Unveiling the Black Box: A Comprehensive Review of Explainable AI Techniques

Anushree G[1[0009−0001−4004−2980]], Suraj B Madagaonkar[2,3[0009−0003−8809−4098]], and Ravili C H[3[0009−0001−2954−7013]]

[1] Artificial Intelligence and Data Science Department, CMR Institute of Technology, Bangalore
anushree.g@cmrit.ac.in
[2] Manipal Institute of Technology Manipal Academy of Higher Education Manipal
suraj.madagaonkar@manipal.edu
[3] CMR Institute of Technology, Bangalore
rach22csds@cmrit.ac.in

Abstract. As artificial intelligence (AI) continues to integrate into various sectors, the complexity and opacity of AI models, particularly in machine learning (ML), pose significant challenges to interpret-ability and trust. This review paper addresses the critical need for explainable AI (XAI) to enhance understanding and transparency in ML models. We provide a comprehensive survey of state-of-the-art XAI techniques, including feature importance methods such as LIME (Local Interpret- able Model-agnostic Explanations) and SHAP (Shapely Additive expla- nation), as well as perturbation and attention-based mechanisms, to elucidate model decisions. Our analysis spans a diverse range of applications, including finance, education, and healthcare, showcasing the practical utility and impact of XAI methods. We discuss crucial issues such as the trade-offs between model accuracy and interpret ability, the de- sign of user-friendly explanations, and the development of comprehensive evaluation metrics. Furthermore, we explore the implications of XAI on user trust and decision-making, emphasizing the importance of reliable and ethical AI systems. This review contributes to the ongoing efforts to make AI systems more interpret- able, reliable, and aligned with societal needs, providing a robust foundation for future research and practical implementations of XAI.

Keywords: Explainable AI · Machine Learning · Interpret-ability · Transparency · Ethical AI · XAI Techniques.

1 Introduction

Explainable AI is an AI system that explains their decision making which is re- ferred as Explainable AI or XAI. The goal of XAI is to provide verifiable expla- nations of how machine learning systems makes decisions and letting humans to be in the loop. There are two ways to provide explainable AI. Use Machine learn- ing approaches that are inherently explainable such as decision trees, knowledge graphs and similarity models. Develop new approaches to explain complicated neural networks.

1.1 Evolution of AI

AI has evolved significantly over the years through different phases. Each wave represents a different approach and capability in AI development. The first wave of AI, which primarily focuses on using logic rules to represent knowledge. These systems were effective for well-defined problems but lacked learning capabilities and struggled with handling uncertainty. Statistical AI is the second wave, char- acterized by the use of statistical models and machine learning [2][7]. These systems learned from large datasets, making them more adaptable and power- ful. However, they often acted as "black boxes," offering little explain-ability or understanding of the context. Explainable AI Represents the third wave of AI, focusing on making AI systems more understandable and interpret-able. Ex- plainable Artificial Intelligence (XAI) has emerged as a crucial field within AI, aiming to enhance the transparency and interpret-ability of machine learning models by providing description to the decision made by AI.

1.2 Evolution of AI

As AI systems are increasingly deployed in various domains such as health- care, finance, education, and autonomous systems, the need for understanding and trusting these models becomes paramount [1]. XAI addresses this need by providing insights into how AI models make decisions, thereby fostering trust, accountability, and ethical use of AI.XAI can be used in multiple domains, each benefiting from different types of explain-ability methods.
The choice of XAI methods depends on the specific scenario and the stake- holders involved:

– **Model-specific methods:** These are preferred when transparency and ease of understanding are critical, such as in regulatory contexts and scenarios requiring direct human interpretation[5].
– **Post-hoc methods:** Suitable for explaining complex models after they have been trained, these methods are ideal for applications where high predictive accuracy is required alongside interpret-ability[1].
– **Visual explanation methods:** These are beneficial in domains where stakeholders can better understand graphical representations, such as in au- tonomous systems and education[1].

