applied sciences

MDPI

Review

Making Sense of Machine Learning: A Review of Interpretation Techniques and Their Applications

Ainura Tursunalieva, David L. J. Alexander, Rob Dunne, Jiaming Li, Luis Riera and Yanchang Zhao *

Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, ACT 2601, Australia; ainura.tursunalieva@data61.csiro.au (A.T.); david.alexander@data61.csiro.au (D.L.J.A.); rob.dunne@data61.csiro.au (R.D.); jiaming.li@data61.csiro.au (J.L.); luis.rieragarcia@data61.csiro.au (L.R.)
* Correspondence: yanchang.zhao@csiro.au

Abstract: Transparency in AI models is essential for promoting human–AI collaboration and ensuring regulatory compliance. However, interpreting these models is a complex process influenced by various methods and datasets. This study presents a comprehensive overview of foundational interpretation techniques, meticulously referencing the original authors and emphasizing their pivotal contributions. Recognizing the seminal work of these pioneers is imperative for contextualizing the evolutionary trajectory of interpretation in the field of AI. Furthermore, this research offers a retrospective analysis of interpretation techniques, critically evaluating their inherent strengths and limitations. We categorize these techniques into model-based, representation-based, post hoc, and hybrid methods, delving into their diverse applications. Furthermore, we analyze publication trends over time to see how the adoption of advanced computational methods within various categories of interpretation techniques has shaped the development of AI interpretability over time. This analysis highlights a notable preference shift towards data-driven approaches in the field. Moreover, we consider crucial factors such as the suitability of these techniques for generating local or global insights and their compatibility with different data types, including images, text, and tabular data. This structured categorization serves as a guide for practitioners navigating the landscape of interpretation techniques in AI. In summary, this review not only synthesizes various interpretation techniques but also acknowledges the contributions of their original authors. By emphasizing the origins of these techniques, we aim to enhance AI model explainability and underscore the importance of recognizing biases, uncertainties, and limitations inherent in the methods and datasets. This approach promotes the ethical and practical use of interpretation insights, empowering AI practitioners, researchers, and professionals to make informed decisions when selecting techniques for responsible AI implementation in real-world scenarios.

Keywords: Artificial Intelligence; explainable AI; interpretable machine learning; interpretation techniques

Citation: Tursunalieva, A.; Alexander, D.L.J.; Dunne, R.; Li, J.; Riera, L.; Zhao, Y. Making Sense of Machine Learning: A Review of Interpretation Techniques and Their Applications. *Appl. Sci.* **2024**, *14*, 496. https://doi.org/10.3390/app14020496

Academic Editor: Mohamed Benbouzid

Received: 6 October 2023
Revised: 24 November 2023
Accepted: 25 November 2023
Published: 5 January 2024

1. Introduction

The growing demand for transparency and accountability in AI decision-making has made eXplainable Artificial Intelligence (XAI) an increasingly important area of research. Although interpretation techniques are not essential for the operation of AI models, they can be viewed as supplements that introduce an extra layer of transparency and explainability. This supplementary aspect could prove pivotal in ensuring the adoption and success of AI applications and the capacity of being more responsive to the needs of intended users.

In the context of the escalating significance of AI, Padovan et al. [1] propose that interpretability entails grasping the mechanics of a system without necessarily plunging into causality (i.e., comprehending how it operates), whereas explainability involves deciphering the reasons (i.e., understanding why it operates). These interpretations heighten the comprehensibility, appropriateness, and usability of AI-generated guidance for its

intended users. Consequently, techniques for interpretation have emerged, aiming to facilitate human comprehension of the underlying rationale behind a model's conclusions. This becomes notably vital in high-stakes domains such as healthcare, criminal justice, environment, and finance. Instances where users fail to apprehend the logic underpinning a model's suggestions can undermine trust in its decisions, potentially leading to user exasperation, limited adoption, and even ethical and legal dilemmas.

From a technical perspective, integrating interpretability into a model serves several essential purposes. It aids developers in identifying and rectifying biases within the training dataset, allowing for targeted corrective measures. Moreover, interpretability plays a pivotal role in fortifying model robustness by shedding light on potential adversarial perturbations capable of altering model outputs. Additionally, it ensures that the model relies solely on relevant variables to derive its outputs, as underscored by Barredo et al. [2] in their work on explainable AI. In a retrospective analysis, Gunning et al. [3] delve into the objectives, accomplishments, challenges, and insights of the XAI program.

In the context of XAI, a variety of interpretation techniques emerge, encompassing representation-based, model-based, post hoc, and hybrid approaches. These strategic approaches are skilfully developed to shed light on the decision-making processes of AI systems, ultimately contributing to improved human comprehension and trust in their outputs. This advancement holds immense promise for enhancing the effectiveness and dependability of AI applications across diverse sectors like healthcare, finance, and legal systems. Consequently, XAI research aims to leverage AI's capabilities while concurrently furnishing comprehensible explanations for its outcomes. These explanations might pertain to the data or systems themselves, or the decision-making mechanisms grounded in fitted models. This review conducts an in-depth exploration of the principal techniques within the XAI domain, while also addressing the challenges and opportunities inherent in their implementation.

1.1. Motivation and Purpose of the Review

The motivation behind our review stems from a nuanced understanding of the existing literature surrounding interpretable ML and AI. While there are commendable reviews addressing interpretability within specific domains, such as healthcare [4], and in specific areas of AI such as deep neural networks [5], and also in general machine learning [6,7], our review aims to contribute a distinctive perspective.

While acknowledging the shared goal among these reviews—to pursue interpretability in AI and ML—our work distinguishes itself through a comprehensive exploration of foundational interpretation techniques in AI. Moreover, it contributes to the scholarly discourse by providing a historical context, delving into publication trends that have shaped the landscape of interpretable AI over time. In essence, our review offers a holistic understanding that goes beyond domain-specific nuances, revealing the broader evolution and foundational principles underpinning interpretability in the realm of AI and ML.

More specifically, our review introduces a structured categorization of interpretation techniques, encompassing model-based, representation-based, post hoc, and hybrid methods. The retrospective analysis and publication trends featured in our review reveal the evolution of different types of interpretations in AI, showcasing a preference shift towards data-driven approaches. The inclusion of a retrospective analysis allows for a deeper understanding of the evolution of interpretation in AI, and the analysis of publication trends sheds light on the evolving landscape of interpretability research.

Overall, our review not only fills specific gaps in the existing literature but also enhances it through a unique consolidation of structured categorization, retrospective analysis, and publication trends, thereby shedding light on the advancements, challenges, and emerging trends in this dynamic field of explainable AI (XAI) interpretation techniques.

This comprehensive review caters to researchers, practitioners, and policymakers seeking a holistic understanding of the interpretability landscape in ML. It specifically targets professionals and researchers in the fields of AI, ML, data science, and related

disciplines, offering an in-depth exploration of the development and application of XAI techniques. Moreover, it provides valuable insights into the historical context and evolution of these methodologies, making it a pertinent resource for policymakers and individuals concerned with the ethical and regulatory implications of AI.

By emphasizing the pivotal contributions of the original authors, this review aims to provide a thorough understanding of the foundational principles of interpretability and transparency in AI systems. Its primary objective is to facilitate a deeper comprehension of the historical significance and progression of interpretability techniques within the field.

This review not only offers a comprehensive overview of foundational interpretability techniques, highlighting the key contributions made by prominent and pioneering authors, but also seeks to address the following critical questions, including a research question centered on trend analysis:

1. What are the foundational interpretation techniques in the AI domain, and how have they influenced the development of contemporary interpretation methodologies?
2. How do diverse interpretation techniques, including model-based, representation-based, post hoc, and hybrid methods, address the challenge of explainability in AI models, thereby fostering effective human–AI collaboration and regulatory compliance?
3. What are the inherent strengths and limitations associated with the utilization of various interpretation techniques in real-world applications, particularly across different domains such as images, text, and tabular data?
4. How has the adoption of advanced computational methods within various categories of interpretation techniques influenced the evolving landscape of AI interpretability over time?

By providing comprehensive insights into these critical questions, along with a dedicated focus on the trend analysis, this review serves as an invaluable resource for individuals seeking a holistic understanding of the complexities and implications associated with interpretability in AI systems.

1.2. Methodology of the Review

To attain a comprehensive understanding of explainable AI techniques, our methodological framework adopts a synthesis approach. Leveraging Google Scholar, we systematically retrieved titles using carefully chosen keywords, specifically "explainable AI model-based techniques", "explainable AI representation-based techniques", "post hoc explainable AI interpretation techniques", and "Hybrid explainable AI interpretation techniques." Following this, we meticulously shortlisted titles by incorporating category-specific keywords, resulting in the identification of 15 pertinent documents sourced from academic journals, conference proceedings, and various Web of Science indexed platforms. The exploration of each technique type involved an inclusive process, encompassing surveys and comparative studies within academic literature. To refine our understanding, we compiled an initial list of interpretation techniques, further enriched by references derived from the short-listed titles. This systematic and inclusive approach serves as the foundation of our methodology, ensuring both relevance and document quality, and facilitating a thorough exploration of the dynamic landscape of explainable AI techniques.

We structured the review by methodically segmenting it into distinct components for comprehensive analysis. Emphasizing historical context, we meticulously traced the evolution of each method by examining references and identifying the pioneering researchers responsible for each approach's inception. Notably, our review only considered publications up to 20 March 2023.

This review is organized as follows: Section 2 provides a foundation for understanding the historical progression of AI. In Section 3, we review the origins, basic principles, and suitability of existing Machine Learning (ML) interpretation techniques for different types of data. The following two sections respectively analyse trends in annual publications on different types of ML interpretation techniques; and cloud-based implementation strategies for XAI. Section 6 explores examples and use cases of XAI in various fields, including

healthcare, finance, environment, criminal justice, and autonomous systems. Section 7 discusses the challenges and limitations of XAI. Finally, in Section 8, we conclude with a summary of the findings and a discussion of future research directions.

