Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

EXPLAINABLE AI TECHNIQUES IN DATA SCIENCE A SYSTEMATIC LITERATURE REVIEW

Muhammad Naeem Akhtar[*1], Munir Ahmad[2], Muhammad Rizwan[3]

*[*1,3]Faculty of Computer Science & Information Technology, The Superior University, Lahore, Pakistan*
[2]University College, Korea University, Seoul, 02841, Republic of Korea

[*1]mnapucit892@gmail.com, [2]munirahmad@korea.ac.kr, [3]mu.rizwan815@gmail.com

[1]ORCID: 0000-0001-9362-2025, [2]ORCID: 0000-0002-5240-0984, [3]ORCID: 0009-0004-9123-7747

DOI: http://doi.org/10.5281/zenodo.21350853

Keywords
Explainable Artificial Intelligence, XAI, Machine Learning, Data Science, Interpretability, Systematic Literature Review, PRISMA, Deep Learning, CNN, LIME, SHAPE.

Article History
Received: 24 January 2026
Accepted: 16 March 2026
Published: 31 March 2026

Corresponding Author: *
Muhammad Naeem Akhtar

Abstract

The increasing adoption of Artificial Intelligence (AI) in data science has enhanced predictive capabilities in various fields, including healthcare, cybersecurity, finance, education, e-commerce, and industrial analytics. Many current AI systems, however, are black-boxes, which lack transparency, interpretability and trust [1]. To address these challenges, Explainable Artificial Intelligence (XAI) has become an important research area that offers explanations of model predictions that are understandable. This systematic literature review (SLR) aims to collate the research works related to XAI techniques in data science from 2020 to 2026. Based on the PRISMA methodology, relevant studies were identified by conducting structured searches in Scopus, IEEE Xplore, ACM Digital Library and Web of Science, which resulted in 21 studies being selected. The review classifies XAI techniques into model-agnostic, intrinsic, visualization-based, attention-based, and counterfactual approaches and discusses various application areas, evaluation metrics, tools, challenges, and directions for future research. Results reveal that the most prevalent approaches in the existing literature are SHAP, LIME, Grad-CAM, and attention-based methods, and that there are limitations with respect to the standardized metrics for evaluation, performance-interpretability trade-offs, scalability, and understanding by humans. The review results offer a systematic and reproducible synthesis and suggest future research directions for trusted and explainable AI systems.

I. INTRODUCTION

In the present times, AI and ML systems are being extensively adopted in the applications of data science [4], [11], [20]. In the healthcare sector, the use of sophisticated predictive patient care models is becoming more prevalent for patient diagnosis, financial risk analysis, fraud detection, cyber security analytics, recommendation systems, natural language processing, and intelligent automation. In various domains, deep learning models have led to improved predictive performance, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers [11] and [12].

Despite these advances, the modern AIs are often black box systems with opaque decision-making processes, that can trigger concerns about transparency, trustworthiness, accountability, fairness, debugging, and compliance with regulation. AI-driven decisions are often crucial in sensitive industries such as healthcare and finance, where human lives and financial investments are involved, and the impact of wrongful or missing decisions can be profound.

In response to these challenges, Explainable Artificial Intelligence (XAI) has become a rapidly evolving research field [4], [5], [20] to increase the understandability of AI systems by generating interpretable explanations of AI model predictions and behaviours. In recent years a large amount of research has been conducted on how to interpret the output of neural networks, such as SHAP (SHapley Additive explicability) [1], [2], [3], [29], [36], LIME (Local interpretable models of the model) [1], [2], [3], [29], [36], Grad-CAM (Gradient based Class Activation Mapping) [1], [2], [3], [29], [36], Saliency mapping [1], [2], [3], [29], [36], Integrated gradients, Counterfactual explanations, and attention-based interpretation. But, there is no uniformity across the literature on XAI since it's distributed across application domains, methodologies, datasets and evaluation frameworks. In the past, research has largely been conducted on model predictive accuracy, providing a diversity of methods to measure interpretability and explanations, and little consensus or evaluation regarding how different explainability techniques affect different sorts of ML architectures.

