International Journal of Innovative Computer Science and IT Research

Volume 02, Issue 04 – Jul-Aug 2026 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

Explainable Data Analytics for Transparent Computational Decision Support

Ritesh Ajoodha

Associate Professor, University of the Witwatersrand, South Africa

Abstract

The increasing use of computational systems for prediction, classification, recommendation, risk assessment, and resource allocation has created a growing requirement for decision processes that are not only accurate but also understandable and transparent. Conventional data analytics frequently produces complex computational outputs without adequately explaining how particular predictions or recommendations are generated. Explainable Data Analytics addresses this limitation by integrating analytical models with techniques that communicate the factors, relationships, and patterns contributing to computational decisions. This paper examines the role of explainable data analytics in developing transparent computational decision-support systems, focusing on model interpretability, feature importance, local and global explanations, surrogate modeling, rule-based explanations, visualization, uncertainty, and human-centered decision support. A conceptual methodological framework is developed consisting of data preparation, analytical modeling, explanation generation, explanation validation, and decision-support integration.

The analysis considers both inherently interpretable models and post-hoc explanation techniques for complex machine-learning systems. The study identifies important application areas including cybersecurity, financial risk assessment, healthcare analytics, education, business intelligence, public administration, and intelligent information systems. Two analytical tables compare explainability approaches and application requirements for transparent decision support. The findings indicate that explainability can improve user understanding, facilitate model auditing, support error identification, and strengthen confidence in computational recommendations.

However, transparency is affected by several challenges, including explanation fidelity, computational complexity, conflicting stakeholder requirements, data quality, model opacity, explanation instability, and the potential for misleading interpretations. The study argues that explainability should be treated as an integral component of analytical system design rather than as an optional layer added after model development. The paper concludes that combining interpretable analytical methods, post-hoc explanation techniques, uncertainty communication, human oversight, and continuous evaluation can contribute to more accountable and trustworthy computational decision-support environments.

Keywords: Explainable Data Analytics, Explainable Artificial Intelligence, Transparent Decision Support, Model Interpretability, Computational Decision-Making, Data Science, Feature Importance, Machine Learning, Explainability, Intelligent Systems.

International Journal of Innovative Computer Science and IT Research

Volume 02, Issue 04 – Jul-Aug 2026 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

1. Introduction

The rapid development of data-driven computing has transformed the way organizations generate predictions, evaluate risks, allocate resources, and support operational decisions. Modern analytical systems can process large and complex datasets and identify patterns that may be difficult for human analysts to recognize manually. Machine-learning models are increasingly integrated into information systems for classification, forecasting, recommendation, anomaly detection, and optimization. Despite these capabilities, many advanced models operate as complex computational structures whose decision processes are difficult to understand. This lack of transparency can limit the practical acceptance and responsible use of analytical systems.

Traditional data analytics generally emphasizes predictive accuracy and computational efficiency. While these characteristics remain important, decision-support systems often operate in contexts where users need to understand why a particular result was produced. A financial analyst may need to understand why a transaction has been classified as suspicious, while a cybersecurity professional may need to know which network characteristics contributed to an alert. Similarly, business managers may require explanations for forecasts before incorporating them into strategic planning. Explainable data analytics therefore extends conventional analytical practice by emphasizing both computational results and understandable reasoning.

The concept of explainability is particularly important when complex machine-learning models are used. Deep neural networks, ensemble models, gradient-boosting algorithms, and other sophisticated approaches can capture nonlinear relationships and interactions within large datasets. However, their internal decision processes may be difficult to communicate to nontechnical users. Explainable artificial intelligence research has consequently developed techniques such as feature attribution, local surrogate models, partial dependence analysis, counterfactual explanations, and rule-based representations. These techniques attempt to provide information about the factors influencing model outputs while preserving the predictive capabilities of complex models.

