

[171] M. T. Ribeiro, S. Singh and C. Guestrin, "Nothing else matters: model-agnostic explanations by identifying prediction invariance," *arXiv preprint arXiv:1611.05817,* 2016.

[172] M. T. Ribeiro, S. Singh and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386,* 2016.

[173] X. Renard, T. Laugel, M.-J. Lesot, C. Marsala and M. Detyniecki, "Detecting Potential Local Adversarial Examples for Human-Interpretable Defense," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2018.

[174] A. Ray, Y. Yao, R. Kumar, A. Divakaran and G. Burachas, "Can you explain that? Lucid explanations help human-AI collaborative image retrieval," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2019.

[175] G. Ras, M. Gerven and P. Haselager, "Explanation methods in deep learning: Users, values, concerns and challenges," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Springer, 2018, pp. 19-36.

[176] Y. Ramon, D. Martens, F. Provost and T. Evgeniou, "Counterfactual explanation algorithms for behavioral and textual data," *arXiv preprint arXiv:1912.01819,* 2019.

[177] D. V. Pynadath, M. J. Barnes, N. Wang and J. Y. C. Chen, "Transparency communication for machine learning in human-automation interaction," in *Human and Machine Learning*, Springer, 2018, pp. 75-90.

[178] V. Putnam and C. Conati, "Exploring the Need for Explainable Artificial Intelligence (XAI) in Intelligent Tutoring Systems (ITS).," in *IUI Workshops*, 2019.

[179] L. Pusztová, F. Babič, J. Paralič and Z. Paraličová, "How to Improve the Adaptation Phase of the CBR in the Medical Domain," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 2019.

[180] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig and Z. C. Lipton, "Learning to deceive with attention-based explanations," *arXiv preprint arXiv:1909.07913,* 2019.

[181] A. Preece, D. Braines, F. Cerutti and T. Pham, "Explainable AI for Intelligence Augmentation in Multi-Domain Operations," *arXiv preprint arXiv:1910.07563,* 2019.

[182] A. Preece, "Asking 'Why'in AI: Explainability of intelligent systems--perspectives and challenges," *Intelligent Systems in Accounting, Finance and Management,* vol. 25, pp. 63-72, 2018.

[183] R. Pierrard, J.-P. Poli and C. Hudelot, "Learning fuzzy relations and properties for explainable artificial intelligence," in *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2018.

[184] L. T. K. Phung, V. T. N. Chau and N. H. Phung, "Extracting rule RF in educational data classification: from a random forest to interpretable refined rules," in *2015 International Conference on Advanced Computing and Applications (ACOMP)*, 2015.

[185] V. Petsiuk, A. Das and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," *arXiv preprint arXiv:1806.07421,* 2018.

[186] D. Petkovic, R. B. Altman, M. Wong and A. Vigil, "Improving the explainability of Random Forest classifier-user centered approach.," in *PSB*, 2018.

[187] D. G. Perez and M. M. Lavalle, "Outlier detection applying an innovative user transaction modeling with automatic explanation," in *2011 IEEE Electronics, Robotics and Automotive Mechanics Conference*, 2011.

[188] A. Papenmeier, G. Englebienne and C. Seifert, "How model accuracy and explanation fidelity influence user trust," *arXiv preprint arXiv:1907.12652,* 2019.

[189] D. Olszewski, "Fraud detection using self-organizing map visualizing the user profiles," *Knowledge-Based Systems,* vol. 70, pp. 324-334, 2014.

[190] F. Offert, "" I know it when I see it". Visualization and Intuitive Interpretability," *arXiv preprint arXiv:1711.08042,* 2017.

[191] E. Novikova and I. Kotenko, "Visualization-Driven Approach to Fraud Detection in the Mobile Money Transfer Services," in *Algorithms, Methods, and Applications in Mobile Computing and Communications*, IGI Global, 2019, pp. 205-236.

[192] A. T. Nguyen, A. Kharosekar, S. Krishnan, S. Krishnan, E. Tate, B. C. Wallace and M. Lease, "Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018.