The objectives of the systematic literature review are to address these issues by summarizing recent empirical evidence about the use of XAI techniques in data science (2020 to 2026) with a methodology that is inspired from PRISMA for the identification, screening, quality evaluation and synthesis of evidence. The main aims of this review are:

- *To have the ability to identify some of the commonly used XAI techniques for data science applications.*
- *To categorize the methods of explainability by characteristics.*
- *To explore significant areas of application and machine learning models related to XAI.*
- *To comprehend the evaluation metrics analysis and procedures of validation.*
- *To offer insights on the issues and future research goals.*

II. Background and Related Concepts

A. Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) is a set of methods and techniques which enables human understanding of the decisions made by an AI system [5], [20], [24] that goes beyond prediction and transparency. The XAI techniques can be broadly classified into four types: intrinsic interpretability methods, post-hoc explanation methods, model-specific approaches, model-agnostic approaches, and local versus global explanation techniques.

B. Interpretability vs. Explainability

Interpretability is the extent to which humans can directly comprehend a model's internal logic [13], [14], [34] while explainability is the use of external techniques that allows for the generation of understandable explanations for black-box prediction. For instance, Deep Neural Networks as a model will need some post-hoc explainability methods, while Decision Trees, for instance, can be intrinsically interpretable.

C. Importance of XAI in Data Science

The significance of XAI is growing as AI systems are now used in various critical applications, including medical diagnosis, financial lending, criminal justice, autonomous systems, cybersecurity, and industrial automation. Interpretability boosts user confidence, troubleshooting, detecting bias, accountability, compliance, and transparency of the system.

III. Research Methodology

A. Review Design

The methodology used in this study is Systematic Literature Review (SLR) based on the PRISMA 2020 principles and literature synthesis best practices [14], [20], [24]. The review process involved development of the research question, the development of a search strategy, searching databases, screening and eligibility assessment, quality appraisal, data extraction, and evidence synthesis.

B. Research Questions

RQ1: What are the most widely used AI techniques in data science applications that can be explained?
RQ2: What are the different classification of explainable AI (XAI) methods in the literature?
RQ3: What are the most common machine learning (ML) models and application areas where XAI techniques are used?
RQ4: What are the criteria for measuring explainability and model performance?
RQ5: What are the key challenges and future research questions for XAI?

C. Databases

The relevant studies were found using the Scopus, IEEE Xplore, ACM Digital Library, and Web of Science databases.

D. Search String

("explainable artificial intelligence" OR "explainable AI" OR XAI OR explainability OR interpretability OR "interpretable machine learning") AND ("machine learning" OR "deep learning" OR "artificial intelligence") AND ("data science" OR "predictive analytics" OR "data analytics") NOT ("survey" OR "review")

E. Inclusion and Exclusion Criteria

Studies were included if published between 2020-2026, written in English, focused on XAI methods, contained empirical evaluation, and were related to ML or data science applications. Studies were excluded if they were surveys or reviews, lacked methodological detail, were editorials or conceptual papers, or were unrelated to explainability.

F. PRISMA Screening Process

A total of *1142* records were initially identified from database searches. After removing *287* duplicates, *855* records were screened by title/abstract, of which *681* were excluded. The remaining *174* full-text articles were assessed, with 136 excluded after full-text review, resulting in a final set of *21* included studies.

G. Quality Appraisal

The methodological clarity, reproducibility, transparency of datasets, the rigor of the evaluation, and comparison with baseline methods were used to assess studies using adapted quality-assessment criteria with each study categorized according to quality as high, medium, or low.

IV. Descriptive Analysis of Included Studies

In the included studies, research activity around explainable AI is starting to significantly ramp up from 2020 onward, and has grown extremely quickly since 2022, with a surge of interest driven by the widespread deployment of deep-learning systems, Transformer models, and generative AI applications. There was a significant share of research related to high-stakes contexts like healthcare, cybersecurity, and financial analytics, and recent works increasingly highlighted trustworthy AI as well as fairness and explainability centered on humans, and not just feature-attribution.

