World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Review Article)

Hybrid models combining explainable AI and traditional machine learning: A review of methods and applications

Ranjith Gopalan [1], Dileesh Onniyil [2], Ganesh Viswanathan [3,*] and Gaurav Samdani [3]

[1] *Principal Consultant, Cognizant Technologies Corp, Charlotte, NC, United states.*

[2] *Director Software Engineering, Lytx.inc, Charlotte, NC, United states.*

[3] *Department of Data science and Business Analytics, UNC Charlotte, Charlotte, NC, United states.*

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(02), 1388-1402

Publication history: Received on 28 March 2025; revised on 05 May 2025; accepted on 07 May 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.2.0635

Abstract

The rapid advancements in artificial intelligence and machine learning have led to the development of highly sophisticated models capable of superhuman performance in a variety of tasks. However, the increasing complexity of these models has also resulted in them becoming "black boxes", where the internal decision-making process is opaque and difficult to interpret. This lack of transparency and explainability has become a significant barrier to the widespread adoption of these models, particularly in sensitive domains such as healthcare and finance.

To address this challenge, the field of Explainable AI has emerged, focusing on developing new methods and techniques to improve the interpretability and explainability of machine learning models. This review paper aims to provide a comprehensive overview of the research exploring the combination of Explainable AI and traditional machine learning approaches, known as "hybrid models".

This paper discusses the importance of explainability in AI, and the necessity of combining interpretable machine learning models with black-box models to achieve the desired trade-off between accuracy and interpretability. It provides an overview of key methods and applications, integration techniques, implementation frameworks, evaluation metrics, and recent developments in the field of hybrid AI models.

The paper also delves into the challenges and limitations in implementing hybrid explainable AI systems, as well as the future trends in the integration of explainable AI and traditional machine learning. Altogether, this paper will serve as a valuable reference for researchers and practitioners working on developing explainable and interpretable AI systems.

Keywords: Explainable AI (XAI), Traditional Machine Learning (ML), Hybrid Models, Interpretability, Transparency, Predictive Accuracy, Neural Networks, Ensemble Methods, Decision Trees, Linear Regression, SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), Healthcare Analytics, Financial Risk Management, Autonomous Systems, Predictive Maintenance, Quality Control, Integration Techniques, Evaluation Metrics, Regulatory Compliance, Ethical Considerations, User Trust, Data Quality, Model Complexity, Future Trends, Emerging Technologies, Attention Mechanisms, Transformer Models, Reinforcement Learning, Data Visualization, Interactive Interfaces, Modular Architectures, Ensemble Learning, Post-Hoc Explainability, Intrinsic Explainability, Combined Models

Keywords: Explainable AI; Machine Learning; SHAP; LIME; Hybrid Models; Interpretability

* Corresponding author: Ganesh Viswanathan

1. Introduction

The widespread adoption of machine learning models in various industries has led to a growing demand for interpretable and explainable systems. Traditional machine learning techniques, such as linear regression and decision trees, have inherent explainability due to their relatively simple structure and the ability to trace the decision-making process. However, these models often struggle to achieve the same level of predictive performance as more complex models, such as neural networks and ensemble methods.

On the other hand, the impressive performance of these complex models often comes at the expense of interpretability, as their inner workings are challenging to understand and explain. This lack of transparency has raised concerns about the trustworthiness and reliability of these models, particularly in critical domains where decisions can have significant consequences.

To address this trade-off between predictive accuracy and interpretability, researchers have explored the development of "hybrid models" that combine the strengths of Explainable AI and traditional machine learning techniques. Hybrid models represent an innovative convergence of traditional machine learning techniques and explainable artificial intelligence (AI). These models aim to leverage the strengths of both domains, providing robust predictive capabilities while ensuring interpretability and transparency. Traditional machine learning excels in data-driven predictions and complex pattern recognition but often lacks the mechanisms to explain its decision-making processes. On the other hand, explainable AI focuses on making AI systems understandable to human users, addressing concerns about trust and accountability. The integration of these two paradigms into hybrid models seeks to create systems that are not only effective but also comprehensible to stakeholders.

The definition of hybrid models encompasses various methodologies designed to combine the predictive power of traditional machine learning with the interpretability of explainable AI. These models can take many forms, such as ensemble approaches where multiple algorithms are used to enhance performance while employing explainability techniques to clarify their outputs. Alternatively, some models may integrate explainability directly into the learning process, enabling more transparent decision-making from the outset. By blending these methodologies, hybrid models aim to produce outputs that are both accurate and accessible to users, minimizing the opacity often associated with complex machine learning algorithms.

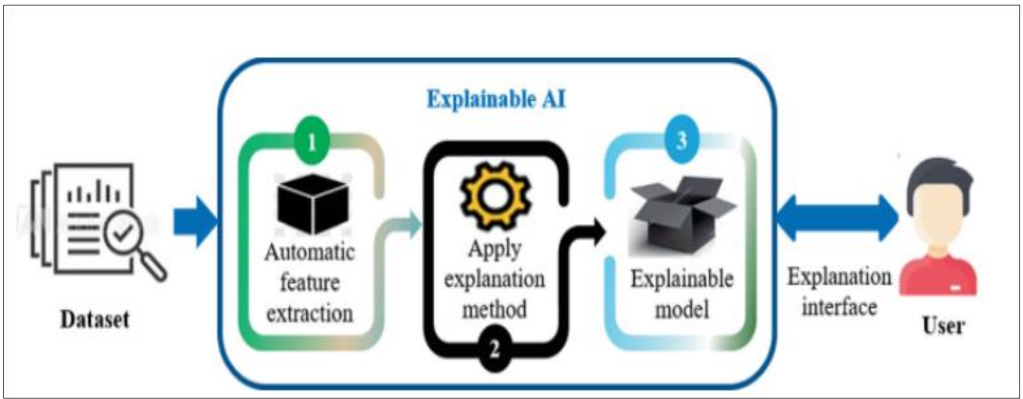

Figure 1 Above diagram explains the sample diagram of explainable AI

1.1. Importance of Explainability in AI

Explainability in AI is vital for fostering trust and understanding in machine learning systems, particularly as these technologies become increasingly integrated into critical domains such as healthcare and finance. As hybrid models combine the strengths of traditional machine learning with explainable AI, the need for transparent decision-making processes becomes even more pronounced. Students and researchers must recognize that the complexity of AI systems can lead to a lack of clarity regarding how decisions are made. This obscurity can hinder user acceptance and limit the effective deployment of AI solutions, making explainability an essential component in the development of robust AI applications.