[193] M. A. Neerincx, J. Waa, F. Kaptein and J. Diggelen, "Using perceptual and cognitive explanations for enhanced human-agent team performance," in *International Conference on Engineering Psychology and Cognitive Ergonomics*, 2018.

[194] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman and F. Doshi-Velez, "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1802.00682,* 2018.

[195] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl and B. Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv preprint arXiv:1901.04592,* 2019.

[196] G. Montavon, W. Samek and K.-R. M{\"u}ller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing,* vol. 73, pp. 1-15, 2018.

[197] T. Mokoena, O. Lebogo, A. Dlaba and V. Marivate, "Bringing sequential feature explanations to life," in *2017 IEEE AFRICON*, 2017.

[198] T. Miller, P. Howe and L. Sonenberg, "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences," *arXiv preprint arXiv:1712.00547,* 2017.

[199] D. A. Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Advances in Neural Information Processing Systems*, 2018.

[200] M. Mejia-Lavalle, "Outlier detection with innovative explanation facility over a very large financial database," in *2010 IEEE Electronics, Robotics and Automotive Mechanics Conference*, 2010.

[201] F. J. Maymí and R. Thomson, "Human-machine teaming and cyberspace," in *International Conference on Augmented Cognition*, 2018.

[202] D. L. Marino, C. S. Wickramasinghe and M. Manic, "An adversarial approach for explainable ai in intrusion detection systems," in *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*, 2018.

[203] P. Madumal, T. Miller, L. Sonenberg and F. Vetere, "Explainable reinforcement learning through a causal lens," *arXiv preprint arXiv:1905.10958,* 2019.

[204] O. Mac Aodha, S. Su, Y. Chen, P. Perona and Y. Yue, "Teaching categories to human learners with visual explanations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[205] Y. Lou, R. Caruana and J. Gehrke, "Intelligible models for classification and regression," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012.

[206] Y. Lou, R. Caruana, J. Gehrke and G. Hooker, "Accurate Intelligible Models with Pairwise Interactions," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2013.

[207] X. Liu, P. Hu, Z. Mao, P.-C. Kuo, P. Li, C. Liu, J. Hu, D. Li, D. Cao, R. G. Mark and others, "Interpretable Machine Learning Model for Early Prediction of Mortality in Elderly Patients with Multiple Organ Dysfunction Syndrome (MODS): a Multicenter Retrospective Study and Cross Validation," *arXiv preprint arXiv:2001.10977,* 2020.

[208] Z. C. Lipton, "The mythos of model interpretability," *Queue,* vol. 16, pp. 31-57, 2018.

[209] Z. C. Lipton, "The Doctor Just Won't Accept That!," *arXiv preprint arXiv:1711.08037,* 2017.

[210] H. Lin, S. Gao, D. Gotz, F. Du, J. He and N. Cao, "Rclens: Interactive rare category exploration and identification," *IEEE transactions on visualization and computer graphics,* vol. 24, pp. 2223-2237, 2017.

[211] B. Y. Lim, Q. Yang, A. M. Abdul and D. Wang, "Why these Explanations? Selecting Intelligibility Types for Explanation Goals.," in *IUI Workshops*, 2019.

[212] S. Li and K.-T. Cheng, "Visualizing the Decision-making Process in Deep Neural Decision Forest.," in *CVPR Workshops*, 2019.

[213] O. Li, H. Liu, C. Chen and C. Rudin, "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[214] B. Letham, C. Rudin, T. H. McCormick, D. Madigan and others, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics,* vol. 9, pp. 1350-1371, 2015.

[215] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein and J. Kuntner, "Visual Analytics for Fraud Detection: Focusing on Profile Analysis.," in *EuroVis (Posters)*, 2016.

[216] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein and J. Kuntner, "Network Analysis for Financial Fraud Detection.," in *EuroVis (Posters)*, 2018.

[217] T. Le, S. Wang and D. Lee, "Why X rather than Y? Explaining Neural Model'Predictions by Generating Intervention Counterfactual Samples," *arXiv preprint arXiv:1911.02042,* 2019.