The need for XAI arises from several key factors: As Artificial Intelligence (AI) systems become increasingly integrated into various aspects of daily life, from healthcare and finance to autonomous vehicles and criminal justice, the demand for transparency and trustworthiness in these systems has grown signif- icantly. Traditional AI models, particularly deep learning algorithms, are often described as "black boxes" due to their complex and opaque decision-making processes. This lack of transparency raises several concerns, making Explainable AI (XAI) not just a desirable feature but a necessity.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 09 | Sept - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

– **Trust and Accountability:** Transparent AI models help build trust among users and stakeholders, ensuring that AI systems are used responsibly and ethically[23].

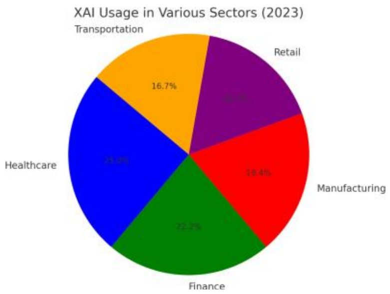
– **Bias Detection and Mitigation:** XAI helps identify and address biases in AI models, promoting fairness and preventing discrimination.

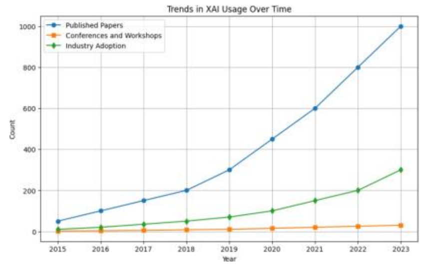
– **Improved Decision-Making:** By understanding the underlying mecha- nisms of AI models, users can make more informed decisions, enhancing the overall effectiveness of AI applications.

– **Regulatory Compliance:** Many industries are subject to regulations that require explanations of automated decisions. XAI facilitates compliance with these regulation [23].

1.3 Trends and Usage of XAI

Fig.a. illustrates the distribution of Explainable AI usage across five different sectors. The sectors and their corresponding usage percentages.From this chart, it is evident that the Healthcare sector leads in the adoption of Explainable AI, accounting for a quarter of the total usage. This indicates a strong emphasis on transparency and interpret-ability in medical decision-making and patient care.





(a) XAI Usage in Various Sectors (2023)

(b) Trends in XAI Usage Over Time (2015-2023)

Fig.b. illustrates Trends in XAI Usage Over Time (2015-2023)" depicts the growth in Explainable AI adoption over an eight-year period. The chart tracks three key metrics: Published Papers, Conferences and Workshops,Industry Adop- tion

Key concept of XAI

Concept	Description
Explainability	The degree to which a human can understand the cause of a decision or can provide an explanation of how a decision is made by an AI system[2].
Transparency	The clarity with which the operations of a system can be understood. Transpar- ent models are those whose workings can be easily comprehended by humans[2] [23].
Interpretability	The extent to which a cause and effect can be observed within a system. In XAI, it refers to the clarity of the model's mechanisms and decision-making process[23].
Trust	The level of confidence that users have in AI systems. Trust is built through explainability and transparency, ensuring the AI behaves as expected[2].
Accountability	The obligation to explain, justify, and take responsibility for the AI's actions. Ensuring AI decisions can be traced back and justified[2].
Causality	Understanding and establishing cause-and-effect relationships within the AI's decision-making process [23].
Fairness	Ensuring that AI systems do not produce biased outcomes. Fairness relates to the ethical dimension of AI, ensuring equitable treatment of all users.
Debugging	The process of identifying, analyzing, and removing errors or bugs within an AI system. Explainability aids in effective debugging.
Model Compression	Techniques used to reduce the size of a model while maintaining its perfor- mance. Often used to improve interpretability by simplifying complex models.
Sensitivity Analysis	A method to determine how different values of an input affect a particular out- put. Useful in understanding model robustness and the importance of features.
Layer-wise Rele- vance Propagation (LRP)	A technique to decompose the prediction of a neural network in order to at- tribute relevance scores to each input feature, highlighting their importance.
Feature Importance	Measures used to identify the contribution of each feature to the prediction made by the model.