2. Unveiling the Evolution of AI: From Symbol Manipulation to Deep Learning Networks

As we delve into the fascinating journey of AI's evolution, this section unveils two pivotal narratives that have shaped the landscape of AI: 'from symbol manipulation to conversational agents' and 'from brain cell models to deep learning networks.' These narratives not only highlight the technological milestones but also lay the foundation for the critical exploration of contemporary AI's interpretability and its implications for transparency and reliability.

2.1. From Symbol Manipulation to Conversational Agents

In the early days of AI, the manipulation of symbols using predicates and logical propositions formed the basis of the technology. Progress in AI was a result of progress in manipulating symbols, which also pushed the boundaries of computing languages [8,9].

An early and noteworthy instance of a conversational agent is ELIZA, which originated in the mid-1960s [10]. This agent employed a domain-specific script to establish a conversational style, such as that of a "DOCTOR", and engaged users in a "therapist" conversation by mostly restating their remarks as questions, thus prompting the user to provide further information and continue an ongoing dialogue. Some users reported benefiting from these therapy sessions and sought continued access to the system, a phenomenon known as "the ELIZA effect". Considering that this occurs despite the strong reliance of each question on the preceding answer and the transparent manner of restatement, it becomes apparent that more sophisticated platforms like ChatGPT [11] are poised to gain widespread acceptance as human conversational partners among a substantial portion of individuals.

Integrating sophisticated symbolic manipulation capabilities, the Lisp family of programming languages has established its prominence. However, it is intriguing to note that ELIZA, despite this trend, was crafted using a distinct language called MAD (Michigan Algorithm Decoder) [12]. This language was enriched with the integration of an accompanying package named SLIP (Symmetric List Processor), effectively infusing Lisp-like features into its functionality. Such a departure from Lisp's convention highlights the innovative approach undertaken in ELIZA's development.

In this unconventional choice of language, ELIZA's creators navigated their way towards designing a conversational agent capable of interactive dialogue. The augmentation of MAD with SLIP not only exemplified a creative adaptation but also paved the way for a system that initiated the journey toward explainable AI and human-computer interaction. This strategic departure from the established norm enabled ELIZA to simulate therapeutic conversations effectively, setting the stage for modern developments in the domain of interpretation techniques for AI systems.

2.2. From Brain Cell Models to Deep Learning Networks

The development of AI has not been a linear climb up a ladder of success. There have been several "AI winters" when interest has faded. The first occurred in the 1970s due to the failure to meet the unrealistic expectations that had been raised for the field. The second AI winter occurred in the late 1980s and early 1990s, and was caused in part by the fact that symbolic manipulation systems, which were based on the Lisp programming language, were not able to handle the complexity of real-world problems. These systems were limited by their reliance on handcrafted rules and lack of ability to learn from data.

Parallel to the symbolic computational stream, there was a stream of research based on computation models of brain cells and a model of learning (Hebbian learning). In 1943 McCulloch and Pitts introduced the "perceptron", an idealized representation of a neuron

cell that sums its inputs and gives an output signal when the sum exceeds a threshold. The first implementation was a machine built by Frank Rosenblatt in 1958 at the Cornell Aeronautical Laboratory.

In 1969 Marvin Minsky and Seymour Papert showed, in their book *Perceptrons*, that it was impossible for these classes of network to learn an XOR function, leading to an AI winter for the computational approach. However two advances (rediscovered several times) overcame this limitation leading to the "multilayer perceptron" (MLP) models (variously called neural networks and deep learning). The MLP model consists of layers of "neurons" fully connected to the neurons in the next layer, with the possibility of some connections skipping layers.

These overcame previous constraints through the use of continuous activation functions, instead of the perceptrons step function, and the "backpropagation" of error terms through the network. Backpropogation can be seen to be the chain rule for differentiation, coupled with the fact that the needed information is always available at each neuron via the established connections. Cybenko's theorem [13] shows the universal function approximation capability of MLPs, establishing their significance as general regression and classification tools.

The MLP model was gradually expanded to architectures aimed at particular problem areas. The convolutional neural network (CNN), for example, was particularly targeted at image problems. There has been a transformative impact of Deep Learning (DL) in text, speech, and image applications. We can now add protein folding to that list. AlphaFold can predict the structure of a protein to an accuracy comparable with the results of laborious experimental determinations and is considered a correct solution that scientists can rely on with some confidence [14].

The evolution depicted in this section embodies AI's transformative narrative—from its rule-driven inception to the intricate landscapes of DL. While early AI systems thrived on explicit rules, the difficulty of interpretation for contemporary AI raises multifaceted ethical and practical concerns. In response, techniques such as model distillation, feature visualization, and attention mechanisms are being explored as promising avenues for enhancing transparency and reliability. By embracing these methods, AI practitioners may usher in a new era of systems that inspire trust and understanding, bridging the gap between human users and the intricacies of machine intelligence.

3. Techniques for Explainable AI

Explainable Artificial Intelligence (XAI) employs a range of fundamental techniques that can be systematically classified into distinct categories, namely, model-based, representation-based, and post hoc interpretation techniques. Additionally, innovative hybrid approaches amalgamate these methods, offering a comprehensive understanding of complex AI systems. This section provides an in-depth exploration of the origins and fundamental principles underpinning these techniques, highlighting their respective merits. The accompanying Figure 1 illustrates the core principles underlying each category of interpretable techniques. Notable interpretation techniques are conveniently compiled in Table 1.

In the following section, our focus shifts to model-based interpretation techniques, where we explore their origins, fundamental principles, and respective merits.

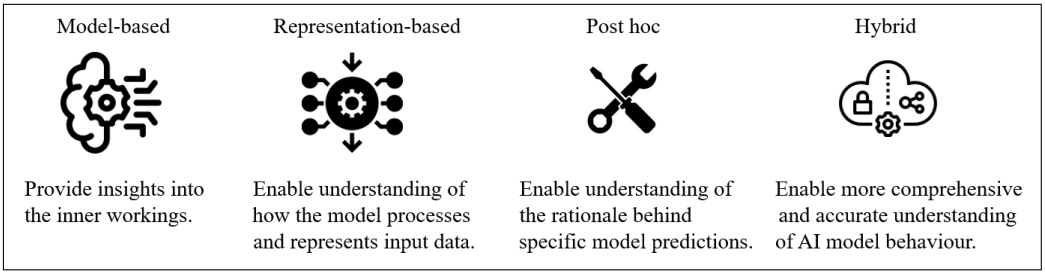

Figure 1. Core principles for interpretable techniques. Icons source: Flaticon. (2023). [Automation, Data Modelling, Settings, Hybrid]. Retrieved from https://www.flaticon.com/ (accessed on 1 November 2023).

Table 1. Overview of Interpretation Techniques for Explainable Artificial Intelligence (XAI).

Technique	Type	Interpretability	Category	Data Type *
SHAP [15]	Agnostic	Local	Post hoc	STR, USTR, TS
LIME [16]	Agnostic	Local	Post hoc	STR, USTR, TS
Linear Models [17]	Agnostic	Local	Model-based	STR
Rule Extraction [18]	Agnostic	Local	Model-based	STR, Text
Decision Trees [19]	Agnostic	Global/Local	Model-based	STR
Anchors [20]	Agnostic	Local	Post hoc	STR, Text, IMG
Kernel SHAP [15]	Agnostic	Global/Local	Post hoc	STR, Text, IMG
Tree SHAP [21]	Tree-based	Global/Local	Post hoc	STR, Text, IMG
Deep SHAP [15]	DL	Global/Local	Post hoc	STR, Text, IMG
DeepLIFT [22]	DL	Global/Local	Post hoc	Image, Text
Grad-CAM [23]	CNN	Local	Post hoc	IMG

* STR—structured data; USTR—unstructured data such as text, IMG—Image, Video, Audio; TS—time series data

3.1. Model-Based Interpretation Techniques

The realm of model-based interpretation techniques serves the fundamental purpose of delving into the internal processes of a model and offering insights into how it operates and arrives at specific predictions. The central premise of these techniques revolves around leveraging the model's structural architecture and parameter configuration to glean insights into its behavior and predictive outputs. This holistic comprehension not only affords users the ability to scrutinize the robustness of the resultant solutions but also bolsters the likelihood of fostering user trust in the yielded outcomes [24].

Rule-based and Bayesian models, decision trees, and random forests, as well as linear models, stand as prominent exemplars of model-based interpretation techniques. Their deployment has effectively contributed to elevating the level of explainability within AI models, paving the way for a more comprehensive understanding of complex phenomena.

3.1.1. Rule-Based Models

McCarthy J. [18] made groundbreaking contributions to the field of AI that significantly influenced the development of rule-based models. Through his visionary work on time-sharing systems, McCarthy laid the groundwork for a more interactive computing environment, allowing researchers to engage in real-time experimentation and iterative development of rule-based reasoning and symbolic computation. Moreover, McCarthy's invention of the LISP programming language revolutionized the field of AI by enabling the manipulation of symbolic expressions and the implementation of complex logical rules. LISP's unique features, including support for recursion and dynamic typing, provided researchers with a powerful tool to explore and refine rule-based systems. McCarthy's pioneering efforts in promoting interactive computing and advancing symbolic computation were pivotal in shaping the trajectory of rule-based modeling and its subsequent applications in various domains of AI research and development.

Rule-based models employ explicit sets of rules to analyze input data and facilitate decision-making. These rules often take the form of "if-then" statements, wherein the

"if" segment outlines conditions that must be met, and the corresponding "then" segment prescribes the appropriate action upon satisfaction of those conditions. In essence, rule-based models offer transparent and human-interpretable rules that guide the decision-making process of AI systems.

Such rules frequently originate from domain experts and require a well-defined problem context, featuring clear cause-and-effect relationships between inputs and outputs. An illustrative example is the domain of expert systems, where a set of rules guides decisions and dispenses advice within a specific field. These models also extend to decision support systems, aiding users in decision-making, and extend even to applications in natural language processing, where rules enable the parsing and comprehension of natural language text.