Transparent computational decision support requires more than simply generating an explanation. An explanation should be relevant to the decision context, sufficiently faithful to the underlying model, understandable to the intended user, and presented at an appropriate level of detail. Different stakeholders may also require different explanations. A data scientist may need detailed technical information about model behavior, whereas a business manager may require a concise description of the most influential factors. Therefore, explainability should be designed around both computational characteristics and human decision-making requirements.

This paper examines explainable data analytics as a framework for improving transparency in computational decision-support systems. The study focuses on analytical interpretability, explanation-generation mechanisms, human-centered explanation design, validation, and practical deployment. A conceptual framework is presented to demonstrate how explainability can be integrated throughout the data analytics lifecycle, from data preparation and model development to explanation generation and decision support.

Volume 02, Issue 04 – Jul-Aug 2026 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

2. Objectives

2.1 Primary Objective

The primary objective of this study is to examine how explainable data analytics can improve transparency and understanding in computational decision-support systems. The research focuses on methods that communicate the factors influencing analytical outputs while maintaining useful predictive performance, thereby supporting informed human interpretation and responsible computational decision-making.

2.2 Examination of Explainability Mechanisms

The first objective is to examine major techniques used to explain computational predictions, including feature-importance methods, local explanations, global explanations, surrogate models, rule-based representations, and counterfactual analysis. These techniques provide different perspectives on model behavior and can help users understand relationships between input information and generated outputs.

A second objective is to evaluate how explanation methods can be integrated into data analytics workflows. Rather than treating explainability as an independent visualization step, the study considers it across data preparation, model development, prediction, explanation generation, and decision evaluation. This integrated perspective supports more systematic transparency.

2.3 Assessment of Transparent Decision Support

The second objective is to investigate how explainable analytics can strengthen computational decision support. Explanations may help users evaluate whether a prediction is reasonable, identify potentially incorrect inputs, recognize unexpected patterns, and determine whether additional human investigation is required.

The study also examines the relationship between predictive performance and interpretability. Highly interpretable models may provide simpler explanations but may not capture all complex relationships within large datasets. Conversely, highly complex models may provide stronger predictive performance but require additional explanation mechanisms. Understanding this trade-off is important for selecting appropriate analytical approaches.

2.4 Identification of Implementation Challenges

The final objective is to identify challenges that can reduce the effectiveness of explainable analytics. These include explanation fidelity, instability, complexity, stakeholder differences, data bias, model opacity, privacy concerns, and difficulties in validating explanations. Identifying these limitations can support the development of more reliable and user-centered computational decision-support systems.

3. Review of Literature

3.1 Foundations of Explainable Data Analytics

Explainability has become an important research area as machine learning has expanded into high-impact decision environments. Earlier analytical systems frequently relied on statistical or rule-based models that were comparatively easier to interpret. The increasing use of complex predictive algorithms has created a need for additional mechanisms capable of describing model behavior.

Doshi-Velez and Kim proposed a framework for understanding interpretability in machine learning and emphasized the importance of evaluating whether explanations genuinely help users understand model behavior. Their work contributed to the broader discussion regarding what constitutes a meaningful explanation and how interpretability should be assessed.

Ribeiro, Singh, and Guestrin introduced LIME, a method for explaining individual predictions of complex models through locally interpretable surrogate models. The approach demonstrated how a complex predictor could be approximated around a particular instance to provide an understandable local explanation.

3.2 Feature Attribution and Model-Agnostic Explanation

Feature attribution techniques attempt to determine how individual variables contribute to model outputs. SHAP, based on Shapley-value principles, provides a unified framework for assigning contribution values to features. Lundberg and Lee demonstrated how such approaches can provide consistent feature-attribution explanations across different machine-learning models.

Permutation-based feature importance provides another approach by measuring changes in model performance when individual features are disturbed. Partial dependence analysis can also help users understand how predicted outcomes vary as a feature changes. These approaches are particularly useful for global model analysis, although they may require careful interpretation when features are strongly correlated.