B. Distribution of XAI Techniques

The most common explainability approaches that have been incorporated were: SHAP, LIME, attention-based mechanism, integrated gradients and counterfactual approach [1], [2], [3], [28], [29]. The popularity of SHAP is due to its solid theory, the ability to use it with several ML models, and its local and global interpretability. The use of Grad-CAM and saliency-based approaches was most prevalent in the computer vision domain, and attention visualization was more closely linked to Transformer-based NLP and generative AI systems.

C. Distribution of Application Domains

Among the included studies, the top application domain had been healthcare [18], [21] driven by the need for transparency in clinical AI for ethical, regulatory, and diagnostic purposes. Some other key areas were finance, cybersecurity, recommendation systems, smart cities, IoT analytics, and industrial automationwith some emerging focus on explainability in edge AI, federated learning, generative AI, and large language models.

D. Machine Learning Models Used in XAI Studies

The reviewed studies utilized explainability techniques with various models such as CNN, Transformer, Random Forest, XGBoost, LSTM, Graph Neural Network, Reinforcement Learning systems, and hybrid deep-learning architectures. The black-box nature of DL models led to their reliance on post-hoc explainability techniques to a much greater extent.

V. Taxonomy of Explainable AI Techniques

A. Model-Agnostic Methods

Model-agnostic approaches can produce explanations without relying on the structure of the machine learning model.

1) LIME

Local Interpretable Model-Agnostic Explanations (LIME) locally approximates a complex model by simpler and interpretable models around specific points in the data. Benefits: model independent, local explanations, easy to implement. Limitations are inconsistencies in sampling (instability, sensitivity to perturbation).

2) SHAP

Cooperative game theory is used to quantify the importance of features in SHAP. It is well founded, supportive, interpretability from global and local perspectives. Limitations are computational complexity and scalability in case of large data sets.

B. Visualization-Based Methods

Techniques for visualization are commonly used in deep learning explainability. Grad-CAM identifies the image areas that affect the output of the CNN and is used in medical imaging, autonomous driving and industry inspection. Saliency maps are a visualization of feature importance that are obtained by computing gradients with respect to the inputs.

C. Attention-Based Explainability

Attention mechanisms, which have been widely employed in Transformer-based architectures, visualize attention distributions, interpret the relationship between tokens, and explain the dependencies between sequences.

D. Counterfactual Explanations

Counterfactual approaches create alternative scenarios that elucidate how predictions might differ. For example, "If annual income rose by 10%, then loan approvals would occur". These methods are not just important, but they are becoming more and more crucial in terms of fairness and actionable explanations.

VI. Application Domains

A. Healthcare

The healthcare sector is the most prevalent application of XAI, where the typical applications span from disease diagnosis, medical imaging, predicting mortality in intensive care units (ICUs), detection of sepsis, and radiology analysis. In medical image analysis, visualization techniques like Grad-CAM are particularly popular.

B. Finance

XAI is being widely adopted by financial institutions in their credit scoring, fraud detection, risk prediction, and regulatory compliance. SHAP is popular because of its feature importance interpretation features.

C. Cybersecurity

XAI enhances intrusions, malware analysis, anomaly detection and network analytics.

D. Natural Language Processing

With the rise of the use of LLMs and generative AI systems, the need for explainability in transformer models has increased significantly.

VII. Overview of Included Studies

Table I presents the summary of the 21 studies included, sorted by their year of publication, type of XAI technique, application domain, model type, and reference. Each study has been analysed and evaluated and the results are shown in Tables II and III (Section VIII-IX) along with the definitions of the evaluation metrics.