While rule-based models excel in providing transparency and comprehensibility, their efficacy might wane in contexts requiring heightened adaptability or when intricate problem domains surpass human experts' capacity to formulate explicit rules. In such scenarios, alternative ML techniques, like neural networks or decision trees, may prove to be more suitable.

3.1.2. Bayesian Rule Lists (BRLs)

Wang et al. [25] were the first to propose a method for learning Bayesian rule sets. This method uses Bayesian model averaging to learn a set of rules that jointly maximize the posterior probability of the class labels. They demonstrated the effectiveness of their method on classification tasks such as predicting adverse drug events and detecting fraudulent insurance claims.

BRLs are a type of ML model that combines the interpretability of rule-based models with the probabilistic reasoning of Bayesian models. The basic idea behind BRLs is to use a set of logical rules for making predictions and to allow for uncertainty in the rules and their weights. The objective of BRLs is to learn a list of if-then rules that define decision boundaries between different classes or regression targets. Each rule in the list consists of a condition and a class or regression target. The conditions are defined in terms of features or attributes of the data, and the class or regression target is assigned a probability based on the conditions of the rules that are satisfied. The weights and probabilities assigned to each rule can be learned from training data or based on expert knowledge. BRLs are often used in natural language processing and other areas where the decision-making process involves multiple competing rules or factors. For example, Wang et al. [26] use BRLs for interpretable classifications of consumer behaviors and Letham et al. [27] apply BRLs to improve the explainability for stroke predictions.

3.1.3. Decision Trees (DTs)

The idea of DTs was introduced by Quinlan, J. R. [19]. His team designed Iterative Dichotomiser 3 (ID3) algorithm that automatically learned trees from data. The ID3 algorithm uses a top-down approach to start from a single node and split the data recursively into smaller subsets using the most informative feature. This splitting continues until all of the instances in a subset belong to the same class or when no more splits are possible. Overall, the DT technique performs well on tabular-style datasets with individually meaningful features and with no strong multi-scale temporal or spatial structures. DTs are considered to be explainable due to their ability to represent the decision-making process of a model in a clear and transparent way [28]. They are particularly useful in applications where transparency and interpretability are essential for ensuring the fairness and trustworthiness of AI-based decision-making systems.

In a similar vein, Westberg et al. [29] explore the utility of decision trees in medical diagnosis, where the trees leverage easily interpretable variables to make decisions, offering straightforward explanations for their outcomes. This approach has potential applicability in other domains that prioritize transparency in decision-making.

However, if the complexity of the relationships in a given dataset is high then the interpretability of a DT decreases. If the relationships between the input features and the output variable are highly nonlinear or complex, a DT may struggle to capture them accurately using a simple set of rules. In such cases, the decision tree may require a larger number of nodes and branches, which can make it difficult to interpret and understand. To overcome this limitation, Norouzi et al. proposed a stochastic gradient descent approach to enable efficient optimization for large-scale datasets [30]. To further improve the model's explainability and generate additional insights, Mishra provides a practical guide for incorporating Python-based explainable AI libraries such as SHAP, LIME, and Skope-Rules into a DT [31].

Another challenge faced by DT algorithms is the use of local optimization functions, meaning that inaccurate splits at the top of the tree may result in poor predictions. Therefore, various techniques have been developed to simplify and optimize DT, such as pruning, ensemble methods, and regularization. These techniques can help to reduce the complexity of the tree and improve its interpretability, while still capturing the essential relationships in the data. Gradient boosting can also be used to improve the interpretability of DT. By combining multiple decision trees into an ensemble model, gradient boosting can create more accurate and robust models that are less prone to overfitting. Costa [32] provided a comprehensive review of the recent advances in the field of optimal DT and their interpretability.

3.1.4. Random Forests (RFs)

Breiman [33] was the first to extend the DT algorithm and introduce RFs, which build a large number of decision trees using bootstrapped samples of the data and random subsets of the features. The final prediction is made by aggregating the predictions of all the trees in the forest. Because the structure of an RF is more complex than the structure of a DT, they can be more challenging to understand and interpret. Moreover, as the complexity of relationships among the variables increases, the number of trees required to achieve good accuracy also increases and with more trees, it becomes challenging to identify important variables and evaluate their contribution to the predictions.

RFs have several advantages over DTs, including reduced over-fitting, improved accuracy, and the ability to handle high-dimensional data. RFs, like DTs, also provide local interpretability by explaining the model's predictions for individual instances. To improve the interpretability of tree-based models for complex tabular datasets, Lundberg et al. [28] proposed computing the SHAP interaction values for trees. These SHAP interaction values capture not only individual predictions, but also the interaction effects among all pairs of those predictions.

3.2. Linear and Logistic Regression

Francis Galton [17] created regression analysis as a method to forecast one continuous variable based on another. He hypothesized that the connection between two variables could be represented by a straight line. He collected data on the heights of parents and their children and to test this idea, he plotted the data on a graph. He observed that the children's heights tended to deviate less from the average height of all children than the heights of their parents. Galton called a straight line drawn through the data points the "regression line", and used it to predict the height of a child based on the height of their parents. Berkson, J. [34] invented logistic regression by using the logistic function for statistical modeling. Both linear and logistic Regression are popular ML algorithms with a wide range of applications in AI. Linear regression predicts a continuous numerical output variable, while logistic regression predicts a categorical one.

It is widely accepted that linear and logistic regression models have a clear and interpretable structure. These properties are valuable for designing AI-based decision-making systems that are trustworthy and can be used in explainable AI.

Schneider et al. [35] provided a thorough overview of linear regression analysis and its strengths and limitations, discussed common interpretation mistakes and emphasized the necessary aspects of reporting the results to ensure the validity and reliability of the research findings.

Peng et al. [36] provided a clear and accessible introduction to logistic regression analysis and emphasized the importance of transparent reporting and careful interpretation of the results. These considerations are particularly important in the context of explainable AI, where the interpretability of models is essential for ensuring their trustworthiness and reliability.

To improve the accuracy, stability, and generalization of regression models, they are frequently used in ensemble methods. Powerful ensemble models can be developed by combining bagging, boosting, stacking, or random forests with regression models [37,38].

Global Sensitivity Analysis (GSA)

GSA has become integral in understanding complex ML models, with Saltelli et al. [39] highlighting its impact. GSA enables the quantification of input parameter importance, aiding in model optimization and critical feature identification and enhancing interpretability across domains. Additionally, GSA facilitates sensitivity assessment in hybrid models, evaluating constituent model contributions and ensuring robust, informed decision-making for improved performance. Furthermore, in engineering and risk assessment, GSA integrated into the GSAS (Global sensitivity analysis-enhanced surrogate) method allows for efficient reliability analysis, identifying critical parameters and enhancing understanding of system interactions and uncertainties, crucial for system design and maintenance [40]. The work of Friedman et al. [41] on gradient boosting machines (GBMs) influenced GSA application in ML. GBMs emphasized individual model contributions, aligning with GSA's focus on input impact evaluation. This integration enhanced model interpretability and robustness, particularly in ensemble-based learning techniques, contributing to improved model performance and understanding in the field of ML.

The pioneering work of Judea Pearl and his colleagues in developing the Bayesian network (BN) framework [42,43] has significantly shaped the landscape of probabilistic reasoning in ML. Pearl's research laid the groundwork for comprehending intricate probabilistic relationships among variables, enabling the representation of complex data structures in an intuitive manner. Emphasizing the utilization of graphical models to depict conditional dependencies, Pearl's contributions facilitated the creation of transparent and interpretable models, fostering a more profound understanding of the underlying data structure. BNs, conceptualized by Pearl, serve as a robust framework that seamlessly integrates prior knowledge and data-driven learning, allowing the incorporation of expert insights into the model development process. This emphasis on transparent modeling has had a transformative impact on the advancement of interpretable ML, leading to the creation of models that not only yield accurate predictions but also offer valuable insights into the reasoning behind these predictions. Pearl's significant contributions have paved the way for the integration of causal reasoning and probabilistic inference in interpretable ML, thus promoting the development of models that are both accurate and comprehensible across various applications.

Russell and Norvig [44] emphasize the pivotal role of BNs as essential tools for effectively representing and reasoning with uncertainty in the field of AI. The book highlights how BN offer a formal and intuitive representation of uncertainty by capturing the probabilistic relationships among variables. It accentuates the graphical depiction of conditional dependencies among variables, enabling the construction of intricate probabilistic models in a more manageable and interpretable manner, thereby facilitating effective decision-making processes in the presence of uncertainty.

3.3. Representation-Based Interpretation Techniques

3.3.1. Sparse Linear and Shallow Decision Tree ML Models

These are known for their ease of interpretation, requiring only minimal human effort. However, the pursuit of greater predictive accuracy in ML has led to the adoption of deeper and more complex models, albeit at a cost. These sophisticated models often involve millions of interconnected deep neural networks, making them exceptionally challenging to interpret. They appear as black boxes to human observers, eroding trust in their predictions. While there are techniques available for visually inspecting the first layer of a neural network, understanding the intricate workings of deep learning models remains a significant challenge. In recent years, there has been a growing research focus on developing methods to gain deeper insights into these models. Among these techniques, saliency maps, as introduced by Itti et al. [45], and activation maximization [46] have emerged as common tools for interpreting image classification models. Saliency maps have also found utility in text classification tasks. The interpretability of attention mechanisms also sparks considerable debate and ongoing research efforts.

3.3.2. Saliency Maps (SMs)

Saliency maps are visual attention explanatory techniques, motivated by the visual functionality of the early primate visual system [45]. The idea of saliency maps in AI dates back to the early 1990s and aims to illustrate the relative importance of each element among its neighbours by measuring its contribution to a particular class. Morch et al. [47] proposed "Visualisation of Neural Networks Using Saliency Maps" to understand and visualize the non-linearities in feed-forward neural networks. Simonyan et al. [48] expanded the idea to be used by classification ConvNets models. In the article "Graying the black box", Zahavy et al. [49] implemented saliency maps to explain deep reinforcement learning models by mapping the entire model as a simplified Markov decision process (MDP). Kindermans et al. [50] argue that this technique "lacks reliability when the explanation is sensitive to factors that do not contribute to the model prediction".