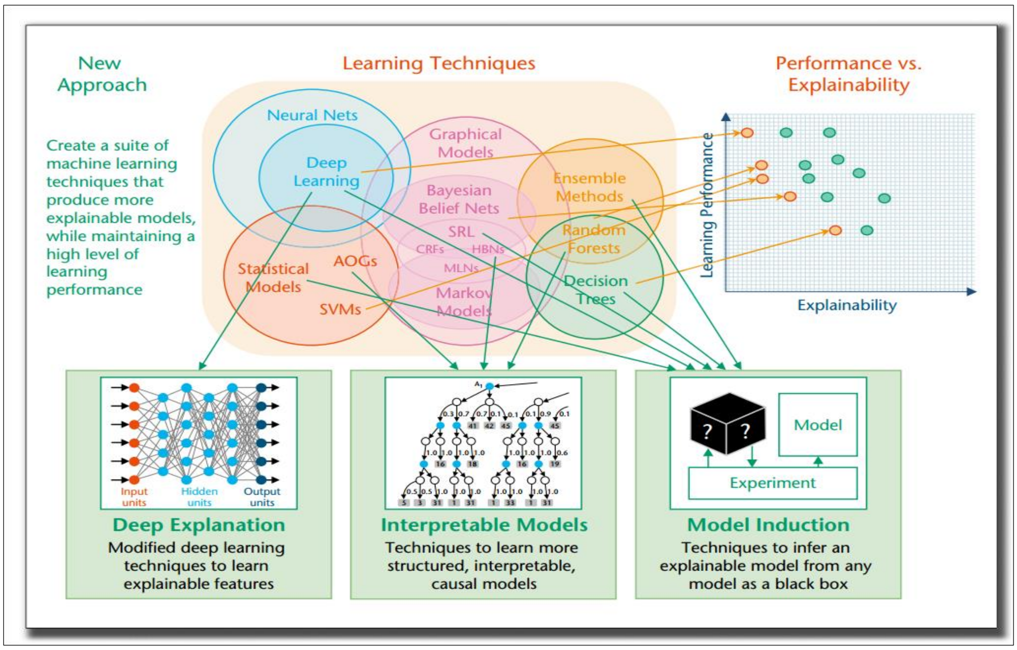

Figure 2 Strategies for Developing Explainable Models.

The evaluation of explainability in combined AI approaches demands the establishment of robust metrics. Students and researchers must engage with existing frameworks and contribute to the development of new evaluation criteria tailored to hybrid models. This involves understanding how to measure the quality of explanations and their effectiveness in improving user comprehension. The challenge lies in balancing the trade-offs between model performance and interpretability, which often necessitates innovative strategies that leverage the strengths of both traditional machine learning and explainable AI.

As hybrid models continue to evolve, the challenges associated with their implementation will also require careful consideration. Researchers must explore the technical and organizational barriers that impede the integration of explainable AI techniques within existing machine learning frameworks. This includes addressing issues related to data quality, model complexity, and user engagement. By conducting case studies that illustrate successful applications in areas like financial risk assessment, scholars can provide valuable insight into overcoming these challenges, thereby paving the way for more widespread adoption of explainable hybrid AI systems.

2. Methodology

Hybrid models that combine Explainable AI (XAI) and traditional machine learning (ML) are gaining traction due to their ability to balance performance with interpretability. These models aim to leverage the strengths of both approaches to create systems that are not only powerful but also transparent and trustworthy.

2.1. Key Methods and Applications

Hybrid models that combine Explainable AI and traditional machine learning can be broadly categorized into three main methods:

- **Post-Hoc Explain ability**: After a traditional ML model has made its predictions, post-hoc methods are used to interpret and explain those results. Tools like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) fall into this category. These techniques provide insights into how a model's outputs were derived by examining feature contributions or approximating the model's behavior locally.
- **Intrinsic Explain ability**: This approach integrates explainability directly into the model architecture. Examples include decision trees and linear models that are inherently interpretable. Hybrid models in this category might use simple, interpretable models for certain parts of the problem while leveraging more complex algorithms where necessary.

- **Combined Models**: This method involves using explainable models in conjunction with traditional ML models. For instance, a complex neural network might be paired with a simpler, interpretable model to provide explanations for the predictions. This can involve using rule-based systems or other transparent methods to oversee or augment the predictions of the opaquer model.

These approaches help ensure that the predictions and decisions made by AI systems are understandable, which is crucial for trust, transparency, and compliance with regulations.

3. Applications

Hybrid models combining Explainable AI and traditional machine learning have been explored in a variety of application domains, including:

- Healthcare: Hybrid models have been applied to medical diagnosis and treatment recommendations, providing both accurate predictions and interpretable explanations to support clinical decision-making.
- Financial risk management: Explainable AI models have been used to enhance the transparency and accountability of AI-based decision-making in the financial industry, where interpretability is crucial for regulatory compliance and consumer trust.
- Autonomous systems: In the development of self-driving cars and other autonomous systems, hybrid models can help ensure the safety and reliability of the AI-powered decision-making process by providing transparency and explainability.
- Industrial applications: Hybrid models have been used in manufacturing, predictive maintenance, and quality control to enhance decision-making and enable better human-AI collaboration.

These hybrid approaches are essential for developing AI systems that are both effective and transparent, fostering greater trust and adoption in various industries.

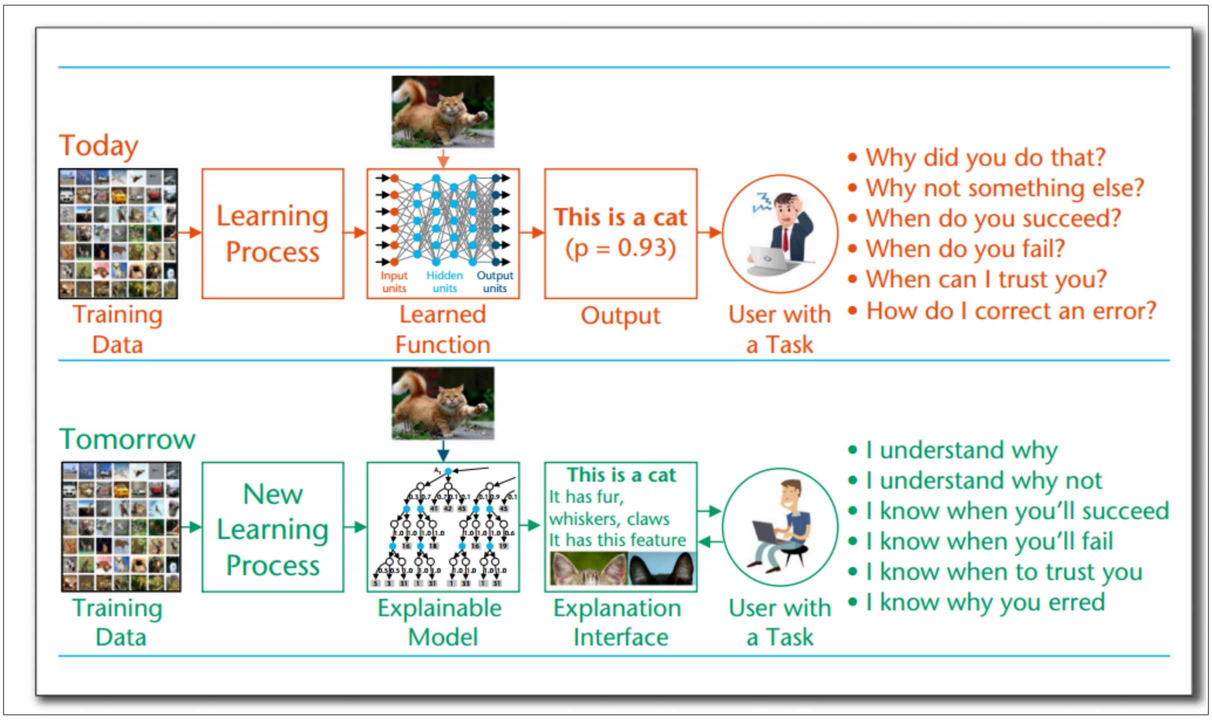

Figure 3 The XAI Concept.

For instance, decision trees can be integrated with more complex models like neural networks. In this scenario, the decision tree serves as the first layer to provide initial predictions that are easily interpretable, while the neural network refines these predictions through deeper learning. Such a hybrid approach allows researchers to leverage the strengths of both methodologies, yielding models that are not only robust but also offer insights into the decision-making process.

