References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*(52), 138–160. doi:10.1109/access.2018.2870052

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the Robustness of Interpretability Methods. *2018 ICML Workshop on Human Interpretability in Machine Learning*. doi:10.48550/arXiv.1806.08049

Anowar, F., & Sadaoui, S. (2020). Incremental Neural-Network Learning for Big Fraud Data. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, *1*(1), 1–4. doi:10.1109/smc42975.2020.9283136

Batageri, A., & Kumar, S. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, *2*(1), 35–41. doi:10.1016/j.gltp.2021.01.006

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8). doi:10.3390/electronics8080832

Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I think I get your point, ai! the illusion of explanatory depth in explainable ai. *26th International Conference on Intelligent User Interfaces*. doi:10.1145/3397481.3450644

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, *41*(10), 4915–4928. doi:10.1016/j.eswa.2014.02.026

Darias, J. M., Caro-Martínez, M., Díaz-Agudo, B., & Recio-Garcia, J. A. (2022). Using case-based reasoning for capturing expert knowledge on explanation methods. *Case-Based Reasoning Research and Development*, *13405*, 3–17. doi:10.1007/978-3-031-14923-8\_1

ElShawi, R., Sherif, Y., Al‐Mallah, M., & Sakr, S. (2020). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633–1650. doi:10.1111/coin.12410

Evans, B. P., Xue, B., & Zhang, M. (2019). What’s inside the black-box? *Proceedings of the Genetic and Evolutionary Computation Conference*. doi:10.1145/3321707.3321726

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, *34*(6), 14–23. doi:10.1109/mis.2019.2957223

Hanafy, M., & Ming, R. (2022). Classification of the insureds using Integrated Machine Learning Algorithms: A comparative study. *Applied Artificial Intelligence*, *36*(1). doi:10.1080/08839514.2021.2020489

Honegger, M. (2018). Shedding light on Black Box Machine Learning Algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions. Retrieved from https://arxiv.org/abs/1808.05054v1

Ignatiev, A. (2020). Towards trustable explainable ai. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 5154–5158. doi:10.24963/ijcai.2020/726

Jacob, V., Song, F., Stiegler, A., Rad, B., Diao, Y., & Tatbul, N. (2021). Exathlon: A Benchmark for Explainable Anomaly Detection over Time Series. *Proceedings of the VLDB Endowment*, *14*(11), 2613–2626. doi:10.14778/3476249.3476307 "ED Benchmark. While a user study may be the best way to evaluate
the usefulness of explanations, it is not always available and
may come at a high cost. Therefore, our benchmark aims to provide
automated evaluation of ED methods based on intuitive metrics,
namely, conciseness, consistency, and accuracy, as well as their various
variants."

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021). How can I choose an explainer? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. doi:10.1145/3442188.3445941

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. doi:10.1145/3313831.3376219

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1675–1684. doi:10.1145/2939672.2939874

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (Vol. 30). essay, NeurIPS Proceedings.

Marcilio, W. E., & Eler, D. M. (2020). From explanations to feature selection: Assessing shap values as feature selection mechanism. *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340–347. doi:10.1109/sibgrapi51738.2020.00053

Martínez, M. A., Nadj, M., Langner, M., Toreini, P., & Maedche, A. (2023). Does this explanation help? designing local model-agnostic explanation representations and an experimental evaluation using eye-tracking technology. *ACM Transactions on Interactive Intelligent Systems*. doi:10.1145/3607145 "Artificial Intelligence (AI) is playing an increasingly significant role in high-stake domains such as Finance...This problem is reinforced by the popularity of deep learning models, which are hard to understand even by experts [23]. Due to the need for users to understand the logic of these systems, new regulations have been enacted..."