3.3.3. Activation Maximization (AM)

This is a technique used to visualize how specific unit neuron i, at a given hidden layer j, is activated by an input x of a given class to produce an output image $h_{ij}(\theta, x)$, where θ denotes the neural weights and biases as exhibited by Equation (1). The idea was first published in 2009 by Erhan et al. [46], enabling us to shine a light inside the model black box. The logic of this idea is that while training a model, the objective is to adjust its weight to minimize the input to the output losses. On the other hand, when using activation maximization, the model weights and a chosen output of a given class are kept constant, and the input is modified to activate specific neurons that match the selected output.

$$X^* = \arg \max_{x \text{ s.t. } ||X||=\rho} h_{ij}(\theta, x). \quad (1)$$

3.3.4. Attention Mechanisms (AtMes)

AtMes have garnered significant attention in the field of AI in recent years, particularly due to their remarkable contributions to natural language processing (NLP). These mechanisms are designed to identify and highlight important information that models consider crucial for various tasks, which in turn enhances the interpretability of these models. A notable feature of attention mechanisms is their ability to reveal the "reasoning" behind each output generated by the model, as discussed by Rigotti et al. [51]. The journey towards interpreting natural language inference (NLI) models using attention mechanisms was first embarked upon by Ghaeini et al. [52]. However, it is important to note that the debate surrounding the interpretability of attention mechanisms is far from settled [51]. Some researchers have asserted that models incorporating attention mechanisms lack interpretability [53,54]. Conversely, there are researchers who assert that they have developed

architectures that perform exceptionally well in various scenarios, pushing the boundaries of state-of-the-art performance [51,55,56]. This ongoing discourse reflects the dynamic nature of research in this field, with the quest for improved interpretability remaining a central focus.

3.4. Post Hoc Interpretation Techniques

Post hoc interpretation techniques generate explanations for models that are not inherently explainable, making it possible to understand more of the logic behind black-box or complex ML models. These techniques include LIME [16], SHAP [15], and Grad-CAM [23], which generate local explanations by highlighting the features that contribute most to a model's decision. Some methods, such as LIME and Kernel SHAP [15], are model-agnostic and can be used on any ML models, while others, such as Tree SHAP [21], Deep SHAP [15], DeepLIFT [22] and Grad-CAM [23], are model-specific by leveraging extra knowledge of the specific models.

Ribeiro et al. proposed Local interpretable model-agnostic explanations (LIMEs) to explain the predictions of any classifier by learning an interpretable model locally around the prediction [16]. To explain the prediction for an instance, a LIME generates new fake instances around it by sampling its neighborhood, applies the original complex or black-box model to those new instances to produce predictions, and trains an interpretable model (e.g., a linear model) that captures the behaviors of the complex model in that neighborhood. Based on that, a representative set of individual instances can be selected and explained to achieve a global understanding of the model. LIMEs are claimed to be capable of explaining the predictions of any classifier in an interpretable and faithful manner.

Lundberg et al. presented a unified framework named Shapley additive explanations (SHAPs) for interpreting model predictions [15]. Shapley values [57] are a concept for measuring each player's contribution to the game in cooperative game theory and are comparable to the importance of features in ML model prediction. The Shapley value of a feature is calculated as a weighted average of the model prediction differences between models trained with and without the feature under all possible combinations of the other features. Formally, it is calculated as

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\ (|F| - |S| - 1)!}{|F|!} \Big[f_{S'}(x_{S'}) - f_S(x_S) \Big], \tag{2}$$

where ϕ_i is the Shapley value of feature i, F is the set of all features, S is a subset of $F \setminus \{i\}$, x_S is the values of input features in set S, f_S is a model trained with feature set S, $f_S(x_S)$ is the model prediction with set S, and $S' = S \cup \{i\}$. The Shapley values are very challenging to calculate, in that there are $2^{|F|}$ differences to be computed. Lundberg et al. proposed Kernel SHAP and Deep SHAP to approximate them efficiently. Kernel SHAP is a model-agnostic method that can be used on any model, and Deep SHAP is a model-specific method for deep neural networks models. Lundberg et al. also proposed Tree SHAP [21] for estimating the SHAP values of tree ensembles, such as random forests and XGBoost, in polynomial time.

Framling et al. [58] compared LIMEs and SHAPs in generating explanations for a deep learning model trained on a medical dataset. The results showed that both techniques were useful in generating explanations for the model's decision-making process.

3.4.1. Deep Learning Important Features

DeepLIFT [22] is a recursive prediction explanation method for deep learning models. For a given output, it assigns importance scores to the inputs in a backpropagation-like way, where an importance signal is propagated from an output neuron backwards to the input in one pass. It features framing importance in terms of differences from a reference state and allowing separate consideration of the effects of positive and negative contributions, which

enables it to work even when gradients are discontinuous and discover dependencies possibly missed by other methods.

3.4.2. Gradient-Weighted Class Activation Mapping

Grad-CAM [23] was proposed by Selvaraju et al. for producing visual explanations for convolutional neural network (CNN)-based models. It uses the gradients of a target concept to produce a localization map and highlights the important regions in the image for the corresponding prediction. Furthermore, Woods et al. [59] discussed how Grad-CAM was used to generate visual explanations for the decision-making process of a deep learning model trained on image classification. The generated explanations helped to improve the model's robustness to adversarial attacks.

Altmann et al. [60] proposed permutation feature importance (PFI) to measure the decrease in model performance after shuffling a feature's values. A feature is important if shuffling its values increases the model error, because the model relies on the feature for the prediction. In contrast, a feature is not so important if shuffling its values has no or little effect on the model error.

3.5. Hybrid Interpretation Techniques

Kim et al. [61] proposed a mind-the-gap model(MGM) for interpretable feature extraction and selection. By building interpretability criteria directly into the model, it can optimize parameters related to interpretability and directly report a global set of distinguishable dimensions for further data exploration and hypothesis generation. In particular, it discovers a global set of distinguishable dimensions when clustering high dimensional data. It draws parameters for important dimensions from distributions with multiple modes, while unimportant ones are drawn from a uni-modal distribution. The MGM method is particularly used for data clustering, rather than for prediction. When clustering data, it is interesting and very useful to simultaneously produce a list of dimension sets that are important for distinguishing between the clusters. However, a limitation of the method is that it focuses on the clustering of binary data only.

Singh et al. [62] proposed a framework for evaluating counterfactual explanation methods based on explainability metrics and properties. These explanations work by finding the minimal set of changes to input that will change the model's output. Jentzsch et al. [63] discuss the use of conversational agents to provide explanations for the outputs of AI systems. The conversational agent engages in a dialogue with the user to understand their questions and provide appropriate explanations. The approach provides a user-centred and interactive way to understand AI decision-making.

Lundberg et al. [28] proposed a technique that combines local explanations generated by LIMEs with global explanations generated by SHAPs to provide a more comprehensive understanding of the decision-making process of decision trees. The proposed technique generates both local and global explanations, which could be used to understand the model's decision-making process at different levels of granularity.

Furthermore, contextual importance and utility and case-based reasoning post hoc explanation-by-example [64] are also techniques that provide local and global explanations.

4. Trends in Literature for Different Types of ML Interpretation Techniques

4.1. Analysis of Trends in Literature for Different Types of ML Interpretation Techniques

Figure 2 plots the numbers of publications on various types of interpretation technique each year. In analyzing the trend lines, it becomes evident that each category of interpretation techniques experienced varying degrees of interest and utilization over time. Notably, the hybrid and model-based categories exhibit an initial period of stronger growth, followed by a period of much more gradual growth, before a final spurt of interest around 2020. The post hoc and representation-based categories demonstrate fairly constant exponential growth, starting from a lower base. The growth of the representation-based category is strongest, though currently it still achieves the lowest number of publications annually.

The trends in the logarithm of Google Scholar references from 1986 to 2022 were notably influenced by pivotal developments in the respective techniques. For instance, the introduction of the mind-the-gap model (MGM) in 2015 and counterfactual explanations (CEs) in 2022 within the hybrid category contributed to its sustained upward trajectory. Similarly, the inclusion of decision trees (DTs) in 1986 and random forests in 2001 within the model-based category contributed to its consistent yet slightly slower growth. Within the post hoc category, the combined introduction of Bayesian rule lists (BRLs) and Shapley additive explanations (SHAPs) in 2017, permutation feature importance (PFI) in 2010, local interpretable model-agnostic explanations (LIMEs) in 2016, and deep learning important features (DeepLIFT) and gradient-weighted class activation mapping (Grad-CAM) in 2019 accounts for the observed fluctuations, denoting varying levels of interest over time. Moreover, the incorporation of saliency maps (SMs) in 1995, activation maximization (AM) in 2009, and attention mechanisms (AtMes) in 2018 within the representation-based category contributes to its modest yet gradually increasing trend, reflecting sustained but relatively moderate scholarly attention to these interpretation techniques over the years. In summary, the trends in the logarithm of Google Scholar references from 1986 to 2022 were significantly shaped by pivotal developments within each technique category, demonstrating the nuanced evolution of ML interpretation techniques. The introduction of specific techniques not only influenced their individual trajectories but also led to spill-over effects, reshaping the broader landscape of scholarly interest and research focus in the field. These insights underscore the intricate interplay between technological advancements and scholarly attention, emphasizing the dynamic nature of the ML interpretation landscape and the continued evolution of research priorities within the scholarly community. These findings provide valuable insights into the dynamic landscape of interpretation techniques and the evolving focus within the scholarly community.