3.3 Human-Centered Computational Decision Support

Explainability becomes meaningful when explanations support human reasoning and decision-making. A technically accurate explanation may still be ineffective if it contains excessive information, uses specialized terminology, or does not address the user's decision context. Human-centered research therefore emphasizes explanation simplicity, relevance, timing, and user expertise.

Counterfactual explanations represent another important direction because they communicate how an outcome might change if specific input conditions were different. For example, a decision-support system may indicate which characteristics would need to change for a classification outcome to be different. Such explanations can provide actionable information rather than simply describing the factors associated with an existing prediction.

4. Methodology

4.1 Research Design and Conceptual Framework

This study adopts a conceptual and analytical research design to investigate explainable data analytics for transparent computational decision support. The framework integrates five principal stages: data preparation, predictive modeling, explanation generation, explanation evaluation, and decision-support integration. The objective is to establish a continuous relationship between computational prediction and human interpretation rather than treating explainability as an isolated post-processing activity.

The conceptual process begins with structured or unstructured data and proceeds toward analytical representation and model development. Let a predictive model be represented as $f(X)=Yf(X)=Y$, where XX represents input features and YY represents the computational output. An

Volume 02, Issue 04 – Jul-Aug 2026 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

explanation mechanism EE is then applied to describe the relationship between XX and f(X)f(X). The resulting explanation is communicated to decision-makers according to their information requirements.

4.2 Data Preparation and Analytical Modeling

The first methodological stage involves preparing data for analytical processing. Data may originate from enterprise databases, transaction systems, sensor networks, cybersecurity logs, customer interactions, public datasets, or other digital information sources. Data quality procedures include missing-value assessment, duplicate removal, normalization, feature encoding, outlier assessment, and consistency checking. Careful preprocessing is important because explanations generated from poor-quality or biased data may provide misleading interpretations.

The second component involves selecting an analytical model appropriate to the decision-support problem. Interpretable models such as decision trees, linear models, generalized additive models, and rule-based systems can provide direct explanations. Complex models such as neural networks, random forests, and gradient-boosting methods may be selected when predictive performance is prioritized. Model selection should therefore consider accuracy, complexity, explainability requirements, computational resources, and decision consequences.

4.3 Explanation Generation and Interpretation

Explanation generation is performed after or alongside model prediction. Global explanations describe general model behavior, including overall feature importance, learned relationships, or response patterns. Local explanations focus on individual predictions and identify the factors that contributed most strongly to a specific computational outcome. Methods such as SHAP, LIME, permutation importance, partial dependence, and counterfactual analysis can be selected according to the analytical model and decision context.

The interpretation stage converts technical explanation outputs into understandable information. Explanation content should be presented using appropriate terminology, visual summaries, rankings, rules, or concise textual descriptions. A useful explanation should communicate not only which features were influential but also the direction and relative significance of their contribution when such information can be reliably established. This stage should account for user expertise and avoid presenting uncertain explanations as definitive causal statements.

Key methodological considerations:

- Selection of appropriate explanation technique
- Global versus local explanation requirements
- Feature attribution analysis
- Counterfactual explanation generation
- Explanation fidelity assessment
- Explanation stability assessment
- User-specific explanation design
- Uncertainty communication
- Privacy and sensitive-feature protection

International Journal of Innovative Computer Science and IT Research

Volume 02, Issue 04 – Jul-Aug 2026 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

4.4 Explanation Validation and Decision-Support Evaluation

The fourth methodological stage evaluates the quality of generated explanations. Explanation fidelity determines how accurately an explanation represents the underlying model behavior. Stability examines whether similar observations produce reasonably consistent explanations. Comprehensibility assesses whether intended users can understand the explanation, while usefulness evaluates whether the explanation improves decision-making.

The final evaluation examines the overall decision-support system rather than the explanation mechanism alone. Measures may include predictive accuracy, precision, recall, F1-score, explanation fidelity, user understanding, decision consistency, response time, and error identification capability. A transparent system should demonstrate that explanations contribute meaningful information without creating unjustified confidence or encouraging users to treat model outputs as unquestionable decisions.