Table I: Summary of Included Studies

ID	Year	Technique	Domain	Model Type	Ref
S1	2020	LIME	Healthcare	CNN	[1]
S2	2020	SHAP	Finance	XGBoost	[2]
S3	2021	Grad-CAM	Medical Imaging	CNN	[3]
S4	2021	Attention Maps	NLP	Transformer	[12]
S5	2021	Integrated Gradients	Healthcare	Deep Learning	[14]
S6	2021	Counterfactuals	Credit Scoring	Random Forest	[29]
S7	2022	SHAP	Cybersecurity	Ensemble Models	[22]
S8	2022	LIME	Recommendation System	Neural Networks	[31]
S9	2022	Saliency Mapping	Smart Cities	CNN	[20]
S10	2022	Explainable Boosting	Fraud Detection	Gradient Boosting	[17]
S11	2021	Grad-CAM	Autonomous Driving	Vision Transformer	[23]
S12	2023	Attention Visualization	NLP	Transformer	[25]
S13	2023	Counterfactual AI	Finance	ML Models	[27]
S14	2023	Integrated Gradients	IoT Analytics	Deep Learning	[21]
S15	2024	Explainable RL	Robotics	RL Models	[32]
S16	2024	SHAP	Smart Manufacturing	XGBoost	[33]
S17	2024	Rule-Based Explanations	Healthcare	Decision Trees	[34]
S18	2025	SHAP	Predictive Analytics	Ensemble ML	[28]
S19	2025	Saliency Methods	Medical Diagnosis	CNN	[35]
S20	2025	Counterfactual Explanations	Banking	ML Models	[36]
S21	2025	Explainable Time-Series	Finance	LSTM	[19]

VIII. Comparative Analysis of Selected Studies

Table II: Comparative Analysis of Selected XAI Studies

ID	Technique	Model	Domain	Key Findings	Limitations
S1	LIME	CNN	Healthcare	Improved local interpretability for disease prediction	Explanation instability across perturbations
S2	SHAP	XGBoost	Finance	Strong feature-attribution consistency in credit risk	High computational complexity
S3	Grad-CAM	CNN	Medical Imaging	Accurate localization of diagnostic regions	Limited to CNN-based architectures
S4	Attention Maps	Transformer	NLP	Improved token-level interpretation	Attention weights not always meaningful
S5	Integrated Gradients	Deep Learning	Healthcare	Reduced gradient saturation problem	Computationally intensive for large models
S6	Counterfactual Expl.	Random Forest	Credit Scoring	Actionable explanations for loan decisions	Difficulty generating realistic counterfactuals
S7	SHAP	Ensemble Models	Cybersecurity	Enhanced intrusion-detection transparency	Reduced scalability for real-time systems
S8	LIME	Neural Networks	Recommendation Sys.	User-friendly local explanations	Inconsistent explanations across runs
S9	Saliency Mapping	CNN	Smart Cities	Effective visualization of urban models	Sensitive to noisy gradients
S10	Explainable Boosting	Gradient Boosting	Fraud Detection	Balanced interpretability and performance	Limited deep-learning compatibility
S11	Grad-CAM	Vision Transformer	Autonomous Driving	Better visual interpretability for road scenes	Reduced explainability for attention layers
S12	Attention Visualization	Transformer	NLP	Improved contextual interpretation	Lack of standardized attention evaluation

ID	Technique	Model	Domain	Key Findings	Limitations
S13	Counterfactual AI	ML Models	Cross-domain	Improved fairness-oriented decision analysis	Sensitive to data distribution changes
S14	Integrated Gradients	Deep Learning	IoT Analytics	Improved feature attribution for sensors	Gradient sensitivity issues
S15	Explainable RL	RL	Robotics	Enhanced agent decision transparency	Difficult for non-experts to interpret
S16	SHAP	XGBoost	Smart Manufacturing	High interpretability in predictive maintenance	Computational cost at scale
S17	Rule-Based Expl.	Decision Trees	Healthcare	Highly interpretable clinical predictions	Lower predictive accuracy
S18	SHAP	Ensemble ML	Predictive Analytics	Consistent feature ranking	Expensive for high-dimensional data
S19	Saliency Methods	CNN	Medical Diagnosis	Effective visualization of pathology regions	Susceptible to adversarial perturbations
S20	Counterfactual Expl.	ML Models	Banking	Actionable customer-level explanations	Counterfactual realism issues
S21	Explainable Time-Series	LSTM	Finance	Improved financial forecasting transparency	Temporal explanation inconsistency

ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

IX. Evaluation Metrics and Performance Analysis

Table III: Common Evaluation Metrics in XAI Studies

Metric	Purpose
Accuracy	Predictive performance
Precision	Classification reliability
Recall	Detection capability
F1-Score	Balanced classification
AUC	Classification quality
Fidelity	Explanation consistency
Interpretability Score	Human understanding
Robustness	Stability of explanations
Computational Cost	Efficiency evaluation

In the comparative evaluation, most studies mainly focused on predictive-performance metrics (accuracy, precision, recall, F1-score, and AUC), whereas relatively fewer studies focused on explainability-specific evaluation using fidelity, robustness, interpretability scoring, or human-centered assessment. While explainability is naturally meant to be understood by humans, explainability in human-centered evaluation is still far under-studied, and deep-learning explainability methods like SHAP, Grad-CAM, and attention-based methods tend to add significant computational overhead, particularly for large-scale and real-time systems.

Table IV: Evaluation Results and Metrics Reported Across Selected Studies

ID	Technique	Metrics Used	Reported Results	Interpretation
S1	LIME	Accuracy, F1	Acc = 91.2%, F1 = 0.89	Improved local interpretability, minor performance loss
S2	SHAP	Accuracy, AUC	Acc = 94.5%, AUC = 0.96	Strong feature-attribution consistency
S3	Grad-CAM	Accuracy, Recall	Recall +6.4% in image localization	Improved diagnostic transparency
S4	Attention Maps	Precision, F1	F1 = 0.92 in NLP classification	Improved contextual explainability
S5	Integrated Gradients	Accuracy, Fidelity	Fidelity = 0.88	Reduced gradient saturation, more stable explanations

ID	Technique	Metrics Used	Reported Results	Interpretation
S6	Counterfactual Expl.	Accuracy, Human Eval.	User satisfaction = 82%	Actionable explanations improved trust
S7	SHAP	Precision, Recall	Precision = 93%, Recall = 91%	Improved intrusion-detection interpretability
S8	LIME	Accuracy, Human Satisf.	User trust +%	Improved recommendation transparency
S9	Saliency Mapping	Accuracy	Accuracy = 90.4%	Effective visual interpretation for urban analytics
S10	Explainable Boosting	Accuracy, AUC	AUC = 0.94	Balanced interpretability and predictive capability
S11	Grad-CAM	Accuracy, Robustness	Robustness +11%	Better visual explanations for autonomous driving
S12	Attention Visualization	Precision, Interp. Score	Interpretability = 0.79	Improved semantic understanding in NLP
S13	Counterfactual AI	Fairness Metrics	Bias reduction = 14%	Counterfactuals improved fairness transparency
S14	Integrated Gradients	Accuracy, Comp. Cost	Acc = 92.6%, higher latency	Improved attribution with computational overhead
S15	Explainable RL	Human Satisfaction	Understanding +21%	RL explanations remained hard for non-experts
S16	SHAP	Precision, Recall	Precision = 95%, Recall = 92%	Effective predictive-maintenance explanations
S17	Rule-Based Expl.	Accuracy	Accuracy = 82.1%	High interpretability, lower predictive performance
S18	SHAP	Accuracy, Fidelity	Fidelity = 0.90	Consistent feature-importance explanations
S19	Saliency Methods	Robustness	Decreased under adversarial noise	Vulnerability to perturbation attacks identified
S20	Counterfactual Explanation	Human Satisfaction	Satisfaction = 84%	Actionable explanations enhanced customer trust

ID	Technique	Metrics Used	Reported Results	Interpretation
S21	Explainable Time-Series	RMSE, Accuracy	RMSE -8%	Improved transparency in financial forecasting

X. Discussion

A review of the selected studies shows that explainable AI has moved beyond being just a facet of research into being a core component of reliable data science systems, with its need for transparent and interpretable models growing along with the use of AI in critical applications. Among the findings was that post-hoc explainability methods like SHAP, LIME and Grad-CAM [1, 2, 3, 28] continued to be popular, despite the fact that they do not change the predictive architectures of complex black-box models.