[218] T. Laugel, X. Renard, M.-J. Lesot, C. Marsala and M. Detyniecki, "Defining locality for surrogates in post-hoc interpretablity," *arXiv preprint arXiv:1806.07498,* 2018.

[219] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard and M. Detyniecki, "Comparison-based inverse classification for interpretability in machine learning," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2018.

[220] N. Lauffer and U. Topcu, "Human-understandable explanations of infeasibility for resource-constrained scheduling problems," in *ICAPS 2019 Workshop XAIP*, 2019.

[221] H. Lakkaraju, S. H. Bach and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

[222] H. Lakkaraju, E. Kamar, R. Caruana and J. Leskovec, "Interpretable & explorable approximations of black box models," *arXiv preprint arXiv:1707.01154,* 2017.

[223] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[224] V. Lai, J. Z. Cai and C. Tan, "Many Faces of Feature Importance: Comparing Built-in and Post-hoc Feature Importance in Text Classification," *arXiv preprint arXiv:1910.08534,* 2019.

[225] V. Lai, S. Carton and C. Tan, "Harnessing Explanations to Bridge AI and Humans," *arXiv preprint arXiv:2003.07370,* 2020.

[226] I. Lage, A. Ross, S. J. Gershman, B. Kim and F. Doshi-Velez, "Human-in-the-loop interpretability prior," in *Advances in neural information processing systems*, 2018.

[227] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun and J. Choo, "Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records," *IEEE transactions on visualization and computer graphics,* vol. 25, pp. 299-309, 2018.

[228] R. Krishnan, G. Sivakumar and P. Bhattacharya, "Extracting decision trees from trained neural networks," *Pattern recognition,* vol. 32, 1999.

[229] B. Krarup, M. Cashmore, D. Magazzeni and T. Miller, "Model-based contrastive explanations for explainable planning," 2019.

[230] P. W. Koh and P. Liang, *Understanding Black-box Predictions via Influence Functions,* 2017.

[231] M. Kobayashi and T. Ito, "A transactional relationship visualization system in Internet auctions," in *2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'07)*, 2007.

[232] S. Ko, I. Cho, S. Afzal, C. Yau, J. Chae, A. Malik, K. Beck, Y. Jang, W. Ribarsky and D. S. Ebert, "A survey on visual analysis approaches for financial data," in *Computer Graphics Forum*, 2016.

[233] T. W. Kim, "Explainable artificial intelligence (XAI), the goodness criteria and the grasp-ability test," *arXiv preprint arXiv:1810.09598,* 2018.

[234] B. Kim, C. Rudin and J. A. Shah, "The bayesian case model: A generative approach for case-based reasoning and prototype classification," in *Advances in neural information processing systems*, 2014.

[235] B. Kim, R. Khanna and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in neural information processing systems*, 2016.

[236] M. T. Keane and B. Smyth, "Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI)," *arXiv preprint arXiv:2005.13997,* 2020.

[237] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach and J. Wortman Vaughan, "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[238] A. Kasirzadeh, "Mathematical decisions and non-causal elements of explainable AI," *arXiv preprint arXiv:1910.13607,* 2019.

[239] B. Kaplan and J. A. Maxwell, "Qualitative research methods for evaluating computer information systems," in *Evaluating the organizational impact of healthcare information systems*, Springer, 2005, pp. 30-55.

[240] A. J. Johs, M. Lutts and R. O. Weber, "Measuring explanation quality in XCBR," in *Proceedings of ICCBR*, 2018.

[241] D. Janzing, L. Minorics and P. Bl{\"o}baum, "Feature relevance quantification in explainable AI: A causal problem," in *International Conference on Artificial Intelligence and Statistics*, 2020.

[242] S. R. Islam, W. Eberle and S. K. Ghafoor, "Towards Quantification of Explainability in Explainable Artificial Intelligence Methods," in *The Thirty-Third International Flairs Conference*, 2020.

[243] S. R. Islam, W. Eberle, S. K. Ghafoor, A. Siraj and M. Rogers, "Domain knowledge aided explainable artificial intelligence for intrusion detection and response," *arXiv preprint arXiv:1911.09853,* 2019.