5. Analysis

5.1 Explainability Approaches for Computational Models

Explainability approaches can generally be classified into inherently interpretable models and post-hoc explanation methods. Inherently interpretable models provide a direct relationship between input variables and outputs, while post-hoc methods generate explanations after a predictive model has been trained. The appropriate approach depends on the complexity of the analytical problem and the consequences associated with computational decisions.

Model-agnostic techniques are particularly useful because they can be applied across multiple types of predictive models. However, their explanations may approximate model behavior rather than directly expose the internal computational mechanism. Consequently, explanation fidelity and stability should be evaluated before explanations are incorporated into operational decision-support systems.

Table 1: Comparative Analysis of Explainability Approaches

Explainability Approach	Explanation Level	Major Strength	Major Limitation	Suitable Application
Decision Trees	Global and Local	Directly interpretable decision paths	Can become complex	Classification and operational decisions
Linear Models	Global	Clear feature relationships	Limited nonlinear representation	Risk and forecasting
LIME	Local	Explains individual predictions	Local explanations may vary	Black-box model analysis
SHAP	Global and Local	Detailed feature attribution	Computational cost can increase	Model interpretation
Permutation Importance	Global	Model-independent importance assessment	Sensitive to correlated variables	Feature analysis

Volume 02, Issue 04 – Jul-Aug 2026 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

Explainability Approach	Explanation Level	Major Strength	Major Limitation	Suitable Application
Partial Dependence	Global	Shows feature-response relationships	Can be misleading with feature dependence	Predictive analysis
Counterfactual Explanation	Local	Provides actionable alternative conditions	Requires meaningful feasible changes	Decision support

Interpretation: The comparison demonstrates that explainability techniques provide different forms of transparency. Decision trees and linear models offer direct interpretability, whereas LIME and SHAP are useful when complex predictive models must be explained. Counterfactual explanations are particularly valuable when users need actionable information about alternative outcomes. However, no technique should automatically be considered universally reliable. Explanation selection should consider model characteristics, user requirements, computational cost, feature dependencies, and the importance of maintaining fidelity to the underlying analytical model.

5.2 Explainable Analytics for Decision Support

Explainable analytics can strengthen computational decision support by allowing users to examine the evidence underlying a prediction. Instead of receiving only a classification or numerical score, users can receive information about influential features, relevant patterns, alternative scenarios, and model confidence. This additional information can support informed human judgment and make analytical systems easier to audit.

The value of explanations varies across application domains. In cybersecurity, explanations can identify network characteristics associated with an alert. In financial analytics, they can indicate transaction attributes associated with risk predictions. In business intelligence, explanations can help managers understand factors influencing forecasts. Across these applications, the explanation should support rather than replace human evaluation.

Table 2: Application Requirements for Transparent Computational Decision Support

Application Domain	Computational Task	Explanation Requirement	Transparency Benefit
Cybersecurity	Threat and anomaly detection	Influential events and network features	Faster investigation
Financial Analytics	Risk and fraud prediction	Important transaction characteristics	Improved assessment
Healthcare Analytics	Prediction and classification	Relevant clinical variables	Better professional review

Volume 02, Issue 04 – Jul-Aug 2026 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

Application Domain	Computational Task	Explanation Requirement	Transparency Benefit
E-Commerce	Recommendation	Preference and interaction factors	Greater user understanding
Business Intelligence	Forecasting	Important drivers of outcomes	Improved managerial decisions
Education Analytics	Performance prediction	Relevant learning factors	Targeted intervention
Public Administration	Risk assessment	Evidence supporting classifications	Greater accountability

Interpretation: The table illustrates that transparent decision support has different explanation requirements across application domains. Technical users may require detailed feature and event information, while decision-makers may prefer concise summaries and actionable alternatives. Healthcare, finance, and public-sector environments may require stronger transparency because decisions can have significant consequences for individuals. Therefore, explanation design should be context-sensitive and should provide enough information to support review without overwhelming the intended user.

