algorithms

MDPI

Article

Explainable AI Frameworks: Navigating the Present Challenges and Unveiling Innovative Applications

Neeraj Anand Sharma [1,*], Rishal Ravikesh Chand [1], Zain Buksh [1], A. B. M. Shawkat Ali [1], Ambreen Hanif [2] and Amin Beheshti [2,*]

[1] Department of Computer Science and Mathematics, University of Fiji, Lautoka P.O. Box 42458, Fiji; rishalc@unifiji.ac.fj (R.R.C.); zainb@unifiji.ac.fj (Z.B.); shawkata@unifiji.ac.fj (A.B.M.S.A.)
[2] School of Computing, Macquarie University, Balaclava Rd, Macquarie Park, NSW 2109, Australia; ambreen.hanif@hdr.mq.edu.au
* Correspondence: neerajs@unifiji.ac.fj (N.A.S.); amin.beheshti@mq.edu.au (A.B.)

Abstract: This study delves into the realm of Explainable Artificial Intelligence (XAI) frameworks, aiming to empower researchers and practitioners with a deeper understanding of these tools. We establish a comprehensive knowledge base by classifying and analyzing prominent XAI solutions based on key attributes like explanation type, model dependence, and use cases. This resource equips users to navigate the diverse XAI landscape and select the most suitable framework for their specific needs. Furthermore, the study proposes a novel framework called XAIE (eXplainable AI Evaluator) for informed decision-making in XAI adoption. This framework empowers users to assess different XAI options based on their application context objectively. This will lead to more responsible AI development by fostering transparency and trust. Finally, the research identifies the limitations and challenges associated with the existing XAI frameworks, paving the way for future advancements. By highlighting these areas, the study guides researchers and developers in enhancing the capabilities of Explainable AI.

Keywords: artificial intelligence; black box; explainable AI; framework; techniques; XAI

Citation: Sharma, N.A.; Chand, R.R.; Buksh, Z.; Ali, A.B.M.S.; Hanif, A.; Beheshti, A. Explainable AI Frameworks: Navigating the Present Challenges and Unveiling Innovative Applications. *Algorithms* **2024**, *17*, 227. https://doi.org/10.3390/a17060227

Academic Editors: Nan-Run Zhou and Hua-Lei Yin

Received: 19 April 2024
Revised: 19 May 2024
Accepted: 21 May 2024
Published: 24 May 2024

1. Introduction

As Artificial Intelligence (AI) rapidly evolves, new and powerful technologies emerge alongside unforeseen challenges. This rapid growth necessitates the continuous development of tools and methods to ensure AI's responsible and ethical use. One such crucial development is Explainable Artificial Intelligence (XAI) [1]. Imagine a machine learning (ML) model that predicts loan approvals. Traditionally, these models might be like a black box—where someone could input data and obtain an output (approved/rejected) without understanding why the model made that decision. XAI techniques aim to shed light on this process, helping users to understand the factors influencing the model's predictions. Some benefits of XAI are trust enhancement with transparency, improved model development, and better decision-making.

XAI is a field dedicated to making the decisions and outputs of AI models understandable and transparent for humans. This transparency is crucial for several reasons. First, it helps to build trust. When people understand how AI systems reach conclusions, they are more likely to trust and accept those results. This is particularly important in critical domains like healthcare, finance, and law, where AI decisions can significantly impact people's lives. Second, XAI fosters improved understanding. Deeper insights into AI models' inner workings can help in understanding the decision-making processes. This knowledge is invaluable for debugging and improving models, also ensuring that they align with human values and ethical guidelines. Finally, XAI techniques play a crucial role in identifying and mitigating bias. These techniques can be used to detect and address

the potential biases that may be present in AI systems, leading to fairer and more ethical development [2].

According to [3], the future of XAI lies in methods that capture causal relationships. This necessitates the development of metrics to assess the explanation quality, ensuring that users grasp causality effectively, efficiently, and with satisfaction in specific use cases. New human–AI interfaces are foreseen, enabling contextual understanding and user interaction through questions and counterfactuals. By fostering explainability, XAI plays a vital role in navigating the challenges of the rapidly evolving AI landscape. It paves the way for building trust, ensuring responsible AI development, and, ultimately, unlocking AI's full potential for society's benefit.