Integration techniques also play a crucial role in the development of hybrid AI systems. These techniques can range from modular designs, where components of XAI and traditional ML are distinctly defined but work collaboratively, to tighter integrations where the outputs of one model directly inform the other. This flexibility allows for tailored solutions that can be adapted to specific application needs, whether in healthcare analytics, financial risk assessment, or natural language processing. The choice of integration method greatly influences the model's effectiveness and interpretability, underscoring the need for careful consideration during the design phase.the evaluation of hybrid AI models necessitates the establishment of metrics that capture both performance and explainability. Traditional evaluation metrics, such as accuracy and F1 score, must be complemented by measures that assess the interpretability of the model's outputs. The development of new metrics that quantify the clarity and usefulness of explanations provided by hybrid models is essential for advancing the field. By focusing on these evaluation criteria, researchers and practitioners can ensure that hybrid AI systems not only perform well but also align with user expectations and ethical standards, ultimately fostering a more responsible and impactful deployment of AI technologies.

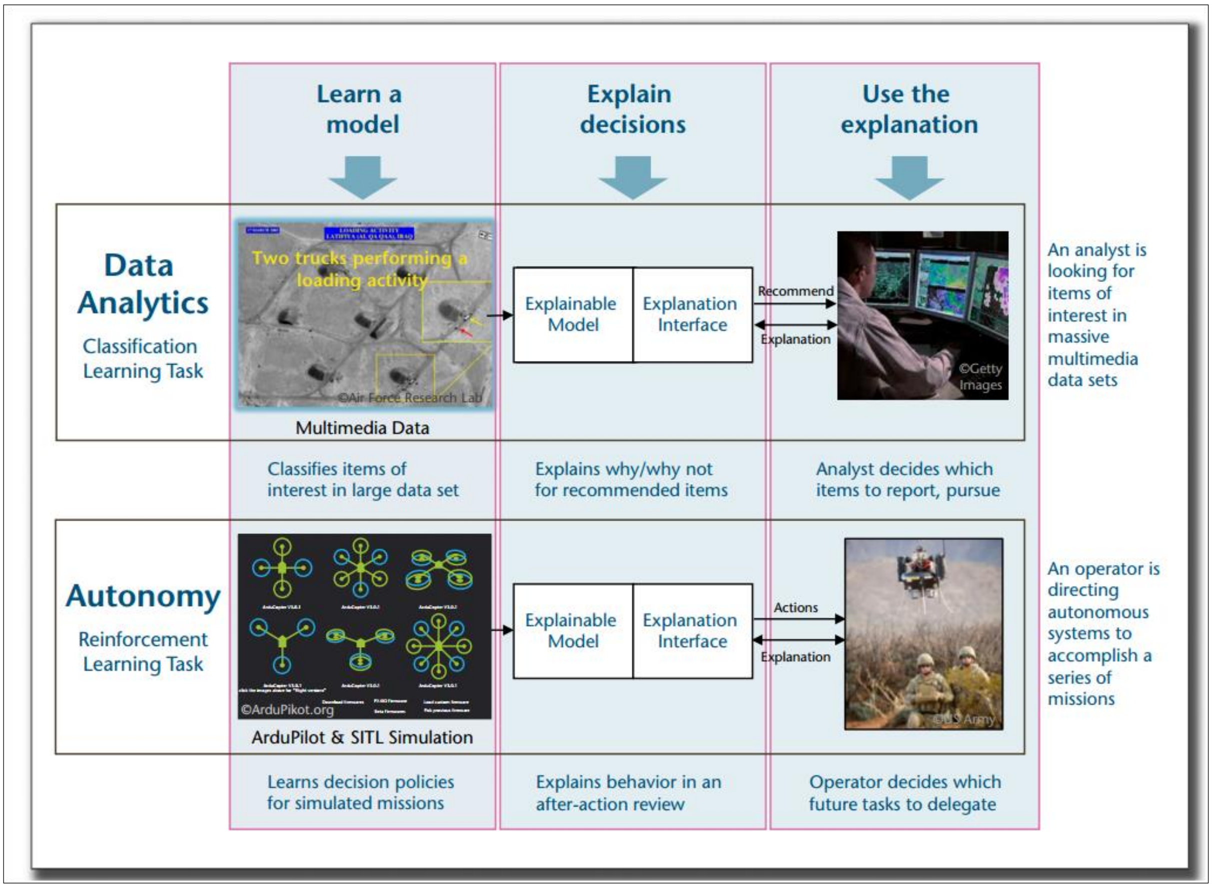

Figure 4 XAI Challenge Problem Areas.

4. Integration Techniques for Explainable AI in Hybrid Models

Integrating model strategies is crucial for merging Explainable AI (XAI) with conventional machine learning (ML) techniques effectively. This combination seeks to boost model precision while preserving the clarity and comprehensibility of AI decision-making processes. A variety of strategies, from ensemble methods to intricate architectures that capitalize on both paradigms' strengths, can be utilized for this integration. These techniques not only offer a structure for comprehending AI decisions but also aid in crafting more stable and dependable systems across various sectors.

A common strategy is ensemble learning, where several models are trained on the same problem, and their predictions are amalgamated to enhance performance. This is especially useful when combining explainable models like decision trees or rule-based systems with opaque models such as deep neural networks. Techniques like bagging or boosting allow for the creation of hybrid systems that retain interpretability while gaining from the advanced ML methods' accuracy. This method fosters a deeper insight into the factors influencing predictions, thus improving complex models' explainability.

Another effective method is developing modular architectures that segregate the explainability and predictive components. In this structure, a conventional ML model manages prediction tasks, while an XAI module interprets the model's results. This separation ensures clear role demarcation, allowing researchers to enhance the ML component's predictive capacity without affecting the XAI module's interpretability. This modular concept has been promising, particularly in fields like healthcare.