5.3 Relationship Between Accuracy and Interpretability

Accuracy and interpretability should not necessarily be considered opposing objectives. In some analytical problems, inherently interpretable models can achieve satisfactory predictive performance while providing direct explanations. In more complex environments, however, sophisticated models may capture relationships that simpler models cannot represent adequately. Post-hoc explanation methods can then provide additional transparency.

The appropriate balance depends on decision risk. For low-risk applications, a highly accurate black-box model with effective explanations may be acceptable. For high-impact decisions, stronger requirements for transparency, validation, human oversight, and explanation stability may be necessary. Consequently, explainability should be evaluated as part of the complete decision-support architecture.

6. Discussion

6.1 Explainability as a Component of Data Analytics

The findings indicate that explainability should be incorporated throughout the analytical lifecycle. Data quality directly affects model behavior, and model behavior determines the reliability of generated explanations. Therefore, explanation cannot completely compensate for biased, incomplete, or inappropriate data.

Embedding explainability into analytical workflows can also improve debugging and model development. Analysts can examine influential variables and identify unexpected relationships that may indicate data leakage, feature errors, or model weaknesses. Explainability therefore serves not only end-users but also developers, data scientists, and system administrators.

Volume 02, Issue 04 – Jul-Aug 2026 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

6.2 Human Oversight and Computational Decisions

Transparent decision support should maintain an appropriate balance between automated analysis and human judgment. Explanations can help users evaluate computational outputs, but they should not automatically be interpreted as proof that a model's prediction is correct. Human review remains important when decisions involve uncertainty, conflicting evidence, or high consequences.

Effective human oversight requires explanations that are concise enough to understand but sufficiently detailed to support meaningful evaluation. User interfaces should distinguish between model predictions, supporting evidence, uncertainty, and human decisions. Such separation can reduce the risk of users confusing computational recommendations with definitive conclusions.

6.3 Trust, Accountability, and Responsible Analytics

Explainable analytics can contribute to accountability by making analytical processes more inspectable. When organizations can identify which variables influenced a decision, they can more effectively investigate errors and evaluate whether computational outcomes align with organizational policies.

However, transparency does not automatically guarantee fairness or correctness. An explanation can accurately describe a biased model without resolving the underlying bias. Therefore, explainability should be combined with fairness assessment, data-quality controls, model validation, security mechanisms, and governance processes. Responsible decision support requires these elements to operate together.

7. Challenges

One major challenge in explainable analytics is determining whether an explanation accurately represents the underlying model. Post-hoc techniques may approximate complex model behavior rather than expose its complete internal reasoning. An explanation that appears intuitive may therefore provide an incomplete representation of the computational process. Explanation fidelity should consequently be evaluated systematically. In addition, explanations may change when similar observations are processed, creating concerns about stability. These issues can reduce confidence in explanation-based decision support.

Another challenge involves communicating explanations to users with different levels of technical expertise. A detailed feature-attribution output may be useful to a data scientist but difficult for a manager or operational user to understand. Conversely, an overly simplified explanation may hide important uncertainty or structural relationships. Explanation systems must therefore balance technical accuracy, simplicity, relevance, and context. User-centered design is essential for preventing information overload and misleading interpretations.

Data quality and bias also present significant challenges. If an analytical model is trained using incomplete, biased, or historically distorted data, explainability techniques may simply reveal the factors associated with those biases. The explanation itself may be technically accurate while the decision remains inappropriate. Explainable analytics should therefore be accompanied by data-quality assessment, fairness evaluation, feature auditing, and continuous model monitoring.

Volume 02, Issue 04 – Jul-Aug 2026 • E-ISSN: 3067-1108 | P-ISSN: 3067-1094

8. Future Research

8.1 Interactive Explainable Analytics

Future research should investigate interactive explanation systems that allow users to explore computational decisions dynamically. Instead of providing a single static explanation, systems could allow users to examine alternative features, compare predictions, investigate counterfactual scenarios, and request explanations at different levels of detail. Such interfaces could improve analytical understanding while accommodating different user roles and expertise levels.