[244] M. L. Huang, J. Liang and Q. V. Nguyen, "A visualization approach for frauds detection in financial market," in *2009 13th International Conference Information Visualisation*, 2009.

[245] H. Horacek, "Requirements for Conceptual Representations of Explanations and How Reasoning Systems Can Serve Them," in *Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017)*, 2017.

[246] A. Holzinger, C. Biemann, C. S. Pattichis and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?," *arXiv preprint arXiv:1712.09923,* 2017.

[247] J. Hois, D. Theofanou-Fuelbier and A. J. Junk, "How to Achieve Explainability and Transparency in Human AI Interaction," in *International Conference on Human-Computer Interaction*, 2019.

[248] R. Hoffman, T. Miller, S. T. Mueller, G. Klein and W. J. Clancey, "Explaining explanation, part 4: a deep dive on deep nets," *IEEE Intelligent Systems,* vol. 33, pp. 87-95, 2018.

[249] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele and T. Darrell, "Generating visual explanations," in *European Conference on Computer Vision*, 2016.

[250] L. A. Hendricks, R. Hu, T. Darrell and Z. Akata, "Generating counterfactual explanations with natural language," *arXiv preprint arXiv:1806.09809,* 2018.

[251] S. Hara and K. Hayashi, "Making tree ensembles interpretable," *arXiv preprint arXiv:1606.05390,* 2016.

[252] M. C. Hao, U. Dayal, R. K. Sharma, D. A. Keim and H. Janetzko, "Visual analytics of large multidimensional data using variable binned scatter plots," in *Visualization and Data Analysis 2010*, 2010.

[253] N. Gupta, D. Eswaran, N. Shah, L. Akoglu and C. Faloutsos, "Beyond outlier detection: Lookout for pictorial explanation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2018.

[254] S. Guo, Z. Jin, Q. Chen, D. Gotz, H. Zha and N. Cao, "Visual Anomaly Detection in Event Sequence Data," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019.

[255] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri and F. Turini, "Factual and counterfactual explanations for black box decision making," *IEEE Intelligent Systems,* vol. 34, pp. 14-23, 2019.

[256] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR),* vol. 51, pp. 1-42, 2018.

[257] O. Gomez, S. Holter, J. Yuan and E. Bertini, "ViCE: visual counterfactual explanations for machine learning models," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 2020.

[258] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg and A. Holzinger, "Explainable AI: the new 42?," in *International cross-domain conference for machine learning and knowledge extraction*, 2018.

[259] B. Ghai, Q. V. Liao, Y. Zhang, R. Bellamy and K. Mueller, "Explainable Active Learning (XAL): An Empirical Study of How Local Explanations Impact Annotator Experience," *arXiv preprint arXiv:2001.09219,* 2020.

[260] A. H. Gee, D. Garcia-Olano, J. Ghosh and D. Paydarfar, "Explaining deep classification of time-series data with learned prototypes," *arXiv preprint arXiv:1904.08935,* 2019.

[261] G. Gal, K. Singh and P. Best, "Interactive visual analysis of anomalous accounts payable transactions in SAP enterprise systems," *Managerial Auditing Journal,* 2016.

[262] C. Frye, I. Feige and C. Rowat, "Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability," *arXiv preprint arXiv:1910.06358,* 2019.

[263] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[264] A. Ferrario, M. Loi and E. Vigan{\`o}, "In AI we trust incrementally: a multi-layer model of trust to analyze human-artificial intelligence interactions," *Philosophy \& Technology,* pp. 1-17, 2019.

[265] C. Fernandez, F. Provost and X. Han, "Explaining data-driven decisions made by ai systems: The counterfactual approach," *arXiv preprint arXiv:2001.07417,* 2020.

[266] S. Feng and J. Boyd-Graber, "What can AI do for me? evaluating machine learning interpretations in cooperative play," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019.

[267] N. El Bekri, J. Kling and M. F. Huber, "A study on trust in black box models and post-hoc explanations," in *International Workshop on Soft Computing Models in Industrial and Environmental Applications*, 2019.