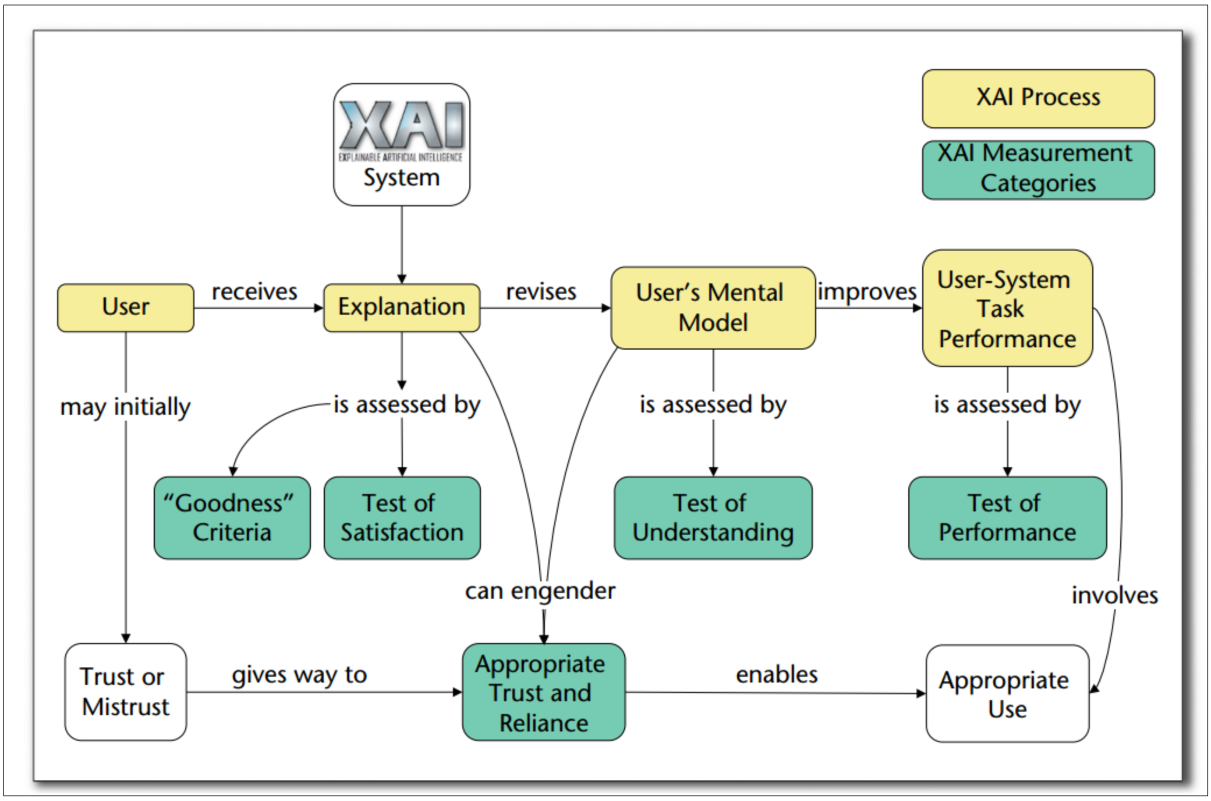

Figure 5 Initial Model of the Explanation Process and Explanation Effectiveness Measurement Categories

5. Frameworks for Implementation

Developing a structured framework is crucial for effectively creating hybrid models that combine Explainable AI and traditional machine learning. These frameworks guide researchers and practitioners through the complexities of integrating the two approaches. A well-defined framework helps identify appropriate hybridization methods and techniques, emphasizing the importance of transparency and interpretability in AI models. This is especially crucial in fields where decision-making processes must be clear to stakeholders and regulatory bodies, such as healthcare and finance.

Selecting appropriate integration techniques is a crucial aspect of the implementation framework. Researchers should evaluate various approaches, such as post-hoc interpretability methods and inherently interpretable models. Post-hoc techniques, like LIME and SHAP, provide insights into model predictions after training, while inherently interpretable models, such as decision trees, offer transparency from the outset. The choice of integration techniques directly affects the model's ability to meet explainability requirements while maintaining performance standards, emphasizing the need for a thorough evaluation process.

Examples in healthcare analytics highlight the practical use of these frameworks. Hybrid models in this field have shown the potential to improve diagnostic accuracy while providing essential interpretability for clinical decision-making. For example, combining deep learning for image analysis with rule-based systems can create models that outperform traditional methods and offer clear explanations for their predictions. The implementation framework should guide researchers in designing hybrid models that leverage the strengths of both approaches, ensuring the resulting systems are effective and explainable.

Evaluation metrics are central to assessing the efficacy of hybrid models that combine explainability and traditional machine learning. Measures such as fidelity, robustness, and user satisfaction must be tailored to reflect the dual aims of model performance and interpretability. A well-designed framework should provide guidelines for selecting appropriate evaluation metrics, aligned with the specific objectives of the hybrid model. This not only facilitates measuring the models' effectiveness but also aids in communicating their value to stakeholders, cultivating trust and acceptance in AI-driven solutions.

6. Evaluation Metrics for Explainability in Combined AI Approaches

The evaluation of hybrid AI models necessitates the establishment of metrics that capture both performance and explainability. Accuracy, F1 score, and other traditional machine learning metrics can assess the predictive capabilities of the models. However, evaluating the interpretability of these models requires specialized measures that quantify the clarity, usefulness, and fidelity of the explanations provided.

Measure	Description
ML Model performance	
Various measures (on a per-challenge problem area basis)	Accuracy/performance of the ML model in its given domain (to understand whether performance improved or degraded relative to state-of-the-art nonexplainable baselines)
Explanation Effectiveness	
Explanation goodness	Features of explanations assessed against criteria for explanation goodness
Explanation satisfaction	User's subjective rating of explanation completeness, usefulness, accuracy, and satisfaction
Mental model understanding	User's understanding of the system and the ability to predict the system's decisions/behavior in new situations
User task performance	Success of the user performing the tasks for which the system is designed to support
Appropriate Trust and Reliance	User's ability to know when to, and when not to, trust the system's recommendations and decisions

Figure 6 Measurement Categories

6.1. Metrics for Assessing Explainability

Metrics for assessing explainability play a crucial role in evaluating the effectiveness of hybrid models that combine explainable AI and traditional machine learning. These metrics serve as benchmarks for understanding how well a model conveys its decision-making process to users, which is essential for fostering trust and transparency. With the increasing complexity of AI systems, especially in hybrid frameworks, it becomes imperative to establish clear and measurable criteria that can adequately assess how understandable and interpretable these systems are.

One widely adopted metric is fidelity, which measures the degree to which the explanation provided by the model aligns with its actual decision-making process. In hybrid models, where traditional ML techniques are integrated with explainable AI methods, fidelity becomes particularly significant. It ensures that the explanations generated are not only plausible but also accurate representations of the model's behavior. High fidelity indicates that users can rely on the explanations to make informed decisions, while low fidelity can lead to confusion and mistrust.

Another important metric is consistency, which evaluates how stable the explanations are across similar inputs. In the context of hybrid models, consistency is critical because it reflects the reliability of the explanations provided by the model. If the explanations vary significantly for similar cases, it may suggest that the model is not sufficiently robust or that the integration of explainable AI techniques is flawed. Consistency reassures users that the model's behavior is predictable and helps in mitigating the risks associated with unexpected outcomes.

Interpretability is also a vital metric, encompassing the ease with which users can understand the reasoning behind a model's predictions. In hybrid systems, achieving a balance between the performance of traditional machine learning algorithms and the interpretability offered by explainable AI methods is key. Metrics that quantify interpretability can include the complexity of the explanation and the number of features considered in the model. A model that provides simpler, more concise explanations is often favored, as it allows users to grasp the rationale behind predictions more easily.

7. Trade-offs Between Accuracy and Interpretability

In the realm of artificial intelligence (AI) and machine learning (ML), the balance between accuracy and interpretability presents a critical dilemma. As hybrid models, which combine elements of explainable AI with traditional machine learning techniques, gain traction, understanding the trade-offs between these two facets becomes essential. Accuracy is often prioritized in traditional ML frameworks, where complex algorithms leverage vast datasets to produce highly effective predictive models. However, these models typically operate as "black boxes," offering little insight into their decision-making processes. This lack of transparency can lead to issues, particularly in domains sensitive to ethical considerations and regulatory compliance, such as healthcare and finance.

Interpretability, on the other hand, emphasizes the ability to understand and explain model behavior, often at the expense of some level of predictive accuracy. Models that prioritize interpretability, such as decision trees or linear regressions, provide clear pathways to understanding how input variables influence outcomes. While these models may not achieve the same level of predictive power as more complex algorithms, their transparency can foster trust among users and facilitate better decision-making. In contexts like healthcare analytics, where understanding the rationale behind a diagnosis is crucial, interpretability can significantly enhance stakeholder engagement and compliance with ethical standards.

Hybrid models strive to bridge this gap by integrating the strengths of both traditional ML and explainable AI. By leveraging advanced techniques such as rule-based systems or attention mechanisms, hybrid approaches can deliver high accuracy while providing insights into the model's inner workings. This dual capability allows practitioners to deploy models that not only perform effectively but also meet the demands for transparency. However, achieving this balance requires careful consideration of the model design and evaluation metrics, which must account for both predictive performance and interpretability.