8.2 Explainability for Complex and Multimodal Models

Future analytical systems will increasingly combine text, images, speech, sensor information, and structured data. Explaining multimodal models presents new challenges because predictions may depend on interactions among multiple information types. Research should develop explanation techniques capable of identifying how different modalities contribute to a prediction and how cross-modal relationships affect computational decisions.

8.3 Standardized Evaluation of Explanations

A further research direction is the development of standardized evaluation frameworks for explainability. Current explanation methods can differ substantially in fidelity, stability, comprehensibility, and computational cost. Future studies should establish evaluation protocols that combine technical metrics with human-centered measures. Benchmark datasets, reproducible experiments, user studies, and domain-specific evaluation criteria could support more reliable comparisons among explainability techniques.

9. Conclusion

Explainable Data Analytics represents an important development in the design of transparent computational decision-support systems. As organizations increasingly rely on machine learning and advanced analytics, the ability to generate accurate predictions alone is no longer sufficient for many decision environments. Users may need to understand which factors influenced an outcome, how strongly those factors contributed, and whether the computational result should be accepted, questioned, or investigated further. Explainability addresses these requirements by connecting analytical outputs with understandable representations of model behavior. Techniques including feature attribution, local and global explanations, surrogate models, partial dependence analysis, rule-based interpretation, and counterfactual reasoning provide different mechanisms for exposing relevant information. Their usefulness, however, depends on the characteristics of the model, data, application, and intended users.

The analysis demonstrates that transparent decision support requires a systematic methodology rather than the simple addition of an explanation layer after model deployment. Data quality, model selection, explanation generation, validation, and user interaction are interconnected components of an explainable analytics lifecycle. Inherently interpretable models can provide direct transparency, while post-hoc techniques allow complex predictive systems to be analyzed without completely sacrificing their modeling capabilities. Nevertheless, explanations must be assessed for fidelity, stability, comprehensibility, and usefulness. An explanation that is technically attractive but poorly aligned with the underlying model or decision context may create false confidence. Similarly, explanations cannot

independently eliminate bias or guarantee fairness because they may faithfully describe decisions produced from problematic data or inappropriate analytical assumptions.

Future development should focus on interactive, adaptive, multimodal, and standardized explainability frameworks. Computational systems should provide explanations that are appropriate for different users while clearly communicating uncertainty and distinguishing model outputs from human judgments. Interactive interfaces can allow users to explore alternative scenarios and investigate influential factors, while multimodal explanation methods can address increasingly complex analytical systems that combine text, images, speech, and structured information. Standardized evaluation procedures can further improve the reliability and comparability of explanation techniques. Overall, explainable data analytics should be regarded as a core component of trustworthy computational intelligence. By combining predictive analytics with interpretability, validation, human oversight, and responsible governance, organizations can develop decision-support systems that are not only computationally capable but also transparent, auditable, and practically meaningful.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence. *IEEE Access, 6*, 52138–52160.
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.
3. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics, 8*(8), 832.
4. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
5. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42.
6. Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2020). What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*.
7. Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36–43.
8. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.
9. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*, 56–67.
10. Molnar, C. (2022). *Interpretable Machine Learning*. Independently published.
11. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences, 116*(44), 22071–22080.

12. 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.
13. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1).
14. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215.
15. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer.
16. Shapley, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games II*, 307–317. Princeton University Press.
17. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.
18. Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813.
19. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.
20. Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15.
21. Yang, F., Du, M., & Hu, X. (2021). Evaluating explanation without ground truth in interpretable machine learning. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 176–187.
22. Zhou, Y., Booth, S., Tahaei, M., & Tandon, A. (2021). A survey of explainable AI techniques and applications. *ACM Computing Surveys*.