Knowledge Distilla- tion	Transferring knowledge from a complex model (teacher) to a simpler model (student) to maintain performance while improving interpretability.
Counterfactual Ex- planations	Descriptions of the minimum conditions required to change a decision, helping users understand the decision boundaries of the model[4].
Visual Analytics	The use of data visualization techniques to enhance the interpretability of AI models, making complex data more accessible and understandable.
Human-Computer Interaction (HCI)	The study and design of interactions between humans and computers. In XAI, it involves creating user-friendly interfaces for AI explanations.
Ethical AI	The development and deployment of AI systems in a manner that adheres to ethical standards, ensuring fairness, accountability, and transparency [23].
Social Science Per- spectives	Understanding the social, psychological, and cognitive aspects of AI interac- tions to improve the design and acceptance of explainable AI systems[11].
Regulatory Com - pliance	Ensuring AI systems adhere to laws and regulations, such as the GDPR's "right to explanation," which mandates transparency in automated decision-making [11].
Dark Knowledge	Knowledge distillation concept where the "dark" or less obvious knowledge learned by a complex model is transferred to a simpler model for interpretabil- ity.

2 Taxonomy of Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) encompasses a wide range of meth- ods and techniques designed to make the decision-making processes of AI mod- els transparent and understandable. This taxonomy categorizes XAI methods based on various criteria to help researchers and practitioners choose appro- priate techniques for specific applications as shown in the figure . The primary categories include explanation generation, coverage, chronological hierarchy, and model specificity.

1. Explanation Generation:

– Feature Attribution: These methods compute the relevance or explana- tory power of features with respect to predictions generated by the model. Examples include SHAP and LIME.

– Simplification: Simplifying the original model into an interpret-able form to mimic and explain its behavior, such as using decision trees or linear models.

– Explain-by-Example: Providing explanations by identifying similar sam- ples with similar or different predictions, helping users understand model behavior through comparative analysis.

2. Coverage:

– Global Explain-ability: Methods that provide explanations summarizing patterns learned by the model over a large number of samples [5]. These methods aim to understand the overall behavior of the model across the entire dataset. Examples include Partial Dependence Plots and Feature Importance analysis.

– Local Explain-ability: Methods that provide explanations for individual predictions or small groups of similar samples. These methods focus on understanding the model's behavior for specific instances [5]. Examples include LIME and SHAP for individual predictions

3. Chronological Hierarchy:

– Pre-Model Explain-ability: Techniques applied to the dataset before the modeling process, often for exploratory data analysis and presentation. Examples include data visualization techniques and feature selection methods.

– In-Model Explain-ability: Techniques that produce explanations as part of the model training process. Examples include inherently interpret-able models like decision trees and linear regression.

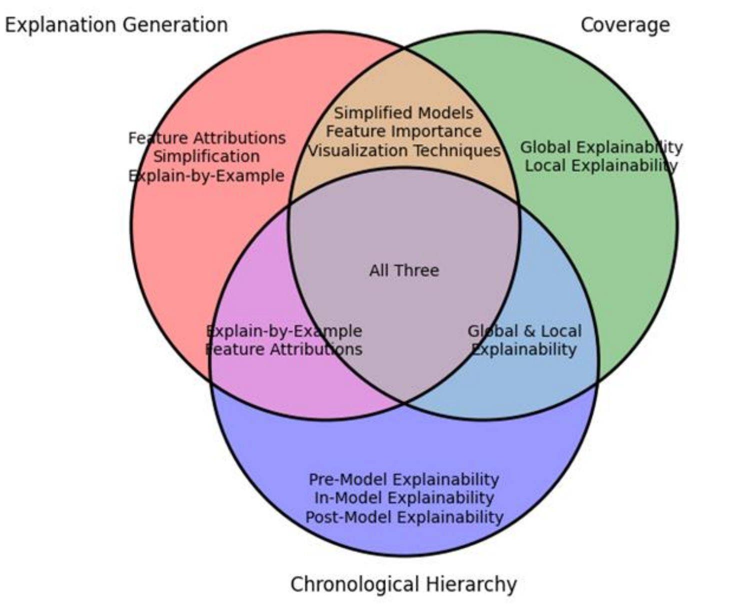
– Post-Model Explain-ability: Techniques applied after the model has been trained to generate explanations for its predictions. Examples include SHAP, LIME, and saliency maps for neural networks.