8. Challenges and Limitations in Implementing Hybrid Explainable AI Systems

The development and deployment of hybrid AI models that combine explainable AI and traditional machine learning face several challenges and limitations.

8.1. Challenges in Integration

Integration of explainable AI (XAI) with traditional machine learning (ML) presents a range of challenges that researchers and practitioners must navigate to create effective hybrid models. One primary challenge is the inherent complexity of combining algorithms from two distinct paradigms. Traditional ML models often prioritize performance and predictive accuracy, while XAI emphasizes interpretability and transparency. Balancing these competing objectives can result in trade-offs that may undermine the strengths of either approach. Consequently, researchers must develop innovative integration techniques that allow for seamless interaction between the two systems without sacrificing the integrity or usability of the model.

Another significant challenge lies in the evaluation of integrated models. The metrics traditionally used to assess ML performance, such as accuracy, precision, and recall, may not adequately capture the explainability aspect of hybrid models. Researchers must establish new evaluation frameworks that encompass both the predictive capabilities and the interpretability of the combined systems. This involves identifying suitable metrics that can quantify the effectiveness of explanations provided by XAI components, ensuring that end-users can trust and understand the output of the hybrid model. Without robust evaluation criteria, it becomes difficult to gauge the success of integration efforts.

The domain-specific nature of applications further complicates the integration of XAI and traditional ML. Different fields, such as healthcare and finance, have unique requirements and expectations regarding explainability and model performance. For instance, in healthcare analytics, stakeholders may demand elevated levels of transparency to ensure patient safety and compliance with regulations. Conversely, in financial risk assessment, the focus may be on predictive accuracy and the ability to handle vast amounts of data. Researchers must tailor their hybrid models to meet these varying demands, which often necessitates extensive domain knowledge and collaboration between experts in AI and the specific application area.

User-centric design also presents a challenge in developing effective hybrid models. The end-users of these systems may include data scientists, business analysts, and domain experts, each with various levels of familiarity with AI technologies and varying expectations regarding model transparency. Designing hybrid solutions that cater to these diverse user needs requires a careful consideration of how explanations are presented and what types of insights are most valuable. This involves iterative testing and refinement to ensure that the explanations generated by the XAI components resonate with users and enhance their understanding of the model's decision-making process.

Finally, regulatory compliance and ethical considerations play a crucial role in the integration of XAI and traditional ML. As regulations around data privacy and algorithmic accountability continue to evolve, researchers must ensure that hybrid models adhere to relevant standards while also providing adequate explanations for their predictions. This necessitates a proactive approach to understanding and implementing ethical guidelines, which can vary significantly across jurisdictions and application areas. Addressing these challenges requires a commitment to developing hybrid systems that not only perform well but also align with societal values and legal requirements, paving the way for responsible AI adoption.

8.2. Technical Challenges

Integrating explainable AI with traditional machine learning frameworks presents multifaceted technical challenges that require a comprehensive understanding of both domains. One key issue stem from the inherent complexity of explainability techniques. Traditional machine learning models often prioritize performance metrics like accuracy and efficiency, while explainable AI methods focus on making model predictions interpretable to users. This divergence can make it difficult to reconcile high-performing algorithms with explainability, particularly when using advanced techniques like deep learning. The main challenge is ensuring that the provided explanations do not undermine the predictive power of the models.

Integrating diverse data sources poses a significant challenge for hybrid models. Explainable AI often relies on structured data to generate clear explanations, while traditional machine learning can handle unstructured data like text and images. This data heterogeneity complicates the development of hybrid systems that maintain consistency and reliability in both prediction and explanation. Researchers must develop integration techniques that can seamlessly process various data formats while preserving the quality of interpretability across the model.

Evaluating the effectiveness of hybrid models poses another technical challenge. While various metrics exist for assessing traditional machine learning performance, evaluating explainability remains less standardized. The subjective nature of interpretability means stakeholders may have diverse expectations regarding the explanations provided by AI systems. Establishing uniform evaluation metrics for explainability across different hybrid models is crucial but still an ongoing effort. This challenge requires collaboration among researchers to develop a framework that can accurately assess both the performance and explainability of these integrated systems.

Implementing hybrid explainable AI systems in real-world applications such as healthcare analytics and financial risk assessment introduces additional technical complexities. These domains often require compliance with strict regulatory standards, which can conflict with the flexibility needed to create interpretable models. Moreover, the integration process itself can be resource-intensive, necessitating advanced computational capabilities and expertise in both AI paradigms. Addressing these implementation challenges requires a concerted effort to develop tools and methodologies that streamline the hybridization process while ensuring adherence to industry regulations.

8.3. Future Trends in Explainable AI and Traditional Machine Learning Integration

The integration of explainable AI and traditional machine learning is a rapidly evolving field that holds significant promise for addressing the growing demand for transparent and trustworthy AI systems.

9. Emerging Technologies

Emerging technologies are rapidly changing the field of artificial intelligence and machine learning, especially in hybrid models that combine explainable AI and traditional machine learning. These advancements improve the interpretability of complex algorithms and address the need for transparency in AI systems. Researchers and students should recognize the importance of incorporating emerging technologies to develop robust hybrid models that can explain decision-making while maintaining high predictive performance.

Advanced neural networks, such as attention mechanisms and transformer models, have significantly impacted natural language processing and image analysis. These technologies enable the creation of hybrid models that combine the strengths of explainable AI and traditional machine learning. By using techniques like layer-wise relevance propagation or SHAP, these hybrid models can offer valuable insights into feature importance and decision-making processes, addressing the interpretability challenges often linked to deep learning.

A crucial area of development is in reinforcement learning, where modern technologies allow for the creation of adaptive AI systems that can learn from their environment while explaining their actions. This adaptability is crucial in fields like healthcare analytics, where hybrid models can predict patient outcomes and explain treatment recommendations. By using these technologies, researchers can better address issues of user trust and regulatory compliance, ensuring that AI systems are effective and responsible.

The integration of explainable AI and traditional machine learning is being improved through new data visualization tools and interactive interfaces. These technologies help users, such as healthcare professionals and financial analysts, engage with AI systems more effectively. As students and researchers explore these interfaces, they will learn how user-centric design principles can create more intuitive and accessible hybrid solutions. This can foster a better understanding of AI outputs and promote wider acceptance of these systems across different sectors.

As hybrid models evolve, researchers must focus on compliance with emerging regulations and ethical standards for AI technologies. They should stay informed about regulatory developments to ensure their work aligns with the best practices in transparency and accountability. By embracing modern technologies, students and researchers can create hybrid AI systems that are both innovative and responsible, enabling AI to serve humanity effectively and ethically.

9.1. Anticipated Developments in the Field

The field of combining explainable AI and traditional machine learning is advancing rapidly. As the need for transparent AI grows, researchers are integrating explainability into traditional machine learning. This aims to improve interpretability without reducing performance, resulting in more trustworthy AI. Expected developments include better algorithms that blend the strengths of both approaches, allowing practitioners to use traditional machine learning's predictive power while gaining the clarity and insight from explainable AI.

Healthcare analytics case studies are expected to be crucial in highlighting practical applications of hybrid models. The complexity of medical data and the need for interpretability in clinical settings make healthcare an ideal field for these advancements. Future developments may focus on creating hybrid models that predict patient outcomes and provide explanations that clinicians can understand and trust. This ability to explain predictions could significantly improve decision-making, leading to better patient care and outcomes.