[268] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison and M. O. Riedl, "Automated rationale generation: a technique for explainable AI and its effects on human perceptions," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019.

[269] M. Dumas, M. J. McGuffin and V. L. Lemieux, "Financevis. net-a visual survey of financial data visualizations," in *Poster Abstracts of IEEE Conference on Visualization*, 2014.

[270] D. Doran, S. Schulz and T. R. Besold, "What does explainable AI really mean? A new conceptualization of perspectives," *arXiv preprint arXiv:1710.00794,* 2017.

[271] Y. Dong, H. Su, J. Zhu and F. Bao, "Towards interpretable deep neural networks by leveraging adversarial examples," *arXiv preprint arXiv:1708.05493,* 2017.

[272] Y. Dong, H. Su, J. Zhu and B. Zhang, "Improving interpretability of deep neural networks with semantic information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[273] J. Dodge, S. Penney, A. Anderson and M. M. Burnett, "What Should Be in an XAI Explanation? What IFT Reveals.," in *IUI Workshops*, 2018.

[274] L. Ding, "Human knowledge in constructing AI systems—Neural logic networks approach towards an explainable AI," *Procedia Computer Science,* vol. 126, pp. 1561-1570, 2018.

[275] W. Didimo, G. Liotta and F. Montecchiani, "Vis4AUI: Visual Analysis of Banking Activity Networks.," in *GRAPP/IVAPP*, 2012.

[276] W. Didimo, G. Liotta, F. Montecchiani and P. Palladino, "An advanced network visualization system for financial crime detection," in *2011 IEEE Pacific visualization symposium*, 2011.

[277] E. Di Giacomo, W. Didimo, G. Liotta and P. Palladino, "Visual analysis of financial crimes: [system paper]," in *Proceedings of the International Conference on Advanced Visual Interfaces*, 2010.

[278] D. Deutch and N. Frost, "Constraints-based explanations of classifications," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 2019.

[279] H. Deng, "Interpreting tree ensembles with intrees," *International Journal of Data Science and Analytics,* vol. 7, pp. 277-287, 2019.

[280] S. Das, M. R. Islam, N. K. Jayakodi and J. R. Doppa, "Active Anomaly Detection via Ensembles: Insights, Algorithms, and Interpretability," *arXiv preprint arXiv:1901.08930,* 2019.

[281] D. Das and S. Chernova, "Leveraging rationales to improve human task performance," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 2020.

[282] P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," *Information Sciences,* vol. 225, pp. 1-17, 2013.

[283] D. Collaris and J. J. Wijk, "ExplainExplore: Visual Exploration of Machine Learning Explanations," in *2020 IEEE Pacific Visualization Symposium (PacificVis)*, 2020.

[284] D. Colla, E. Mensa, D. P. Radicioni and A. Lieto, "Tell me why: Computational explanation of conceptual similarity judgments," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2018.

[285] L. Chu, X. Hu, J. Hu, L. Wang and J. Pei, "Exact and consistent interpretation for piecewise linear neural networks: A closed form solution," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery \& Data Mining*, 2018.

[286] M. Chmielewski and P. St{\k{a}}por, "Money Laundering Analytics Based on Contextual Analysis. Application of Problem Solving Ontologies in Financial Fraud Identification and Recognition," in *Information Systems Architecture and Technology: Proceedings of 37th International Conference on Information Systems Architecture and Technology--ISAT 2016--Part I*, 2017.

[287] M. Chmielewski and P. Stąpor, "Hidden information retrieval and evaluation method and tools utilising ontology reasoning applied for financial fraud analysis.," in *MATEC Web of Conferences*, 2018.

[288] R. Chhatwal, P. Gronvall, N. Huber-Fliflet, R. Keeling, J. Zhang and H. Zhao, "Explainable Text Classification in Legal Document Review A Case Study of Explainable Predictive Coding," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018.

[289] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper and H. Zhu, "Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019.