4. Model Specificity:

– Model-Specific Methods: These methods are tailored to specific types of models and leverage their internal structures to provide explanations. Examples include Layer-wise Relevance Propagation (LRP) for neural networks and decision tree paths for tree-based models[2].

– Model-Agnostic Methods: These methods can be applied to any type of model regardless of its internal workings. They are flexible and widely applicable but may not leverage model-specific details. Examples include SHAP and LIME.

Fig. 2: Taxonomy of XAI



3 SHAP and LIME

LIME is a technique designed to explain the predictions of any machine learning model. It helps you understand why a model made a particular decision for a specific instance.Key Concepts

– Local Explanations: LIME focuses on explaining the prediction for a single instance (or a small, local area around that instance) rather than explaining the entire model. This means LIME helps you understand why the model made a particular prediction for a specific data point[1] [5].

– Model-Agnostic: LIME can be used with any machine learning model, whether it's a simple model like linear regression or a complex model like a neural network. It doesn't rely on the internal workings of the model, making it very versatile[1].

SHAP is a method used to explain the output of machine learning models. It provides a way to understand the contribution of each feature to a particular prediction, based on principles from cooperative game theory [5].Key Concepts

– Shapley Values: Originally from game theory, Shapley values represent a fair way to distribute the "payout" (in this case, the prediction) among all features (players) based on their contribution.

– Model-Agnostic and Model-Specific: SHAP can be used with any model (model-agnostic) or have specific versions optimized for certain models

– Global and Local Explanations: SHAP values can explain individual predic- tions (local) and give an overview of feature importance across all predictions (global).

SHAP Works by calculating perturb, shapely values,and aggregating contribu- tions. Perturb Features Like LIME, SHAP perturbs feature values and observes the effect on the model's output. Calculate Shapley Values For each feature, SHAP calculates how the prediction changes when the feature is included versus when it is excluded. This is done by averaging over all possible combinations of feature subsets. Aggregate Contributions the contributions (Shapley values) of all features are combined to explain the model's prediction for a specific in- stance.

SHAP for linear models is not as common because linear models have intrin- sic interpret-ability and simplicity. Here's why SHAP may not be the preferred choice for linear models:

– Intrinsic Interpret-ability of Linear Models Linear models provide direct co- efficients for each feature. These coefficients indicate the strength and di- rection (positive or negative) of the relationship between each feature and the target variable. Simple Explanation: For a linear regression model, the prediction is simply a weighted sum of the input features. This means that the contribution of each feature to the prediction is directly proportional to its coefficient[1].
– SHAP Complexity and Overhead SHAP calculations can be computation- ally expensive, especially for large datasets or models with many features. Since linear models already offer clear and direct explanations, the added computational cost of SHAP is unnecessary.SHAP adds an extra layer of complexity that may not provide significant additional insights beyond what is already available from the linear model's coefficients [5].
– Redundancy it means providing the same information multiple times or through different means without adding any new insights or value. In the context of using SHAP values with linear models, it means that the expla- nations SHAP provides are essentially duplicating the information that is already available through the model's coefficients.