Nevertheless, the focus on performance is still dominant in the current research on XAI, as most studies focus on predictive accuracy rather than comprehensive and rigorous human-centered evaluation of explanations, and although scores like accuracy, precision, recall, and F1 score are often reported, few studies systematically examine the usefulness of the explanation, its cognitive interpretability, human trust, fairness, or decision transparency.

The review also notes the growing use of Transformer architectures and generative AI systems [12, 23] where traditional ML and CNN explainability techniques typically fail to provide meaningful explanations for large scale transformer models. However, computational complexity is a significant challenge for approaches like SHAP and integrated gradients, especially in large-scale data sets, real-time systems, federated learning, and edge AI applications.

In addition, the thematic synthesis identified increasing emphasis on trustworthy AI, causal explainability, fairness-aware explanations, and human-centered interpretability, suggesting that future XAI systems will be increasingly more than just attribute explanations.

XI. Challenges and Open Research Issues

The review revealed significant disparities in the definition of explainability across the studies, with some focusing on the accuracy of the predictions but failing to provide a proper test of the explainability.

A. The Trade-off between explainability and performance

Highly interpretable models tend to be not very predictive, and highly accurate deep learning systems tend to be black boxes. Interpretability and performance are seen as key challenges.

B. Lack of Standardized Evaluation Metrics

There is no universally accepted framework for measuring explanation quality, and studies frequently use inconsistent evaluation approaches

C. Scalability Issues

For large datasets, real time systems and high dimensional deep learning models, methods like SHAP become expensive.

D. Human-Centered Interpretability

There are lots of explanations that are technically accurate but not easily understood by humans; human-centered explainability is not well-developed.

E. Adversarial Vulnerabilities

Adversarial attacks can be used to manipulate the XAI systems, which has led to concerns about the reliability and trustworthiness of the explanations.

XII. Future Research Directions

A. Explainability for LARGE LANGUAGE MODELS

There is a need for more sophisticated explainability models for modern systems of generative artificial intelligence.

B. Causal Explainability

Future systems should be able to go beyond correlation-based explanations to causal explanations.

C. Real-Time Explainability

Low latency explainability is needed for Edge AI and autonomous systems.

D. Standardized Explainability Benchmarks

There's a need for consistency in the standards used to assess the quality of explanations in the field of explanation. There is a need for consistency in the standards used to assess the quality of explanations in the field of explanation.

E. Human-Centered XAI

Future studies should focus explaining the concept in a way that is comprehensible to non-technical users.

XIII. Conclusion

This systematic literature review aimed to compile recent research on explainable AI techniques used in data science applications published from 2020-2026. The findings prove that explainability is now an essential, must-have requirement for systems that are trustworthy, and that the most commonly used techniques in healthcare, finance, cybersecurity and intelligent analytics are still SHAP, LIME, Grad-CAM, attention-based explainability, and counterfactual approaches.

While progress has been made, there are several challenges that still need to be addressed, such as the absence of standard metrics for evaluation, scalability, explainability-performance trade-offs, and lack of human validation. Future research can be dedicated to explainability of Generative AI, causal reasoning, real-time interpretation, and standardized evaluation frameworks. The review provides a structured synthesis and taxonomy of existing XAI techniques and highlights relevant research gaps.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," Proc. ACM SIGKDD, 2016.

[2] S. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," Proc. NIPS, 2017.

[3] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Proc. ICCV, 2017.

[4] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence Program," AI Magazine, vol. 40, no. 2, pp. 44-58, 2019.