[290] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern and A. Sudjianto, "Wirevis: Visualization of categorical, time-varying data from financial transactions," in *2007 IEEE Symposium on Visual Analytics Science and Technology*, 2007.

[291] A. Chander, R. Srinivasan, S. Chelian, J. Wang and K. Uchino, "Working with Beliefs: AI Transparency in the Enterprise.," in *IUI Workshops*, 2018.

[292] A. Chander and R. Srinivasan, "Evaluating explanations by cognitive value," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 2018.

[293] T. Chakraborti, K. P. Fadnis, K. Talamadupula, M. Dholakia, B. Srivastava, J. O. Kephart and R. K. E. Bellamy, "Visualizations for an explainable planning agent," *arXiv preprint arXiv:1709.04517,* 2017.

[294] D. V. Carvalho, E. M. Pereira and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics,* vol. 8, p. 832, 2019.

[295] S. Carton, "The Design and Evaluation of Neural Attention Mechanisms for Explaining Text Classifiers," 2019.

[296] S. Carton, Q. Mei and P. Resnick, "Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2020.

[297] A. Carrington, P. Fieguth and H. Chen, "Measures of model interpretability for model selection," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 2018.

[298] M. Carminati, R. Caron, F. Maggi, I. Epifani and S. Zanero, "BankSealer: An online banking fraud analysis and decision support system," in *IFIP International Information Security Conference*, 2014.

[299] C. J. Cai, J. Jongejan and J. Holbrook, "The effects of example-based explanations in a machine learning interface," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019.

[300] C. J. Cai, S. Winter, D. Steiner, L. Wilcox and M. Terry, "" Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making," *Proceedings of the ACM on Human-computer Interaction,* vol. 3, pp. 1-24, 2019.

[301] R. M. J. Byrne, "Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning.," in *IJCAI*, 2019.

[302] W. Briguglio and S. Saad, "Interpreting Machine Learning Malware Detectors Which Leverage N-gram Analysis," in *International Symposium on Foundations and Practice of Security*, 2019.

[303] O. Biran and C. Cotton, "Explanation and justification in machine learning: A survey," in *IJCAI-17 workshop on explainable AI (XAI)*, 2017.

[304] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," *The Annals of Applied Statistics,* pp. 2403-2424, 2011.

[305] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen and K.-R. M{\"u}ller, "How to explain individual classification decisions," *The Journal of Machine Learning Research,* vol. 11, pp. 1803-1831, 2010.

[306] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. M{\"u}ller and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one,* vol. 10, p. e0130140, 2015.

[307] K. Böhmer and S. Rinderle-Ma, "Mining association rules for anomaly detection in dynamic process runtime behavior and explaining the root cause to users," *Information Systems,* vol. 90, p. 101438, 2020.

[308] A. Atrey, K. Clary and D. Jensen, "Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep RL," *arXiv preprint arXiv:1912.05743,* 2019.

[309] S. O. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," *arXiv preprint arXiv:1908.07442,* 2019.

[310] S. Anjomshoae, K. Främling and A. Najjar, "Explanations of black-box model predictions by contextual importance and utility," in *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 2019.

[311] K. Amarasinghe and M. Manic, "Improving user trust on deep neural networks based intrusion detection systems," in *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*, 2018.

[312] A. Alqaraawi, M. Schuessler, P. Weiß, E. Costanza and N. Berthouze, "Evaluating saliency map explanations for convolutional neural networks: a user study," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 2020.

[313] N. Alexandrov, "Explainable AI decisions for human-autonomy interactions," in *17th AIAA Aviation Technology, Integration, and Operations Conference*, 2017.

[314] A. R. Akula, S. Todorovic, J. Y. Chai and S.-C. Zhu, "Natural Language Interaction with Explainable AI Models.," in *CVPR Workshops*, 2019.

[315] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and information systems,* vol. 41, pp. 647-665, 2014.

[316] E. Štrumbelj, I. Kononenko and M. R. Šikonja, "Explaining instance classifications with interactions of subsets of feature values," *Data \& Knowledge Engineering,* vol. 68, pp. 886-904, 2009.