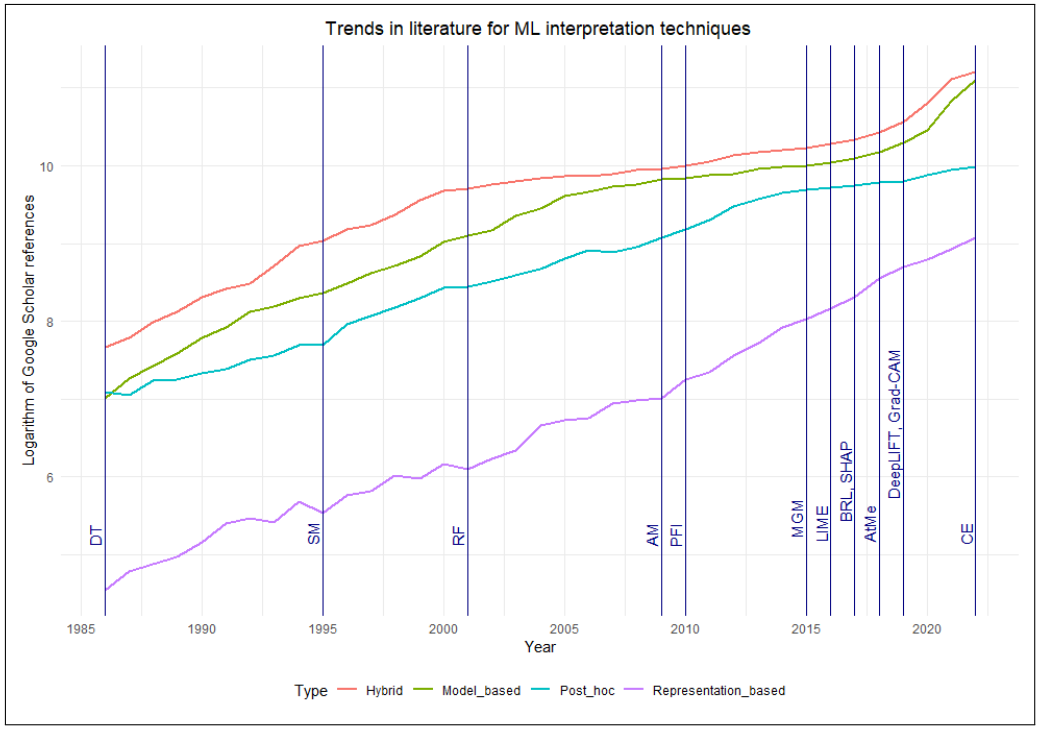

Figure 2. Numbers of annual publications on various interpretation techniques, with key milestones marked.

4.2. Emerging Techniques within the Field of ML Interpretation

As the complexity of neural network models in natural language processing (NLP) continues to grow, the development of novel methods for computing local explanations has emerged as a promising avenue for boosting interpretability within this domain. Concur-

rently, the integration of logic-based systems in contemporary AI research has demonstrated its significance in handling noisy and inconsistent data, emphasizing the critical role of rigorous methodologies in ensuring trustworthy and XAI.

Recent advancements in the field of ML interpretation, as highlighted by [65], underscore the promising potential of a novel method for computing local explanations in neural network models for NLP. This innovative approach demonstrates the critical importance of generating robust and optimal local explanations, particularly in complex neural network models, to bolster transparency and enhance interpretability. By emphasizing the significance of producing reliable explanations, this method paves the way for the development of robust interpretation techniques, especially in the domain of NLP.

Moreover, the integration of logic-based systems in contemporary AI research, as discussed by [66], has proven to be indispensable for establishing trustable XAI. By accounting for the inherent inconsistencies and noise present in real-world data, these logic-based methodologies provide a reliable framework for uncovering intricate relationships within complex datasets. This recognition of the indispensable role of logic-based approaches further accentuates the criticality of robust and rigorous frameworks in ensuring the reliability and interpretability of AI-driven systems.

The evolving landscape of ML interpretation techniques showcases the potential for developing robust and reliable methodologies to enhance the interpretability of complex neural network models. With the integration of innovative methods for computing local explanations and the recognition of the indispensable role of logic-based systems in handling real-world data, the field of AI interpretation is poised for significant advancements in the pursuit of transparent and trustworthy AI systems.

5. Implementation Strategies for Explainable AI Using Cloud-Based Solutions

While various techniques for explainable AI have been explored extensively, the successful implementation and deployment of these techniques in real-world applications require careful consideration of numerous factors. This section delves into the crucial aspects of implementing and deploying explainable AI solutions, focusing on the integration of different interpretation techniques with cloud-based solutions. By understanding the intricacies involved in the implementation process, organizations can leverage the power of explainable AI to enhance decision-making, foster trust, and ensure regulatory compliance within their respective domains.

The implementation of XAI is significantly facilitated by the integration of various cloud-based solutions, which have emerged as pivotal platforms for deploying and managing ML models. Notably, major cloud computing providers such as Amazon, Google, Microsoft, and IBM offer comprehensive toolsets and services tailored to address the growing demand for model transparency and interpretability. This section explores how leading cloud platforms, including Amazon SageMaker, Google Cloud AI Platform, Microsoft Azure Machine Learning, and IBM Watson Studio, have incorporated dedicated features and functionalities to enable users to comprehend and elucidate the decision-making processes of complex ML models. By leveraging these cloud-based solutions, organizations can not only enhance the interpretability of their AI models but also ensure transparency, fairness, and accountability in their AI-driven applications.

Amazon SageMaker[67] is Amazon's machine learning service that supports building, training and deploying ML models. It offers SageMaker Clarify for model explainability and interpretability, which detects and measures potential bias in ML models using a variety of metrics and enables ML developers to address potential bias and explain model predictions.

Google Cloud AI Platformoffers model explainability features to explain the predictions of ML models. It enables users to generate feature attributions, visualise explanations and understand why a particular prediction was made. Specifically, it provides Vertex Explainable AI [68] for feature-based and example-based explanations and a What-If Tool for visually investigating model behavior.

Microsoft Azure Machine Learning [69] provides an Azure ML Interpret package and a Responsible AI dashboard for interpreting and explaining ML models in terms of their fairness, reliability, safety, transparency and accountability. The Azure ML Interpret package provides various explainable AI methods such as SHAPs and LIMEs, as well as feature importance metrics. It is integrated with Azure Machine Learning Studio, and can also be accessed through the Python SDK for Azure Machine Learning.

IBM Watson Studio v8.0.0 [70] is IBM's software platform for data science, which empowers data scientists and analysts to build, run and manage AI models. It brings together open-source frameworks like PyTorch, TensorFlow and scikit-learn and supports programming languages such as Python, R and Scala. Particularly to interpret ML models and make them more transparent, it offers a toolkit named AI Explainability 360 (AIX360) [71], providing various algorithms that cover different dimensions of explanations and explainability metrics.

6. Applications: XAI's Impact across Diverse Domains

XAI is assuming growing significance in facilitating AI systems to elucidate their decision-making methodologies in a manner comprehensible to humans. This not only fosters enhanced trust but also empowers superior decision-making and more effective risk management. XAI's adoption spans diverse applications, and its role is especially pivotal in critical domains such as healthcare, finance, criminal justice, environment, education, and autonomous systems. The ensuing examples spotlight the varied applications and use cases of XAI.

6.1. Healthcare

XAI has emerged as a pivotal component within the healthcare sector, affording healthcare providers the ability to fathom the rationale behind AI algorithms' specific diagnostic conclusions or treatment recommendations. This fosters heightened assurance of the accuracy of these algorithms and consequently elevates the quality of decision-making and patient care. Consequently, XAI finds widespread utility in realms such as medical diagnosis, treatment recommendations, patient monitoring, and clinical decision support.

6.1.1. Example 1: Enhanced Breast Cancer Prognostication

An illustrative case highlighting the pivotal role of XAI in healthcare is the work of Mucaki et al. [72]. They developed an ML approach utilizing biochemically-inspired algorithms to predict outcomes related to hormone and chemotherapy for breast cancer patients. Their predictions are rooted in extensive data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) study. Through this XAI-based model, healthcare providers gain a deeper understanding of the factors contributing to these predictions, ultimately improving the precision and personalization of breast cancer treatment.

6.1.2. Example 2: Augmented Understanding of Breast Cancer Diagnoses

Another compelling application of XAI is demonstrated by Wang et al. [73], who deployed an XAI model to provide explanations for breast cancer diagnoses. This model not only aids medical practitioners in making more informed decisions but also empowers patients by offering them insights into the diagnostic process. By demystifying the AI-driven diagnosis, XAI fosters collaboration between healthcare providers and their patients, improving overall healthcare experiences.

6.1.3. Example 3: Informed Clinical Decisions for Sepsis Prediction

Rajkomar et al. [74] leveraged XAI to provide explanations for sepsis predictions, thereby enhancing clinical decision support. Sepsis is a life-threatening condition, and having transparent explanations for AI-driven predictions can be a critical factor in the timely intervention and treatment of patients. In this case, XAI acts as a vital aid for

healthcare professionals, enabling them to make more confident and effective decisions in emergency situations.

XAI has significantly reshaped healthcare by providing unprecedented insights into AI algorithms' decision-making processes. Its integration not only bolsters medical professionals' confidence in diagnostic and treatment recommendations but also ushers in a new era of informed medical interventions.

6.2. Finance

XAI has assumed a pivotal role within the finance sector, driven by imperatives of regulatory compliance and customer transparency. Through its capacity to offer lucid insights into decision-making processes, XAI equips financial institutions and investors to enhance their choices grounded in a comprehensive grasp of underlying variables. Its applications in finance span a wide spectrum, encompassing critical areas like risk management, trading strategies, and fraud detection. Notably, XAI models have found utility in revealing the risk profiles of financial products and investments, illuminating the intricate nuances that influence financial decision landscapes.

6.2.1. Example 1: Illuminating Credit Risk Assessment

The power of XAI models in finance is exemplified by a groundbreaking study conducted by Bussmann et al. [75], which introduces an XAI-based model tailored for credit risk assessment. This innovative model serves as a potent tool for financial institutions, significantly enhancing their risk management capabilities. Through transparent explanations of the factors influencing credit risk, financial professionals can make more informed decisions, reducing potential financial losses and fostering greater trust in their lending practices.

6.2.2. Example 2: Enhancing Stock Trading Strategies

XAI's influence extends into the realm of trading and investment. Kumar et al. [76] present a compelling perspective by proposing an explainable reinforcement learning approach designed for financial stock trading. This approach integrates SHAP, offering transparency and interpretability to decision-making processes. It unveils the intricate factors underpinning trading choices and portfolio management strategies. By doing so, it equips traders with a deeper understanding of their strategies, which can lead to more effective investment decisions.