SHAP (SHapley Additive exPlanations) is a powerful tool for explaining model predictions, but it is not always the most efficient or necessary choice for all types of models. Here are some models for which SHAP might not be the best choice
– Linear Regression Models in this The contributions of each feature are di- rectly provided by the model's coefficients, making it easy to interpret with- out additional tools. Each coefficient represents the contribution of its cor- responding feature to the prediction, so using SHAP would be redundant.
– Logistic Regression Models is Similar to linear regression, the coefficients directly indicate the contribution of each feature to the log-odds of the target variable.The interpret-ability of the model comes from the coefficients, which can be converted to odds ratios for better understanding.
– Simple Decision Trees which provide a clear and interpret-able structure where each decision path can be traced from root to leaf. The split points and feature importance can be directly observed, making additional explanations from SHAP unnecessary.
– Naive Bayes Classifiers these classifiers are based on the assumption of fea- ture independence, and their probabilistic nature allows straightforward in- terpretation of feature contributions.The conditional probabilities and like- lihoods used in the model are easy to understand without needing SHAP values.
– K-Nearest Neighbors (KNN) which is the non-parametric model that makes predictions based on the closest training examples, making it less clear how to attribute contributions to individual features. The model's predictions are based on the majority class of the nearest neighbors rather than feature contributions, so SHAP might not provide meaningful insights.
– Rule-Based Models use a set of human-readable rules for making predictions, which are inherently interpret-able. The rules themselves provide clear logic for predictions, and SHAP values may not add much additional clarity.
– Simple Ensemble Methods of interpret-able models (like decision trees) can often be interpreted by examining individual models.When the ensemble is not too complex, the feature importance's and decision paths are still relatively clear without needing SHAP.

While SHAP might not be the best choice for the models listed above due to redundancy or lack of added value, there are scenarios where SHAP can still provide benefits: When comparing multiple types of models SHAP provides a consistent framework for understanding feature contributions across models. If even simple models have complex interaction terms or non-linearity's, SHAP can help illuminate these effects.For organizations or projects that use a mix of simple and complex models, using SHAP across the board can provide a unified approach to model interpret-ability.

4 Applications of Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) is pivotal in enhancing transparency, trust, and decision-making capabilities across various domains. By providing in- sights into AI model decisions, XAI fosters a better understanding of how and why specific outcomes are produced. This transparency is essential for ensur- ing ethical and accountable use of AI technologies. Here is a summary of XAI applications across different fields:

– Healthcare:Medical Diagnosis and Treatment Planning: XAI helps clinicians understand AI-driven recommendations, increasing trust and improving pa- tient care through transparent decision-making processes.Example Methods like Decision trees, logistic regression.
– Finance:redit Scoring and Fraud Detection: XAI ensures compliance with regulations by explaining complex models, helping financial institutions main- tain accountability and fairness.Example Methods like SHAP, LIME.
– Education:Personalized Learning and Performance Analysis: XAI supports personalized education by explaining student performance predictions and providing insights into individual learning needs [6] [4].Example Methods like Rule-based systems, model-agnostic methods.
– Autonomous Systems:Self-Driving Cars: XAI aids in understanding and val- idating decisions made by AI in autonomous vehicles, ensuring safety and reliability.Example Methods like Saliency maps, feature importance.
– Legal and Regulatory Compliance:Algorithmic Decision-Making: XAI pro- vides explanations for decisions made by automated systems, ensuring trans- parency and adherence to legal standards[4].
– Manufacturing and Industry:Predictive Maintenance and Quality Control: XAI helps in understanding predictions about equipment failures or product defects, facilitating better maintenance and quality assurance [13].
– Marketing and Customer Insights:Customer Behavior Analysis: XAI explains the factors influencing customer decisions, enabling more targeted and effec- tive marketing strategies[13].
– Human Resources:Employee Performance and Retention: XAI analyzes and explains factors affecting employee performance, aiding in better HR decision- making.