[5] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence," IEEE Access, vol. 6, pp. 52138-52160, 20.

[6] W. Samek, T. Wiegand, and K. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," ITU Journal, 2017.

[7] C. Molnar, Interpretable Machine Learning. Lulu Press, 2020.

[8] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making," AI Magazine, 2017.

[9] A. Holzinger et al., "What Do We Need to Build Explainable AI Systems for the Medical Domain?" arXiv, 2017.

[10] J. Pearl, The Book of Why. Basic Books, 20.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, pp. 436-444, 2015.

[12] A. Vaswani et al., "Attention Is All You Need," Proc. NIPS, 2017.

[13] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," Artificial Intelligence, vol. 267, pp. 1-38, 2019.

[14] D. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv, 2017.

[15] B. Kim et al., "Interpretability Beyond Feature Attribution," arXiv, 2021.

[16] F. Dosilovic, M. Brcic, and N. Hlupic, "Explainable Artificial Intelligence: A Survey," Proc. MIPRO, 20.

[17] A. Rai, "Explainable AI: From Black Box to Glass Box," Journal of the Academy of Marketing Science, vol. 48, pp. 137-141, 2020.

[18] D. Saraswat *et al.*, "Explainable AI for Healthcare 5.0: Opportunities and Challenges," in *IEEE Access*, vol. 10, pp. 84486-84517, 2022.

[19] Kumar, "Explainable AI in Financial Forecasting Using Time Series Analysis," International Journal for Research in Applied Science and Engineering Technology. 13. 7155-7159. 10.22214/ijraset.2025.70080.

[20] M. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges," Information Fusion, vol. 58, pp. 82-115, 2020.

[21] Moss, J.; Gordon, J.; Duclos, W.; Liu, Y.; Wang, Q.; Wang, J. Explainable AI in IoT: A Survey of Challenges, Advancements, and Pathways to Trustworthy Automation. *Electronics* **2025**, *14*, 4622.

[22] S. Bhatt et al., "Explainable Machine Learning in Cybersecurity," Computers & Security, vol. 111, 2022.

[23] Jiqian Dong et al., "Image transformer for explainable autonomous driving system", IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE Press, 2732–2737, 2021

[24] M. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence," ACM Computing Surveys, vol. 55, no. 1, 2022.

[25] Mohseni, Sina, Niloofar Zarei, and Eric D. Ragan. "A multidisciplinary survey and framework for design and evaluation of explainable AI systems." ACM Transactions on Interactive Intelligent Systems (TiiS) 11.3-4 (2021): 1-45.

[26] J. Guidotti et al., "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, 2018.

[27] Nauta, Meike, et al. "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai." *ACM Computing Surveys* 55.13s (2023): 1-42.

[28] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," Entropy, vol. 23, no. 1, 2021.

[29] Y. Zhang et al., "Counterfactual Explanations in AI," Knowledge-Based Systems, vol. 238, 2022.

[30] Gilpin, Leilani H., et al. "Explaining explanations: An overview of interpretability of machine learning." *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018.

[31] Bussone et al, "The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems". 10.1109/ICHI.2015.26., 2015

[32] S. Ras et al., "Explainable AI in Industry 4.0," IEEE Access, vol. 9, 2021.

[33] M. Carvalho et al., "Machine Learning Interpretability: A Survey on Methods and Metrics," Electronics, vol. 8, no. 8, 2019.

[34] Z. Lipton, "The Mythos of Model Interpretability," Queue, vol. 16, no. 3, 2016.

[35] D. Slack et al., "Fooling LIME and SHAP," Proc. AIES, 2020.

[36] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box," Harvard Journal of Law & Technology, vol. 31, 2018.

[37] R. Guidotti et al., "Local Rule-Based Explanations of Black Box Decision Systems," Artificial Intelligence, 2018.

[38] A. Holzinger, "From Machine Learning to Explainable AI," World Symposium on Digital Intelligence for Systems and Machines, 2019.