6.2.3. Example 3: Bolstering Fraud Detection

XAI models have also demonstrated their effectiveness in enhancing fraud detection within the finance sector by introducing heightened transparency into the decision-making process. Ji [77] explores the domain of credit card fraud detection, where the deployment of explainable AI techniques, specifically SHAPs and LIMEs, emerges as a strategic avenue. This approach effectively unravels the contributing factors behind calculated fraud scores, providing clear explanations that empower financial institutions to proactively thwart fraudulent activities. By pinpointing potential fraudulent transactions and offering comprehensible rationales for their determinations, this research resonates as a proactive measure for fraud prevention.

In essence, the multifaceted applications of XAI within the finance sector, spanning from risk assessment and trading to fraud detection, embody the transformative potential of transparent decision-making. As XAI's prominence continues to grow, its role in fostering regulatory compliance, bolstering investor confidence, and refining financial decision landscapes stands resolute, promising a future where transparency and accountability converge to shape the financial industry.

6.3. Environment

XAI is leaving an indelible mark on the realm of environmental sciences by its use in models forecasting and tracking air quality, climate shifts, and water conditions. These

applications of XAI are facilitating a more comprehensive comprehension of the intricate determinants underlying environmental challenges. Below, we present noteworthy instances exemplifying the impact of XAI applications in the field of the environment.

One of the pivotal applications of XAI within the environmental domain is in air quality monitoring. XAI models have been ingeniously employed to furnish elucidations for air pollution predictions, thereby facilitating a profound comprehension of the determinants contributing to air quality degradation. A notable instance is the study conducted by Gu et al. [78], where an XAI model melding deep neural networks with a nonlinear auto regressive moving average model was proposed for air pollution prediction. This amalgamation yielded enhanced predictive accuracy and interpretability, exemplifying the potency of XAI in advancing understanding and forecasting.

Similarly, XAI's impact extends to climate change prediction, where it lends transparency to the intricacies of climate forecasts. By explicating the factors that influence global warming, XAI models contribute significantly to the comprehension of climate dynamics. For example, Straaten et al. [79] have harnessed an XAI model to shed light on the elevated summer temperatures experienced in Western and Central Europe.

In addition, the critical realm of water quality monitoring is profoundly influenced by XAI. XAI models here serve to expound upon water quality predictions, empowering a comprehensive grasp of the variables responsible for water pollution. Wu et al. [80], for instance, devised a water quality prediction model rooted in XAI principles. Applied to the Yellow River in China, this model unveiled insights into the intricate interplay among various water quality parameters. Likewise, Park et al. [81] employed an XGBoost model coupled with Shapley values (SHAPs) to predict water quality, select input variables, and provide model explanations.

In sum, XAI's ascendancy within the environmental domain, spanning air quality, climate change, and water quality, illuminates the trajectory toward a more transparent, comprehensible, and actionable understanding of complex environmental phenomena.

6.4. Criminal Justice

The utilization of XAI has notably showcased its potential to augment transparency, accountability, and equity within the criminal justice system. Extensive research has probed the deployment of XAI tools across diverse domains encompassing risk assessment, bail determinations, and sentencing protocols.

For instance, Dressel and Farid [82] undertook an exploration into XAI's role in predicting recidivism within the criminal justice sphere. Their work accentuated the paramount importance of employing XAI tools judiciously and transparently to ensure fairness in decision-making processes. Counterfactual fairness, a concept introduced by Kusner et al. [83], emerged as a pivotal consideration. They proposed a method to integrate this notion into XAI models, thereby mitigating biases that might otherwise perpetuate.

Moreover, Rudin and Radin [84] provided an incisive discourse on the indispensable value of incorporating XAI models across various applications, including the criminal justice domain. Their insights underscored the pivotal role of XAI in averting the potential pitfalls of biased decisions.

In conclusion, the advent of XAI as a pivotal tool within the criminal justice domain manifests its potency in fostering fair, informed, and unbiased decision-making processes, holding the potential to reshape the landscape of judicial systems for the better.

6.5. Autonomous Systems

XAI has risen as an invaluable asset in enhancing the safety and dependability of autonomous systems such as self-driving cars and drones. By unveiling the intricate decision-making processes of these systems, XAI plays a pivotal role in the identification and rectification of errors or biases, thereby elevating their trustworthiness. Noteworthy instances abound in this realm.

For instance, Lipton's framework [85] stands as an exemplar of XAI's impact. By harnessing attention mechanisms and layer-wise relevance propagation, this framework delves into the depths of deep learning models, illuminating their inner workings and rendering insights accessible.

Equally compelling is the framework proposed by Kim et al. [86]. This innovation ingeniously amalgamates adversarial training, model introspection, and user feedback to cultivate an adaptable and explicable model for self-driving cars. The outcome is a harmonious blend of adaptability and transparency that resonates with real-world applications.

Moreover, XAI finds resonance in the realm of drones. Keneni's evolutionary approach [87] encapsulates its utility by developing an interpretable decision support system for unmanned aerial vehicles in precision agriculture. The experimental outcomes stand as a testament to the promises that XAI holds in this evolving landscape.

In sum, the integration of XAI into the realm of autonomous systems not only upholds the principles of accountability and reliability but also shapes the trajectory toward a safer and more dependable future for these technologies.

7. Challenges and Limitations

The human-centered approach to XAI takes into account the value that users will ultimately derive from the models. The complexity of these models can often be such that humans struggle to understand them, but more interpretable models may not perform as well—in general, there is a trade-off between accuracy and interpretability. If decision-makers cannot understand their models, they may not trust them, which can lead to a lack of adoption and decrease the ultimate effectiveness of the AI system.

Many multiple criteria decision making methods have assumed a fixed importance weighting for each criterion. The ability of LIMEs to discover the most important variables is questionable when the importance of each factor varies with context, or is even thresholded in a rules-based model (whether explicitly or implicitly). One approach to address this challenge is shown by Främling [88], who introduced the notion of contextual importance and utility (CIU), building on past research on the theory of how humans actually reach decisions. The notion here is that the importance to a human of an input may vary given the context.

Several challenges still remain to be addressed to ensure that XAI systems can be deployed efficiently, safely and ethically. These challenges can be categorized into four main areas, namely technical challenges, ethical challenges, social challenges, and regulatory challenges. One of the technical challenges of XAI is that many of the techniques developed for explaining models have been designed for specific types of models. For example, post hoc techniques such as LIMEs and SHAPs are best suited for models that are based on feature importance, whereas counterfactual explanations are better suited to rule-based systems. Rudin [89] argues that there is a need to move away from black box models towards models that are inherently interpretable. The use of interpretable models can provide a more transparent and understandable decision-making process.

One of the ethical challenges of XAI is that it can be difficult to determine what level of explanation is sufficient for different scenarios. Gunning et al. [3] discuss the need to ensure that the explanations provided by XAI systems are appropriate for the context in which they are deployed. The explanations must be able to convey the necessary information to users without overwhelming them with too much technical detail. Furthermore, the use of XAI must not perpetuate existing biases or discrimination in society.

Social challenges of XAI include issues related to human trust and perception. Shin [90] investigated human trust and perception of AI systems. Their study showed that participants were more likely to trust and accept the decisions of systems that provided explanations for their outputs. However, the study also showed that the level of trust and acceptance varied depending on the type of explanation provided.

The regulatory challenges of XAI include issues related to the legal and ethical implications of AI systems. Gunning et al. [3,91] discuss the need for clear regulations

and standards for the use of XAI in different domains. The regulations must be able to provide guidance on the ethical and legal implications of AI systems, as well as the potential risks associated with their use. Liao et al. [92] discuss the need to ensure that AI systems are held accountable for their decisions. They propose a framework for the design of XAI systems that includes mechanisms for detecting and reporting errors in the decision-making process.

The challenges associated with XAI are varied and require a multidisciplinary approach to address them. Technical challenges require the development of techniques that are applicable across different types of models. Ethical challenges require the development of systems that are transparent, fair, and do not perpetuate existing biases in society. Social challenges require the development of systems that are trusted and perceived as fair by humans. Regulatory challenges require the development of guidelines and standards that can ensure the ethical and legal use of XAI systems.

These challenges can also be seen as opportunities—by using XAI techniques, AI systems can be made more transparent, and their decision-making processes can be better understood by humans. This, in turn, can lead to increased trust and adoption of the AI system. XAI can provide improved accountability and potential regulatory compliance in AI systems.

8. Conclusions and Future Research Directions

With the increasing use of ML models in various domains, there is a growing need for these models to be transparent, interpretable, and accountable. However, the complexity of deep learning and other modern AI models makes generating interpretable explanations challenging and there is in general a trade-off between accuracy and explainability. This highlights the importance of XAI in providing tools to ensure AI system outputs can be understood by human users, enhancing trust, adoption, and ultimate effectiveness of these systems. This can also have a positive impact on regulatory compliance and the ethical use of AI in various fields.

This review provides a comprehensive overview of foundational XAI techniques within the AI domain, including model-agnostic methods, post hoc explanations such as LIMEs and SHAPs, counterfactual explanations, and intrinsically interpretable models. By addressing critical questions ranging from the evolution of interpretability methodologies to the impact of diverse interpretation techniques on fostering human–AI collaboration and regulatory compliance, this review has shed light on the intricate facets of explainability in AI models. Through an exploration of the strengths and limitations associated with the application of these techniques across varied data domains, including images, text, and tabular data, this review has emphasized the nuanced challenges and opportunities prevalent in real-world applications. Furthermore, by examining the transformative influence of advanced computational methods on the dynamic landscape of AI interpretability, this review has revealed the evolving trends that shape the future trajectory of this critical field. As a result, this comprehensive analysis serves as a resource for individuals seeking a holistic understanding of the complexities and implications inherent in the realm of interpretability in AI systems. Furthermore, we highlighted the importance of XAI in various fields such as healthcare, finance and law enforcement and highlighted some of the challenges associated with the development and deployment of XAI systems, including technical, ethical, social, and regulatory challenges. These challenges can also be seen as opportunities for further research and development.