Additionally, XAI finds applications in several other fields: Telecommunications: Network optimization and fault detection.
Environmental Monitoring: Climate change impact analysis.
Retail and E-commerce: Inventory management and demand forecasting. Cybersecurity: Threat detection and response.
Transportation and Logistics: Route optimization and fleet management. Insurance: Risk assessment and claim processing.
Agriculture: Precision farming and crop monitoring.
Public Safety and Emergency Response: Disaster management and crime pre- diction.
Energy Sector: Smart grid management and renewable energy forecasting. Social Media and Content Moderation: Content recommendation and modera- tion.
By providing clear and interpretable insights into AI models, XAI enhances the trustworthiness and effectiveness of AI applications, ensuring their ethical and responsible use across diverse sectors. This broad applicability underscores the importance of XAI in driving the adoption of AI technologies in a transparent and accountable manner.

5 Conclusion

In this review paper, we have explored the rapidly evolving field of Explain- able Artificial Intelligence (XAI). As AI systems become increasingly integral to decision-making processes across various domains, the need for transparency and interpret-ability in these systems has become paramount. XAI aims to bridge the gap between complex, opaque machine learning models and the human users who rely on their outputs. The literature presents a wide array of methods and techniques designed to enhance the interpretability of AI models. From model- agnostic approaches, such as LIME and SHAP, to inherently interpretable mod- els, like decision trees and rule-based systems, the diversity of XAI methods reflects the complexity of the challenges faced. Additionally, specific applica- tion areas, such as healthcare, finance, and autonomous systems, have unique requirements and constraints that drive the development of tailored XAI solu- tions. Despite significant advancements, XAI remains a field rich with challenges and opportunities. Ensuring the robustness and reliability of explanations, ad- dressing the trade-offs between interpretability and model performance, and de- veloping standardized evaluation metrics are critical areas for future research. Furthermore, the ethical implications of AI and the need for regulatory frame- works underscore the importance of responsible AI development. XAI is not just a technical challenge but a multidisciplinary endeavor that requires collabora- tion between AI researchers, domain experts, ethicists, and policymakers. By advancing our understanding of XAI and implementing effective solutions, we can build AI systems that are not only powerful but also trustworthy, transpar- ent, and aligned with human values

References

1. V. Arya, R.K.E. Bellamy, P.Y. Chen, et al., *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, 2020.
2. Y. Lu, D. Wang, P. Chen, and Z. Zhang, *Design and Evaluation of Trustworthy Knowledge Tracing Model for Intelligent Tutoring System*.
3. G. Elkhawaga, O. Elzeki, M. Abuelkheir, and M. Reichert, *Evaluating Explainable Artificial Intelligence Methods Based on Feature Elimination: A Functionality- Grounded Approach*, Electronics, Vol. 12, p. 1670, 2023. DOI: 10.3390/electron- ics12071670.
4. Various Authors, *Explainable AI In Education: Current Trends, Challenges, And Opportunities*, 2021.
5. Various Authors, *Earliest Possible Global and Local Interpretation of Students Per- formance in Virtual Learning Environment by Leveraging Explainable AI*, 2022.
6. Various Authors, *Explainable Student Performance Prediction Models: A System- atic Review*, 2021.
7. Various Authors, *From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where*, 2021.
8. Various Authors, *The Importance of Ethical Reasoning in Next Generation Tech Education*, 2022.
9. W. M. P. van der Aalst, *Process mining: A 360 degree overview*, in Process Mining Handbook (LNBIP), Springer-Verlag, New York, NY, USA, 2022, pp. 3–34.
10. A. E. Márquez-Chamorro, M. Resinas, and A. Ruiz-Cortés, *Predictive monitoring of business processes: A survey*, IEEE Trans. Services Comput., vol. 11, no. 6, pp. 962–977, Nov./Dec. 2018.
11. W. Rizzi, et al., *Explainable predictive process monitoring: A user evaluation*, 2022, arXiv:2202.07760.
12. M. Velmurugan, C. Ouyang, C. Moreira, and R. Sindhgatta, *Evaluating fidelity of explainable methods for predictive process analytics*, in Proc. Lecture Notes Bus. Inf. Process., Intell. Inf. Syst., vol. 424, S. N. A. Korthaus, Ed., Springer-Verlag, New York, NY, USA, 2021, pp. 64–72.
13. M. Velmurugan, C. Ouyang, C. Moreira, and R. Sindhgatta, *Evaluating stability of post-hoc*

explanations for business process predictions, in Proc. Int. Conf. Service- Oriented Comput., H. Hacid, O. Kao, M. Mecella, N. Moha, and H.-y. Paik, Eds., Springer-Verlag, Cham, Switzerland, 2021, pp. 49–64.