In financial risk assessment, hybrid models are expected to become more prominent as regulators emphasize the need for explainability in AI-based decision-making. As financial firms adopt these models, anticipated developments will likely include user-focused design approaches prioritizing transparency and usability. By prioritizing the end-user experience, researchers can ensure hybrid AI solutions meet regulatory requirements and user needs, promoting acceptance and adoption within the industry.

9.2. Research Directions and Opportunities

The study of hybrid models that blend Explainable AI with conventional machine learning offers many research avenues and possibilities. As the need for transparency and interpretability in AI grows, researchers can explore methods to effectively combine XAI principles with traditional ML algorithms. This includes creating frameworks to smoothly integrate explainability into existing ML workflows. By concentrating on the design of hybrid models, researchers can contribute to a deeper understanding of how these systems function and how their outcomes can be interpreted by users.

Evaluating hybrid models requires new metrics that capture both predictive performance and explanatory power. Traditional machine learning prioritizes accuracy, often overlooking interpretability. Researchers should develop multi-dimensional evaluation frameworks that assess both model performance and the quality of explanations. These comprehensive metrics will aid model selection and guide the creation of more user-friendly AI applications.

Case studies in specific domains, such as healthcare, insurance, and finance, present valuable research opportunities. Applying hybrid models in healthcare analytics can advance patient care by providing clinicians with interpretable insights from complex data. Similarly, in financial risk assessment, integrating explainable features into predictive models can enhance decision-making and regulatory compliance. Through detailed case studies, researchers can identify the best practices, challenges, and successful strategies for implementing these hybrid approaches in real-world settings.

Researchers should investigate challenges in implementing hybrid explainable AI systems. They should identify barriers to adoption, such as computational complexity, data quality, and user trust. Understanding these challenges will help develop solutions to enable practical deployment of hybrid models. Exploring user-centric design principles will also enhance the usability of these systems, ensuring the explanations are accessible and meaningful to various stakeholders, including non-experts.

The regulatory landscape and ethical considerations around hybrid AI models present valuable research opportunities. As new regulations emerge, researchers should explore how hybrid models can meet evolving standards while preserving their explanatory capabilities. This intersection of technology and ethics is crucial in guiding the future development of AI applications. By addressing these challenges, research can contribute to the design of responsible AI systems that prioritize transparency, accountability, and fairness - paving the way for innovative applications across diverse fields.

10. Conclusion

Hybrid models that combine Explainable AI and traditional machine learning techniques offer a promising approach to address the trade-off between predictive accuracy and interpretability. By leveraging the strengths of both paradigms, these hybrid models can deliver high-performing yet transparent and explainable AI systems that are well-suited for critical application domains. As the field of Explainable AI continues to evolve, further research and development of hybrid models will be crucial to unlocking the full potential of AI in areas where both performance and interpretability are paramount.

Techniques like SHAP and Integrated Gradients have emerged as popular methods for creating explainable hybrid models.

The paper concludes by highlighting the pivotal role of explainability within artificial intelligence. It underscores the imperative to amalgamate interpretable machine learning models with the more enigmatic 'black box' models. This amalgamation aims to strike a balance between accuracy and interpretability. The document presents an exhaustive survey of key methodologies, applications, integration tactics, implementation frameworks, evaluation metrics, and recent progress in the realm of hybrid AI models.

Moreover, the paper also delves into the challenges and limitations encountered in the deployment of hybrid explainable AI systems. It investigates prospective trajectories in the confluence of explainable AI and traditional machine learning methods. In essence, this paper serves as a critical compendium for both researchers and practitioners committed to fostering the development of explainable and intelligible AI systems.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Abad, Z S H., Thieme, A., Yang, Z., Wang, Y., Lin, B., Gevaert, O., Li, L., Jain, R., & Rahmani, A M. (2023, January 1). Foundation Metrics: Quantifying Effectiveness of Healthcare Conversations powered by Generative AI. Cornell University. https://doi.org/10.48550/arxiv.2309.12444

[2] AI data transparency: an exploration through the lens of AI incidents. (2024, September 4). https://doi.org/10.48550/arXiv.2409.03307

[3] Aizenberg, E., & Hoven, J V D. (2020, July 1). Designing for human rights in AI. SAGE Publishing, 7(2), 205395172094956-205395172094956. https://doi.org/10.1177/2053951720949566

[4] Akata, Z., Balliet, D., Rijke, M D., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K V., Hoos, H H., Hung, H., Jonker, C M., Monz, C., Neerincx, M A., Oliehoek, F A., Prakken, H., Schlobach, S., Gaag, L V D., Harmelen, F V., . . . Welling, M. (2020, July 31). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. IEEE Computer Society, 53(8), 18-28. https://doi.org/10.1109/mc.2020.2996587

[5] Antoniadi, A M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B A., & Mooney, C. (2021, May 31). Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. Multidisciplinary Digital Publishing Institute, 11(11), 5088-5088. https://doi.org/10.3390/app11115088

[6] Bajwa, J., Munir, U., Nori, A V., & Williams, B. (2021, July 1). Artificial intelligence in healthcare: transforming the practice of medicine. Royal College of Physicians, 8(2), e188-e194. https://doi.org/10.7861/fhj.2021-0095

[7] Bateni, A., Chan, M., & Eitel-Porter, R. (2022, January 1). AI Fairness: from Principles to Practice. Cornell University. https://doi.org/10.48550/arXiv.2207.

[8] Bruckert, S., Finzel, B., & Schmid, U. (2020, September 24). The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions. Frontiers Media, 3. https://doi.org/10.3389/frai.2020.507973

[9] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020, April 24). Explainable AI in Fintech Risk Management. Frontiers Media, 3. https://doi.org/10.3389/frai.2020.00026

[10] Caudle, K., Hoover, R C., Alphonsus, A., & Shradha, S. (2019, December 1). Advanced Decision Making and Interpretability through Neural Shrubs. , 6, 489-494. https://doi.org/10.1109/icmla.2019.00091

[11] Cheng, L., Varshney, K R., & Liu, H. (2021, August 28). Socially Responsible AI Algorithms: Issues, Purposes, and Challenges. AI Access Foundation, 71, 1137-1181. https://doi.org/10.1613/jair.1.12814

[12] Cinà, G., Röber, T E., Goedhart, R., & Birbil, Ş İ. (2022, January 1). Why we do need Explainable AI for Healthcare. Cornell University. https://doi.org/10.48550/arXiv.2206.

[13] Cockburn, I., Henderson, R., & Stern, S. (2018, March 1). The Impact of Artificial Intelligence on Innovation. https://doi.org/10.3386/w24449

[14] Díaz-Rodríguez, N., Ser, J D., Coeckelbergh, M., Prado, M L D., Herrera-Viedma, E., & Herrera, F. (2023, June 24). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Elsevier BV, 99, 101896-101896. https://doi.org/10.1016/j.inffus.2023.101896

[15] Doshi-Velez, F., & Kim, B. (2017, January 1). Towards A Rigorous Science of Interpretable Machine Learning. Cornell University. https://doi.org/10.48550/arXiv.1702.

[16] Emaminejad, N., North, A M., & Akhavian, R. (2022, January 1). Trust in AI and Implications for the AEC Research: A Literature Analysis. Cornell University. https://doi.org/10.48550/arXiv.2203.

[17] Explain Your AI. (n.d). https://www.h2o.ai/wp-content/uploads/2019/08/An-Introduction-to-Machine-Learning-Interpretability-Second-Edition.pdf

[18] Ferrara, E. (2023, January 1). Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies. Cornell University. https://doi.org/10.48550/arXiv.2304.