It appears that, while deep learning has shown considerable advantages for "unstructured" data, like images and text, for structured data which comes as a table of numbers say, deep learning is far from the best approach at the moment [93]. What are the implications of this? (1) Massive search and optimization may produce accurate models but they may not be interpretable. However, problems that have a large set of possible solutions (i.e., a large Rashomon set) may include models that are simple, interpretable, robust and accurate. Rudin et al. [89,94] notably advocates this approach and has shown in a number of

cases that large black box models can be replaced by remarkably simple and interpretable models with no lack of accuracy. They outline a number of challenges for interpretable ML [95]; (2) modern deep learning architectures have achieved surprising accuracy while appearing to be over-parameterized and un-regularized. Guo et al. [96] showed that, while classification accuracy is very high, many of these approaches cause the model to be very poorly calibrated. This may be an important point for some use cases where we want the outputs to be accurate estimates of probabilities.

The current limitations of deep learning in handling unstructured data have been noted. However, research in the area of interpretable ML has shown that simple, interpretable models can provide accurate results. Finally, we should consider the importance of model calibration, especially in cases where accurate probability estimates are needed.

To ensure the efficient, safe, and ethical deployment of XAI systems, a multidisciplinary approach is required, involving experts from various fields such as computer science, psychology, philosophy, and law. This approach can help address technical challenges by developing techniques that are applicable across different types of models. Ethical challenges can also be addressed by developing systems that are transparent, fair, and do not perpetuate existing biases in society. Social challenges can be addressed by developing systems that are trusted and perceived as fair by humans. Finally, regulatory challenges can be addressed by developing guidelines and standards that ensure the ethical and legal use of XAI systems.

XAI is an exciting and rapidly changing field with significant potential to transform the way we use AI systems. By addressing the challenges associated with XAI, we can develop more trustworthy, transparent, and effective AI systems that can benefit industry and society as a whole.

Author Contributions: Conceptualization, Y.Z.; formal analysis and investigation, A.T.; writing, original draft preparation, review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Acknowledgments: We thank the editors and three anonymous referees for their helpful comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Padovan, P.H.; de Castro, C.M.M.; Reed, C. Black is the new orange: How to determine AI liability. *Artif. Intell. Law* **2023**, *31*, 133–167. [CrossRef]
2. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
3. Gunning, D.; Vorm, E.; Wang, Y.J.; Turek, M. DARPA's explainable AI (XAI) program: A retrospective. *Appl. AI Lett.* **2021**, *2*. [CrossRef]
4. Abdullah, T.A.A.; Zahid, M.S.M.; Ali, W. *A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions*. *Symmetry* **2021**, *13*, 2439. [CrossRef]
5. Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C.J.; Müller, K.R. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proc. IEEE* **2021**, *109*, 247–278. [CrossRef]
6. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef]
7. Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining Explanations: An Overview of Interpretability of Machine Learning. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 1–3 October 2018; pp. 80–89. Available online: https://api.semanticscholar.org/CorpusID:59600034 (accessed on 1 February 2023).
8. The Lisp Approach to AI. Available online: https://medium.com/ai-society/the-lisp-approach-to-ai-part-1-a48c7385a913 (accessed on 16 February 2023).
9. How Lisp Became God's Own Programming Language. Available online: https://twobithistory.org/2018/10/14/lisp.html (accessed on 16 February 2023).

10. ELIZA on Wikipedia. Available online: https://en.wikipedia.org/wiki/ELIZA (accessed on 18 February 2023).
11. OpenAI. ChatGPT (Mar 14 Version) [Large Language Model]. 2023. Available online: https://chat.openai.com/chat (accessed on 1 May 2023).
12. Wikipedia. MAD (Programming Language). 2023. Available online: https://en.wikipedia.org/wiki/MAD_(programming_language) (accessed on 1 January 2023) .
13. Wikipedia. Universal Approximation Theorem. 2023. Available online: https://en.wikipedia.org/wiki/Universal_approximation_theorem (accessed on 1 November 2023).
14. Paul Workman, Reflecting on DeepMind's AlphaFold Artificial Intelligence Success—What's the Real Significance for Protein Folding Research and Drug Discovery. The Institute of Cancer Research. 2021. Available online: https://www.icr.ac.uk/blogs/the-drug-discoverer/page-details/reflecting-on-deepmind-s-alphafold-artificial-intelligence-success-what-s-the-real-significance-for-protein-folding-research-and-drug-discovery (accessed on 16 February 2023).
15. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774. Available online: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (accessed on 20 March 2023).
16. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R., Eds.; ACM: New York, NY, USA, 2016; pp. 1135–1144. . [CrossRef]
17. Galton, F. Regression Towards Mediocrity in Hereditary Stature. *J. Anthropol. Inst. Great Br. Irel.* **1886**, *15*, 246–263. [CrossRef]
18. McCarthy, J. *Recursive Functions of Symbolic Expressions: Their Computation by Machine, Part I*; Massachusetts Institute of Technology: Cambridge, MA, USA, 1960.
19. Quilan, J.R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
20. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
21. Lundberg, S.M.; Erion, G.G.; Lee, S.I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv* **2019**, arXiv:1802.03888.
22. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv* **2019**, arXiv:1704.02685.
23. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2019**, *128*, 336–359. [CrossRef]
24. Fernandez, A.; Herrera, F.; Cordon, O.; Jose del Jesus, M.; Marcelloni, F. Evolutionary Fuzzy Systems for Explainable Artificial Intelligence: Why, When, What for, and Where to? *IEEE Comput. Intell. Mag.* **2019**, *14*, 69–81. [CrossRef]
25. Wang, T.T.; Rudin, C.; Doshi-Velez, F.; Liu, Y.; Klampfl, E.; MacNeille, P. A Bayesian framework for learning rule sets for interpretable classification. *J. Mach. Learn. Res.* **2017**, *18*, 2357–2393. [CrossRef]
26. Wang, T.; Rudin, C.; Velez-Doshi, F.; Liu, Y.; Klampfl, E.; MacNeille, P. Bayesian Rule Sets for Interpretable Classification. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016; pp. 1269–1274. [CrossRef]
27. Letham, B.; Rudin, C.; McCormick, T.H.; Madigan, D. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.* **2015**, *9*, 1350–1371. Available online: https://www.jstor.org/stable/43826424 (accessed on 1 January 2023). [CrossRef]
28. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. . [CrossRef] [PubMed]
29. Westberg, M.; Zelvelder, A.; Najjar, A. A Historical Perspective on Cognitive Science and Its Influence on XAI Research. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*; Calvaresi, D., Najjar, A., Schumacher, M., Främling, K., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 205–219. [CrossRef]
30. Norouzi, M.; Collins, M.D.; Johnson, M.; Fleet, D.J.; Kohli, P. Efficient Non-greedy Optimization of Decision Trees. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 1729–1737.
31. Mishra, P. *Practical Explainable AI Using Python*; APress Media, LLC, Part of Springer Nature: New York, NY, USA, 2022.
32. Costa, V.G.; Pedreira, C.E. Recent advances in decision trees: An updated survey. *Artif. Intell. Rev.* **2022**, *56*, 4765–4800. [CrossRef]
33. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
34. Berkson, J. Application of the logistic function to bioassay. *J. Am. Stat. Assoc.* **1944**, *39*, 357–365. [CrossRef]
35. Schneider, A.; Hommel, G.; Blettner, M. Linear regression analysis—Part 14 of a series on evaluation of scientific publications. *Dtsch. Ärztebl. Int.* **2010**, *107*, 776–782. [CrossRef]
36. Chao-Ying, J.P.; Kuk, L.L.; Gary, M.I. An Introduction to Logistic Regression Analysis and Reporting. *J. Educ. Res.* **2002**, *96*, 3–14. [CrossRef]
37. Berk, R.A. An Introduction to Ensemble Methods for Data Analysis (arXiv:2110.01889). UCLA, Department of Statistics Papers. Available online: https://escholarship.org/content/qt54d6g9gf/qt54d6g9gf.pdf (accessed on 1 January 2023).