1.1. Problem Statement

Achieving interpretable and trustworthy AI systems remains a significant challenge. XAI techniques aim to address this by providing users insights into model decisions. However, several limitations hinder our ability to understand and evaluate XAI frameworks, thereby impeding the progress toward genuinely interpretable AI. This section outlines these limitations, highlighting the areas where further research is crucial for advancing the field of XAI.

Despite the rapid advancements in XAI frameworks, there exists a lack of a comprehensive understanding of the diverse attributes of these frameworks and how they cater to different use cases and technical requirements. This ambiguity hinders the optimal selection and practical application of XAI frameworks within various domains, potentially leading to transparency, trust, and responsible AI development limitations. In [4], the authors identify several limitations that hinder our understanding of XAI frameworks and their performance needs. One key aspect is the lack of specific evaluation metrics. It is difficult to gauge their effectiveness without clear metrics to assess how well XAI frameworks explain model decisions. Additionally, another study emphasizes the importance of developing explanation methods, often neglecting the role of user interpretation [2]. Even if an XAI framework produces an explanation, it is crucial to understand if the users can interpret it meaningfully.

Furthermore, research by [5] points out that the limited details on XAI frameworks themselves create a barrier to understanding. By not delving into the specifics of the individual methods and functionalities, it is difficult to know what kind of performance to expect from these frameworks. Finally, another study by [6] emphasizes the need for empirical evaluation methods to assess XAI frameworks in practice. Developing these methods would be crucial for bridging the gap between the theoretical XAI concepts and the practical understanding of their effectiveness. By addressing these limitations, future research can move us closer to a more comprehensive understanding of XAI frameworks and their role in achieving interpretable and trustworthy AI systems.

This research aims to bridge this knowledge gap by analyzing the key attributes of XAI frameworks, including the following:

- Explanation types (extrinsic or intrinsic);
- Model-dependence (model-agnostic or model-specific);
- Explanation scope (global or local);
- Output format (arguments, text, visualizations, etc.);
- Use cases;
- Programming language compatibility;
- Applicability to different ML and deep learning paradigms;
- Model complexity considerations;
- Open-source availability and license type;
- Associated challenges and limitations;
- Primary strengths and potential benefits.

By providing a structured and comparative analysis of these attributes, this research will contribute to a deeper understanding of the capabilities and limitations of XAI frameworks.

1.2. Objectives

The primary objective of this study is to create a comprehensive knowledge base and comparative analysis of prominent XAI frameworks, enabling researchers, developers, and stakeholders to effectively evaluate and implement XAI solutions that meet their specific interpretability, transparency, and accountability needs.

- Categorize XAI frameworks: Develop a methodical framework for classifying XAI solutions based on their core attributes and mechanisms.
- Comparative Analysis: Perform a detailed comparison of the major XAI frameworks, evaluating their strengths, weaknesses, and suitability for various use cases and ML model types.
- Technical Considerations: Examine the technical implications of different XAI frameworks, including programming language dependencies, compatibility with ML and deep learning models, and performance considerations.
- XAI Selection Guidance: Create guidelines and recommendations for practitioners to aid in selecting the most appropriate XAI framework for their specific application scenarios.
- Identify Challenges and Opportunities: Highlight the existing limitations and challenges associated with the current XAI frameworks and pinpoint areas for future research and development.

This research aims to significantly contribute to the existing body of knowledge in XAI by offering several key elements. Firstly, it will establish a compendium of XAI solutions, providing a valuable resource to understand the diverse characteristics and landscapes of the available frameworks. Secondly, the proposed classification scheme will facilitate an objective evaluation framework, enabling practitioners to effectively assess the suitability of different XAI solutions for their specific needs. Furthermore, the study will yield actionable recommendations (practical insights). Finally, by identifying the current limitations, this research will act as a foundation for future research, paving the way for advancements in the capabilities of Explainable AI.

1.3. Methodology

This study will employ a methodical analysis approach to investigate the diverse XAI frameworks. This will involve the following:

- Literature review: An extensive review of the existing research on XAI frameworks and the established classification schemes.
- Framework selection: Identifying and selecting a representative set of prominent XAI frameworks based on their popularity, application domains, and technical characteristics.
- Attribute analysis: Performing a comprehensive analysis of each selected framework based on predefined attributes, such as explanation type, model dependence, output format, and use cases. This may involve a documentation review, code analysis (if open-source), and potentially direct interaction with the frameworks through application programming interfaces (APIs) or simple test cases.
- Comparative analysis: Evaluating and comparing the capabilities and limitations of different XAI frameworks across various attributes, highlighting their strengths and weaknesses and suitability for different use cases.