Moreira, C., Chou, Y.-L., Velmurugan, M., Ouyang, C., Sindhgatta, R., & Bruza, P. (2020). Linda-BN: An interpretable probabilistic approach for demystifying Black-Box Predictive Models. *Decision Support Systems*, *150*, 113561. doi:10.1016/j.dss.2021.113561

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. doi:10.1145/3351095.3372850

Nascita, A., Montieri, A., Aceto, G., Ciuonzo, D., Persico, V., & Pescape, A. (2021). Unveiling mimetic: Interpreting deep learning traffic classifiers via xai techniques. *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, 455–460. doi:10.1109/csr51186.2021.9527948

Nesvijevskaia, A., Ouillade, S., Guilmin, P., & Zucker, J.-D. (2021). The accuracy versus interpretability trade-off in fraud detection model. *Data &amp; Policy*, *3*. doi:10.1017/dap.2021.3

Nguyen, M.-D., Bouaziz, A., Valdes, V., Rosa Cavalli, A., Mallouli, W., & Montes De Oca, E. (2023). A deep learning anomaly detection framework with explainability and robustness. *Proceedings of the 18th International Conference on Availability, Reliability and Security*. doi:10.1145/3600160.3605052 "Addressing
this need, eXplainable Artificial Intelligence (XAI) [1, 9] has
emerged as a prominent research area, aiming to provide users with
a rationale for understanding the output produced by AI systems
and fostering trust among end-users. Various approaches, such as
the ones proposed in [8, 11], have been developed to enhance global
and local interpretability, shedding light on what the models have
learned and how they make individual predictions.
Existing AI frameworks, such as Shapash1, explainerdashboard2,
and DataRobot3, have made valuable contributions to the field."

Nri, H., Jenkins, S., Paul, K., & Caruana, R. (2019). InterpretML: A unified framework for machine learning interpretability. *arXiv Preprint arXiv:1909.09223*.

Priscilla, C. V., & Prabha, D. P. (2020). Influence of optimizing xgboost to handle class imbalance in credit card fraud detection. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 1309–1315. doi:10.1109/icssit48917.2020.9214206

Psychoula, I., Gutmann, A., Mainali, P., Lee, S. H., Dunphy, P., & Petitcolas, F. (2021). Explainable Machine Learning for Fraud Detection. *Computer*, *54*(10), 49–59. doi:10.1109/mc.2021.3081249

Ras, G., Xie, N., Van Gerven, M., & Doran, D. (2022). Explainable Deep Learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, *73*, 329–397. doi:10.1613/jair.1.13200

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “why should I trust you?” Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. doi:10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). doi:10.1609/aaai.v32i1.11491

Schlegel, U., Arnout, H., El-Assady, M., Oelke, D., & Keim, D. A. (2019). Towards a rigorous evaluation of XAI methods on time series. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. doi:10.1109/iccvw.2019.00516

Sharma, A., & Bathla, N. (2020). *Review on Credit Card Fraud Detection and Classification by Machine Learning and Data Mining Approaches*, *6*(4), 687–692.

Sharma, P., & Priyanka, S. (2020). Credit card fraud detection using Deep Learning based on neural network and auto encoder. *International Journal of Engineering and Advanced Technology*, *9*(5), 1140–1143. doi:10.35940/ijeat.e9934.069520

Sinanc, D., Demirezen, U., & Sağıroğlu, Ş. (2021). Explainable Credit Card Fraud Detection with Image Conversion. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *10*(1), 63–76. doi:10.14201/adcaij20211016376 A new explainable artificial intelligence approach is ... presented. In this way, feature relationships that have a dominant effect on fraud detection are revealed.

Sullivan, R. S., & Longo, L. (2023). Explaining deep Q-learning experience replay with Shapley additive explanations. *Machine Learning and Knowledge Extraction*, *5*(4), 1433–1455. doi:10.3390/make5040072

Vilone, G., & Longo, L. (2021a). A quantitative evaluation of global, rule-based explanations of Post-Hoc, model agnostic methods. *Frontiers in Artificial Intelligence*, *4*. doi:10.3389/frai.2021.717899

Vilone, G., & Longo, L. (2021b). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, *3*(3), 615–661. doi:10.3390/make3030032

Vilone, G., & Longo, L. (2021c). Notions of explainability and evaluation approaches for Explainable Artificial Intelligence. *Information Fusion*, *76*, 89–106. doi:10.1016/j.inffus.2021.05.009

Vouros, G. A. (2022). Explainable deep reinforcement learning: State of the art and Challenges. *ACM Computing Surveys*, *55*(5), 1–39. doi:10.1145/3527448

Wu, T.-Y., & Wang, Y.-T. (2021). Locally interpretable one-class anomaly detection for credit card fraud detection. *2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. doi:10.1109/taai54685.2021.00014