38. Seni, G.; Elder, J.F. *Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions*; Morgan & Claypool Publishers LLC: Kentfield, CA, USA, 2010.
39. Saltelli, A.; Ratto, M.; Andres, T.; Campolongo, F.; Cariboni, J.; Gatelli, D.; Saisana, M.; Tarantola, S. *Global Sensitivity Analysis: The Primer*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
40. Hu, Z.; Mahadevan, S. Global sensitivity analysis-enhanced surrogate (GSAS) modeling for reliability analysis. *Struct. Multidiscip. Optim.* **2016**, *53*, 501–521. [CrossRef]
41. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. Available online: http://www.jstor.org/stable/2699986 (accessed on 1 January 2023) [CrossRef]
42. Pearl, J. Chapter 3—MARKOV AND BAYESIAN NETWORKS: Two Graphical Representations of Probabilistic Knowledge. In *Probabilistic Reasoning in Intelligent Systems*; Pearl, J., Ed.; Morgan Kaufmann: San Francisco, CA, USA, 1988; pp. 77–141. ISBN 978-0-08-051489-5. Available online: https://www.sciencedirect.com/science/article/pii/B9780080514895500096 (accessed on 1 January 2023). [CrossRef]
43. Pearl, J. *Bayesian Networks*; Department of Statistics, UCLA: Los Angeles, CA, USA, 2011. Available online: https://escholarship.org/uc/item/53n4f34m (accessed on 1 January 2023)
44. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Pearson Education: London, UK, 2003; ISBN 0137903952. Available online: http://portal.acm.org/citation.cfm?id=773294 (accessed on 1 January 2023).
45. Itti, L.; Koch, C.; Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
46. Erhan, D.; Bengio, Y.; Courville, A.; Vincent, P. *Visualizing Higher-Layer Features of a Deep Network*; University of Montreal: Montreal, QC, Canada, 2009.
47. Morch, N.J.S.; Kjems, U.; Hansen, L.K.; Svarer, C.; Law, I.; Lautrup, B.; Strother, S.; Rehm, K. Visualization of neural networks using saliency maps. In Proceedings of the ICNN'95-International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December 1995; Volume 4, pp. 2085–2090.
48. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
49. Zahavy, T.; Ben-Zrihem, N.; Mannor, S. Graying the black box: Understanding DQNs. In Proceedings of the 33rd International Conference on Machine Learning (PMLR), New York City, NY, USA, 16–24 June 2016; pp. 1899–1908.
50. Kindermans, P.-J.; Hooker, S.; Adebayo, J.; Brain, G.; Alber, M.; Schütt, K.; Tu-Berlin, S.; Erhan, D.; Brain, K. The (Un) Reliability of Saliency Methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Springer: Cham, Switzerland, 2017; pp. 267–280.
51. Rigotti, M.; Miksovic, C.; Giurgiu, I.; Gschwind, T.; Scotton, P. Attention-based Interpretability with Concept Transformers. *International Conference on Learning Representations*. 2022. Available online: https://www.matrig.net/publications/articles/rigotti2022.pdf (accessed on 1 January 2023).
52. Ghaeini, R.; Fern, X.; Tadepalli, P. Interpreting Recurrent and Attention-Based Neural Models: A Case Study on Natural Language Inference. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018.
53. Jain, S.; Wallace, B. Attention is not Explanation. *arXiv* **2019**, arXiv:1902.10186.
54. Neely, M.; Schouten, S.; Bleeker, M.; Lucic, A. Order in the Court: Explainable AI Methods Prone to Disagreement. *arXiv* **2021**, arXiv:2105.03287.
55. Serrano, S.; Smith, N. Is Attention Interpretable? *arXiv* **2019**, arXiv:1906.03731.
56. Mylonas, N.; Mollas, I.; Tsoumakas, G. Improving attention-based interpretability of text classification transformers: A preprint. *arXiv* **2022**, arXiv:2209.10876.
57. Lipovetsky, S.; Conklin, M. Analysis of regression in game theory approach. *Appl. Stoch. Model. Bus. Ind.* **2001**, *17*, 319–330. [CrossRef]
58. Främling, K.; Westberg, M.; Jullum, M.; Madhikermi, M.; Malhi, A. Comparison of Contextual Importance and Utility with LIME and Shapley Values. In *Explainable and Transparent AI and Multi-Agent Systems*; Calvaresi, D., Najjar, A., Winikoff, M., Främling, K., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; Volume 12688, pp. 39–54. [CrossRef]
59. Woods, W.; Chen, J.; Teuscher, C. Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nat. Mach. Intell.* **2019**, *1*, 508–516. [CrossRef]
60. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef]
61. Kim, B.; Shah, J.A.; Doshi-Velez, F. Mind the Gap: A Generative Approach to Interpretable Feature Selection and Extraction. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28. Available online: https://proceedings.neurips.cc/paper/2015/file/82965d4ed8150294d4330ace00821d77-Paper.pdf (accessed on 1 January 2020).
62. Singh, V.; Cyras, K.; Inam, R. Explainability Metrics and Properties for Counterfactual Explanation Methods. In *Practical Explainable AI Using Python*; APress Media, LLC, Part of Springer Nature: New York, NY, USA, 2022.

63. Jentzsch, S.F.; Höhn, S.; Hochgeschwender, N. Conversational Interfaces for Explainable AI: A Human-Centred Approach. In *Practical Explainable AI Using Python*; APress Media, LLC, Part of Springer Nature: New York, NY, USA, 2019.
64. Keane, M.T.; Kenny, E.M. How Case-Based Reasoning Explains Neural Networks: A Theoretical Analysis of XAI Using Post-Hoc Explanation-by-Example from a Survey of ANN-CBR Twin-Systems. In *Case-Based Reasoning Research and Development*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 155–171. [CrossRef]
65. La Malfa, E.; Zbrzezny, A.; Michelmore, R.; Paoletti, N.; Kwiatkowska, M. On guaranteed optimal robust explanations for NLP models. *arXiv* **2021**, arXiv:2105.03640.
66. Ignatiev, A. Towards Trustable Explainable AI. Electronic proceedings of IJCAI 2020. Available online: https://www.ijcai.org/proceedings/2020/726 (accessed on 1 January 2023).
67. Amazon. Machine Learning Service—Amazon SageMaker. Available online: https://aws.amazon.com/pm/sagemaker/ (accessed on 7 November 2023).
68. Google. Introduction to Vertex Explainable AI. Available online: https://cloud.google.com/vertex-ai/docs/explainable-ai/overview (accessed on 7 November 2023).
69. Microsoft. Azure Machine Learning. Available online: https://azure.microsoft.com/en-au/products/machine-learning (accessed on 7 November 2023).
70. IBM. IBM Watson Studio. Available online: https://www.ibm.com/products/watson-studio (accessed on 7 November 2023).
71. The Linux Foundation. AI Explainability 360: Understand How ML Models Predict Labels. Available online: https://ai-explainability-360.org/ (accessed on 7 November 2023).
72. Mucaki, E.J.; Baranova, K.; Pham, H.Q.; Rezaeian, I.; Angelov, D.; Ngom, A.; Rueda, L.; Rogan, P.K. Predicting outcomes of hormone and chemotherapy in the molecular taxonomy of breast cancer international consortium (METABRIC) study by biochemically-inspired machine learning. *F1000Research* **2016**, *5*, 2124. [CrossRef]
73. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2097–2106.
74. Rajkomar, A.; Oren, E.; Chen, K.; Dai, A.M.; Hajaj, N.; Hardt, M.; Liu, P.J.; Liu, X.; Marcus, J.; Sun, M.; et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **2018**, *1*, 18. [CrossRef]
75. Bussmann, N.; Giudici, P.; Marinelli, D.; Papenbrock, J. Explainable Machine Learning in Credit Risk Management. *Comput. Econ.* **2021**, *57*, 203–216. [CrossRef]
76. Kumar, S.; Vishal, M.; Ravi, V. Explainable Reinforcement Learning on Financial Stock Trading using SHAP. *arXiv* **2022**, arXiv:2208.08790.
77. Ji, Y. Explainable AI Methods for Credit Card Fraud Detection: Evaluation of LIME and SHAP through a User Study. Master's Thesis, University of Skövde, School of Informatics, Skövde, Sweden, 2021.
78. Gu, Y.; Li, B.; Meng, Q. Hybrid interpretable predictive machine learning model for air pollution prediction. *Neurocomputing* **2022**, *468*, 123–136. Available online: https://www.sciencedirect.com/science/article/pii/S0925231221014296 (accessed on 1 January 2023). [CrossRef]
79. Van Straaten, C.; Whan, K.; Coumou, D.; Van den Hurk, B.; Schmeits, M. Using Explainable Machine Learning Forecasts to Discover Subseasonal Drivers of High Summer Temperatures in Western and Central Europe. *Mon. Weather Rev.* **2022**, *150*, 1115–1134. [CrossRef]
80. Wu, X.; Zhang, Q.; Wen, F.; Qi, Y. *A Water Quality Prediction Model Based on Multi-Task Deep Learning: A Case Study of the Yellow River, China*. *Water* **2022**, *14*, 3408. [CrossRef]
81. Park, J.; Lee, W.H.; Kim, K.T.; Park, C.Y.; Lee, S.; Heo, T.Y. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. *Sci. Total Environ.* **2022**, *832*, 155070. [CrossRef] [PubMed]
82. Dressel, J.; Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **2018**, 4. [CrossRef] [PubMed]
83. Kusner, M.J.; Loftus, J.; Russell, C.; Silva, R. Counterfactual Fairness. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4066–4076.
84. Rudin, C.; Radin, J. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harv. Data Sci. Rev.* **2019**, *1*. [CrossRef]
85. Lipton, Z.C. The Mythos of Model Interpretability. *arXiv* **2018**, arXiv:1606.03490.
86. Kim, J.; Rohrbach, A.; Akata, Z.; Moon, S.; Misu, T.; Chen, Y.T.; Darrell, T.; Canny, J. Toward explainable and advisable model for self-driving cars. *Appl. AI Lett.* **2021**, **4**, 415–422. [CrossRef]
87. Keneni, B.M. Evolving Rule-Based Explainable Artificial Intelligence for Decision Support System of Unmanned Aerial Vehicles. Master's Thesis, University of Toledo, Toledo, OH, USA, 2018. Available online: http://rave.ohiolink.edu/etdc/view?acc_num=toledo1525094091882295 (accessed on 1 January 2020).
88. Främling, K. Decision Theory Meets Explainable AI. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*; Calvaresi, D., Najjar, A., Winikoff, M., Främling, K., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 57–74. ISBN 978-3-030-51923-0/978-3-030-51924-7. Available online: http://link.springer.com/10.1007/978-3-030-51924-7_4 (accessed on 1 January 2022). [CrossRef]
89. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]

90. Shin, D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int. J. Hum. Comput. Stud.* **2021**, *146*, 102551. [CrossRef]
91. Gunning, D.; Aha, D.W. DARPA's Explainable Artificial Intelligence Program. *AI Mag.* **2019**, *40*, 44–58. [CrossRef]
92. Liao, Q.V.; Gruen, D.; Miller, S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; Available online: https://api.semanticscholar.org/CorpusID:210064344 (accessed on 20 March 2023).
93. Borisov, V.; Leemann, T.; Seßler, K.; Haug, J.; Pawelczyk, M.; Kasneci, G. Deep Neural Networks and Tabular Data: A Survey. *arXiv* **2022**, arXiv:2110.01889. Available online: http://arxiv.org/abs/2110.01889 (accessed on 29 June 2022). [CrossRef].
94. Fisher, A.; Rudin, C.; Dominici, F. All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *arXiv* **2019**, arXiv:1801.01489.
95. Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *arXiv* **2021**, arXiv:2103.11251.
96. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. *arXiv* **2017**, arXiv:1706.04599.