[19] Gilpin, L H., Bau, D., Yuan, B Z., Bajwa, A., Specter, M., & Kagal, L. (2018, January 1). Explaining Explanations: An Overview of Interpretability of Machine Learning. Cornell University. https://doi.org/10.48550/arXiv.1806.

[20] Gilpin, L H., Bau, D., Yuan, B Z., Bajwa, A., Specter, M., & Kagal, L. (2018, January 1). Explaining Explanations: An Overview of Interpretability of Machine Learning. Cornell University. https://doi.org/10.48550/arxiv.1806.00069

[21] Giuste, F., Shi, W., Zhu, Y., Naren, T., Isgut, M., Sha, Y., Li, T., Gupte, M., & Wang, M D. (2022, June 23). Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review. Institute of Electrical and Electronics Engineers, 16, 5-21. https://doi.org/10.1109/rbme.2022.3185953

[22] Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J P., Yordanova, K., Vered, M., Nair, R., Abreu, P H., Blanke, T., Pulignano, V., Prior, J O., Lauwaert, L., Reijers, W., Depeursinge, A., Andrearczyk, V., & Müller, H. (2022, September 6). A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. Springer Science+Business Media, 56(4), 3473-3504. https://doi.org/10.1007/s10462-022-10256-8

[23] Guo, A., Kamar, E., Vaughan, J W., Wallach, H., & Morris, M R. (2019, January 1). Toward Fairness in AI for People with Disabilities: A Research Roadmap. Cornell University. https://doi.org/10.48550/arXiv.1907.

[24] Hall, P B., Gill, N., & Schmidt, N. (2019, January 1). Proposed Guidelines for the Responsible Use of Explainable Machine Learning. Cornell University. https://doi.org/10.48550/arxiv.1906.03533

[25] Hanif, A. (2021, January 1). Towards Explainable Artificial Intelligence in Banking and Financial Services. Cornell University. https://doi.org/10.48550/arxiv.2112.08441

[26] Hernández-Orallo, J. (2014, January 1). AI Evaluation: past, present, and future. Cornell University. https://doi.org/10.48550/arXiv.1408.

[27] Hernández, E. (2024, May 2). Towards an Ethical and Inclusive Implementation of Artificial Intelligence in Organizations: A Multidimensional Framework. Cornell University. https://doi.org/10.48550/arXiv.2405.

[28] Holstein, K., Aleven, V., & Rummel, N. (2020, January 1). A Conceptual Framework for Human–AI Hybrid Adaptivity in Education. Springer Science+Business Media, 240-254. https://doi.org/10.1007/978-3-030-52237-7_20

[29] Hulsen, T. (2023, August 10). Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare. Multidisciplinary Digital Publishing Institute, 4(3), 652-666. https://doi.org/10.3390/ai4030034

[30] Human-Centered Responsible Artificial Intelligence: Current & Future Trends. (2023, April 18). https://export.arxiv.org/pdf/2302.08157v1.pdf

[31] Karabacak, M., & Margetis, K. (2023, May 21). Embracing Large Language Models for Medical Applications: Opportunities and Challenges. Cureus, Inc.. https://doi.org/10.7759/cureus.39305

[32] Kaur, D., Islam, S N., Mahmud, M A., & Dong, Z. (2020, January 1). Energy Forecasting in Smart Grid Systems: A Review of the State-of-the-art Techniques. Cornell University. https://doi.org/10.48550/arXiv.2011.

[33] Kazemi, M., Moradkhani, D., & Alipour, A A. (2023, January 1). Application of Random Forest and Support Vector Machine for Investigation of Pressure Filtration Performance, a Zinc Plant Filter Cake Modeling. Cornell University. https://doi.org/10.48550/arXiv.2307.

[34] Kelly, C., Karthikesalingam, A., Suleyman, M., Corrado, G S., & King, D. (2019, October 29). Key challenges for delivering clinical impact with artificial intelligence. BioMed Central, 17(1). https://doi.org/10.1186/s12916-019-1426-2

[35] Khalifa, M., Albadawy, M., & Iqbal, U. (2024, January 1). Advancing Clinical Decision Support: The Role of Artificial Intelligence Across Six Domains. Elsevier BV, 5, 100142-100142. https://doi.org/10.1016/j.cmpbup.2024.100142

[36] Kim, S., Ko, B C., & Nam, J. (2021, April 25). Model Simplification of Deep Random Forest for Real-Time Applications of Various Sensor Data. Multidisciplinary Digital Publishing Institute, 21(9), 3004-3004. https://doi.org/10.3390/s21093004

[37] Kim, S., Yoon, S M., Yang, M C., Choi, J., Akay, H., & Burnell, E. (2019, January 1). AI for design: Virtual design assistant. Elsevier BV, 68(1), 141-144. https://doi.org/10.1016/j.cirp.2019.03.024

[38] Kuiper, O X., Berg, M V D., Burgt, J V D., & Leijnen, S. (2022, January 1). Exploring Explainable AI in the Financial Sector: Perspectives of Banks and Supervisory Authorities. Springer Science+Business Media, 105-119. https://doi.org/10.1007/978-3-030-93842-0_6

[39] Lai, T. (2024, January 8). Interpretable Medical Imagery Diagnosis with Self-Attentive Transformers: A Review of Explainable AI for Health Care. , 4(1), 113-126. https://doi.org/10.3390/biomedinformatics4010008

[40] Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019, January 27). Faithful and Customizable Explanations of Black Box Models. https://doi.org/10.1145/3306618.3314229

[41] Leech, G., Garfinkel, S., Yagudin, M., Briand, A., & Zhuravlev, A. (2024, February 6). Ten Hard Problems in Artificial Intelligence We Must Get Right. Cornell University. https://doi.org/10.48550/arxiv.2402.04464

[42] Leitão, D., Saleiro, P., Figueiredo, M A T., & Bizarro, P. (2022, January 1). Human-AI Collaboration in Decision-Making: Beyond Learning to Defer. Cornell University. https://doi.org/10.48550/arxiv.2206.13202

[43] Li, B., Qi, P., Bo, L., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2022, August 18). Trustworthy AI: From Principles to Practices. Association for Computing Machinery, 55(9), 1-46. https://doi.org/10.1145/3555803

[44] Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., Bian, J., & Dou, D. (2022, September 14). Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. Springer Science+Business Media, 64(12), 3197-3234. https://doi.org/10.1007/s10115-022-01756-8

[45] Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A K., & Tang, J. (2021, January 1). Trustworthy AI: A Computational Perspective. Cornell University. https://doi.org/10.48550/arxiv.2107.06641

[46] Loh, H W., Ooi, C P., Seoni, S., Barua, P D., Molinari, F., & Acharya, U R. (2022, September 27). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Elsevier BV, 226, 107161-107161. https://doi.org/10.1016/j.cmpb.2022.107161

[47] Lu, S., Swisher, C L., Chung, C., Jaffray, D A., & Sidey-Gibbons, C. (2023, February 28). On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. Frontiers Media, 13. https://doi.org/10.3389/fonc.2023.1129380

[48] Mandala, S K. (2023, January 1). XAI Renaissance: Redefining Interpretability in Medical Diagnostic Models. Cornell University. https://doi.org/10.48550/arxiv.2306.01668

[49] Markus, A F., Kors, J A., & Rijnbeek, P R. (2020, December 10). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Elsevier BV, 113, 103655-103655. https://doi.org/10.1016/j.jbi.2020.103655