14. D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, *Machine learning interpretability: A survey on methods and metrics*, Electronics, vol. 8, no. 8, 2019, Art. no. 832.

15. R. Wilming, C. Budding, K.-R. Müller, and S. Haufe, *Scrutinizing XAI using linear ground-truth data with suppressor variables*, Mach. Learn., vol. 111, no. 5,
pp. 1903–1923, 2022.

16. I. Verenich, M. Dumas, M. L. Rosa, F. M. Maggi, and I. Teinemaa, *Survey and cross-benchmark comparison of remaining time prediction methods in business pro- cess monitoring*, ACM Trans. Intell. Syst. Technol., vol. 10, pp. 1–34, Jul. 2019.

17. I. Teinemaa, M. Dumas, M. L. Rosa, and F. M. Maggi, *Outcome-oriented predictive process monitoring: Review benchmark*, ACM Trans. Knowl. Discov. Data, vol. 13,
pp. 1–57, Mar. 2019.

18. D. Apley and J. Zhu, *Visualizing the effects of predictor variables in black box supervised learning models*, J. Roy. Statist. Soc. B, vol. 82, no. 4, pp. 1059–1086, 2020.

19. International Data Corporation IDC, *Worldwide Semiannual Cognitive Artificial Intelligence Systems Spending Guide*, Accessed: Jun. 6, 2018. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS43662418

20. Statista, *Revenues From the Artificial Intelligence (AI) Market World- wide From 2016 to 2025*, Accessed: Jun. 6, 2018. [Online]. Available: https://www.statista.com/statistics/607716/worldwide-artificial-intelligence- market-revenues/

21. Gartner, *Top 10 Strategic Technology Trends for 2018*, Accessed: Jun. 6, 2018. [Online]. Available: https://www.gartner.com/doc/3811368?srcId=1-6595640781

22. S. Barocas, S. Friedler, M. Hardt, J. Kroll, S. Venkatasubramanian, and
H. Wallach, *The FAT-ML Workshop Series on Fairness, Accountability, and Transparency in Machine Learning*, Accessed: Jun. 6, 2018. [Online]. Available: http://www.fatml.org/

23. G. Elkhawaga, O. M. Elzeki, M. Abu-Elkheir, and M. Reichert, *Why Should I Trust Your Explanation? An Evaluation Approach for XAI Methods Applied to Predictive Process Monitoring Results*, Member, IEEE.

24. B. Kim, K. R. Varshney, and A. Weller, *2018 Workshop on Human Interpretabil- ity in Machine Learning (WHI)*, Accessed: Jun. 6, 2018. [Online]. Available: https://sites.google.com/view/whi2018/

25. A. G. Wilson, B. Kim, and W. Herlands, *Proceedings of NIPS 2016 Workshop on Interpretable Machine Learning for Complex Systems*, 2016. [Online]. Available: https://arxiv.org/abs/1611.09139

26. D. W. Aha, T. Darrell, M. Pazzani, D. Reid, C. Sammut, and P. Stone, *Proc. Workshop Explainable AI (XAI)*, IJCAI, 2017.

27. M. P. Farina and C. Reed, *Proc. XCI, Explainable Comput. Intell. Workshop*, 2017.

28. I. Guyon, et al., *Proc. IJCNN Explainability Learn. Mach.*, 2017.

29. A. Chander, et al., *Proc. MAKE-Explainable AI*, 2018.