[50] Martínez-Garcia, M., & Hernández-Lemus, E. (2022, January 25). Data Integration Challenges for Machine Learning in Precision Medicine. Frontiers Media, 8. https://doi.org/10.3389/fmed.2021.784455

[51] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad-Khasmakhi, N., Chenaghlu, M., & Gao, J. (2021, April 17). Deep Learning-based Text Classification. Association for Computing Machinery, 54(3), 1-40. https://doi.org/10.1145/3439726

[52] Misheva, B H., Jaggi, D., Posth, J., Gramespacher, T., & Osterrieder, J. (2021, December 21). Audience-Dependent Explanations for AI-Based Risk Management Tools: A Survey. Frontiers Media, 4. https://doi.org/10.3389/frai.2021.794996

[53] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I D., & Gebru, T. (2019, January 9). Model Cards for Model Reporting. https://doi.org/10.1145/3287560.3287596

[54] Nazer, L., Zatarah, R., Waldrip, S., Ke, J X C., Moukheiber, M., Khanna, A K., Hicklen, R S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P. (2023, June 22). Bias in artificial intelligence algorithms and recommendations for mitigation. Public Library of Science, 2(6), e0000278-e0000278. https://doi.org/10.1371/journal.pdig.0000278

[55] Ochella, S., & Shafiee, M. (2021, February 1). Performance Metrics for Artificial Intelligence (AI) Algorithms Adopted in Prognostics and Health Management (PHM) of Mechanical Systems. IOP Publishing, 1828(1), 012005-012005. https://doi.org/10.1088/1742-6596/1828/1/012005

[56] Petković, D. (2023, January 30). It is Not "Accuracy vs. Explainability"—We Need Both for Trustworthy AI Systems. Institute of Electrical and Electronics Engineers, 4(1), 46-53. https://doi.org/10.1109/tts.2023.3239921

[57] Prakash, N., & Mathewson, K W. (2020, January 1). Conceptualization and Framework of Hybrid Intelligence Systems. Cornell University. https://doi.org/10.48550/arXiv.2012.

[58] Prakash, N., & Mathewson, K W. (2020, January 1). Conceptualization and Framework of Hybrid Intelligence Systems. Cornell University. https://doi.org/10.48550/arxiv.2012.06161

[59] Qi, P., Chiaro, D., Guzzo, A., Ianni, M., Fortino, G., & Piccialli, F. (2023, September 11). Model aggregation techniques in federated learning: A comprehensive survey. Elsevier BV, 150, 272-293. https://doi.org/10.1016/j.future.2023.09.008

[60] Rastogi, C., Leqi, L., Holstein, K., & Heidari, H. (2022, April 21). A Unifying Framework for Combining Complementary Strengths of Humans and ML toward Better Predictive Decision-Making. https://export.arxiv.org/pdf/2204.10806v2.pdf

[61] Ribeiro, J L P., Cardoso, L., Silva, R., Cirilo, V., Carneiro, N., & Alves, R. (2022, January 1). Explanations Based on Item Response Theory (eXirt): A Model-Specific Method to Explain Tree-Ensemble Model in Trust Perspective. Cornell University. https://doi.org/10.48550/arXiv.2210.

[62] Rossi, D., & Zhang, L. (2022, April 25). Landing AI on Networks: An Equipment Vendor Viewpoint on Autonomous Driving Networks. Institute of Electrical and Electronics Engineers, 19(3), 3670-3684. https://doi.org/10.1109/tnsm.2022.3169988

[63] Sadeghi, Z., Alizadehsani, R., Çifçi, M A., Kausar, S., Rehman, R., Mahanta, P., Bora, P K., Almasri, A., Alkhawaldeh, R S., Hussain, S., Alataş, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S., & Pardalo, P M. (2023, January 1). A Brief Review of Explainable Artificial Intelligence in Healthcare. https://doi.org/10.2139/ssrn.4600029

[64] Samek, W., & Müller, K. (2019, January 1). Towards Explainable Artificial Intelligence. Springer Science+Business Media, 5-22. https://doi.org/10.1007/978-3-030-28954-6_1

[65] Samek, W., Wiegand, T., & Müller, K. (2017, January 1). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. Cornell University. https://doi.org/10.48550/arxiv.1708.08296

[66] Scholtz, J. (2006, January 28). Metrics for evaluating human information interaction systems. Oxford University Press, 18(4), 507-527. https://doi.org/10.1016/j.intcom.2005.10.004

[67] Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022, March 15). Towards a standard for identifying and managing bias in artificial intelligence. https://doi.org/10.6028/nist.sp.1270

[68] Shams, R A., Zowghi, D., & Bano, M. (2023, January 1). Challenges and Solutions in AI for All. Cornell University. https://doi.org/10.48550/arxiv.2307.10600

[69] Sherson, J., Rabecq, B., Dellermann, D., & Rafner, J. (2023, June 22). A Multi-Dimensional Development and Deployment Framework for Hybrid Intelligence. https://doi.org/10.3233/faia230119

[70] Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020, January 1). Explainable deep learning models in medical image analysis. Cornell University. https://doi.org/10.48550/arxiv.2005.13799

[71] Sivaraman, V., Bukowski, L A., Levin, J., Kahn, J M., & Perer, A. (2023, April 19). Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. https://doi.org/10.1145/3544548.3581075

[72] Swamy, V., Radmehr, B., Krčo, N., Marras, M., & Käser, T. (2022, July 18). Evaluating the Explainers: Black-Box Explainable Machine Learning for Student Success Prediction in MOOCs. European Organization for Nuclear Research. https://doi.org/10.5281/zenodo.6852964

[73] Tsaih, R., Chang, H., Hsu, C., & Yen, D C. (2023, February 22). The AI Tech-Stack Model. Association for Computing Machinery, 66(3), 69-77. https://doi.org/10.1145/3568026

[74] Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S D., Tegmark, M., & Nerini, F F. (2020, January 13). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Portfolio, 11(1). https://doi.org/10.1038/s41467-019-14108-y

[75] Welsch, G., & Kowalczyk, P. (2023, January 1). Designing Explainable Predictive Machine Learning Artifacts: Methodology and Practical Demonstration. Cornell University. https://doi.org/10.48550/arXiv.2306.

[76] Weng, Y., Wu, J., Kelly, T., & Johnson, W. (2024, September 19). Comprehensive Overview of Artificial Intelligence Applications in Modern Industries. Cornell University. https://doi.org/10.48550/arxiv.2409.13059

[77] Wondimu, N A., Buche, C., & Visser, U. (2022, January 1). Interactive Machine Learning: A State-of-the-Art Review. Cornell University. https://doi.org/10.48550/arxiv.2207.06196

[78] Wu, C., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F A., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Мельников, A C., Candido, S., Brooks, D J., Chauhan, G., Lee, B., Lee, H S., . . . Hazelwood, K. (2021, January 1). Sustainable AI: Environmental Implications, Challenges and Opportunities. Cornell University. https://doi.org/10.48550/arxiv.2111.00364

[79] Yeo, W J., Heever, W V D., Mao, R., Cambria, E., Satapathy, R., & Mengaldo, G. (2023, January 1). A Comprehensive Review on Financial Explainable AI. Cornell University. https://doi.org/10.48550/arxiv.2309.11960

[80] Zhang, T., Hemmatpour, M., Mishra, S., Linguaglossa, L., Zhang, D., Chen, C S., Mellia, M., & Aghasaryan, A. (2023, January 1). Operationalizing AI in Future Networks: A Bird's Eye View from the System Perspective. Cornell University. https://doi.org/10.48550/arXiv.2303.