1.4. Related Works

The field of XAI has witnessed a surge in research, as evidenced by the numerous review studies (see Table 1). However, a gap exists in comprehensively analyzing the existing XAI frameworks. By identifying the limitations in the prior research, we aim to address these challenges and contribute to a deeper understanding of XAI frameworks. Our work focuses on reviewing the recent XAI papers and dissecting their contributions and limitations to pave the way for future advancements.

Table 1. Summary of contributions and limitations of existing studies on XAI in the last five years.

Paper	Year	Contribution	Limitation
[4]	2023	The research highlights that current toolkits only cover a limited range of explanation quality aspects. This work guides researchers in selecting evaluation methods that offer a broader assessment of explanation quality. The research suggests cross-evaluation between toolkits/frameworks and original metric implementations to identify and address inconsistencies.	This research focuses on the limitations of existing toolkits and does not propose a solution (e.g., a new evaluation framework). The specific details of identified evaluation gaps (modalities, explanation types, and missing metrics) are not explicitly mentioned in the paper.
[1]	2023	The research provides a broad overview of various XAI techniques, encompassing white-box models (interpretable by design) and black-box models with techniques like LIMEs and SHAPs to enhance interpretability. The research highlights the need for a deeper understanding of XAI limitations and explores the importance of considering data explainability, model explainability, fairness, and accountability in tandem. It emphasizes the need for a holistic approach to XAI implementation and deployment.	While the research is comprehensive, it might not delve deep into the specifics of each XAI technique or provide detailed comparisons. The research focuses on identifying limitations and considerations, but it might not offer specific solutions or address how to address the challenges raised.
[7]	2022	The chapter overviews available resources and identifies open problems in utilizing XAI techniques for user experience design. The chapter highlights the emergence of a research community dedicated to human-centered XAI, providing references for further exploration.	The chapter focuses on the need for bridging design and XAI techniques but does not provide specific details on how to achieve this in practice. The chapter also focuses on the high-level importance of human-centered design but does not delve into specific design techniques for XAI user experiences.
[8]	2021	The authors reviewed existing research and categorized various design goals and evaluation methods for XAI systems. To address the need for a comprehensive approach, the authors propose a new framework that connects design goals with evaluation methods.	The framework provides a high-level overview of a multidisciplinary approach. Still, it does not delve into specific details of interface design, interaction design, or the development of interpretable ML techniques. It does not cover all aspects like interface design and the development of interpretable models.
[2]	2021	The research demonstrates how the framework can be used to compare various XAI methods and understand the aspects of explainability addressed by each. The framework situates explanations and interpretations within the bigger picture of an XAI pipeline, considering input/output domains and the gap between mathematical models and human understanding.	The review of current XAI practices reveals a focus on developing explanation methods, often neglecting the role of interpretation. It acknowledges the current limitations of XAI practices and highlights areas for future research, such as explicitly defining non-functional requirements and ensuring clear interpretations accompany explanations.
[6]	2021	The research acknowledges the growing emphasis on human stakeholders in XAI research and the need for evaluating explainability approaches based on their needs. The research emphasizes the importance of empirically evaluating explainability approaches to assess how well they meet stakeholder needs.	The research emphasizes the need for empirical evaluation but does not discuss specific methods for conducting such evaluations. While advocating for a broader view, the research does not delve into specific considerations for different stakeholder groups.
[9]	2021	The research proposes a classification framework for explanations considering six aspects: purpose, interpretation method, context, presentation format, stakeholder type, and domain. The research acknowledges the need to expand the framework by including more explanation styles and potentially sub-aspects like user models.	The research focuses on a review of existing explanations, and the provided classification might not encompass all possible XAI explanation styles. While acknowledging the need for personalization, the research does not delve deep into how the framework can be used to personalize explanations based on user needs.

Table 1. *Cont.*

Paper	Year	Contribution	Limitation
[10]	2020	The research emphasizes the importance of XAI in today's ML landscape, considering factors like data bias, trustability, and adversarial examples. The research provides a survey of XAI frameworks, with a focus on model-agnostic post hoc methods due to their wider applicability.	The research specifically focuses on XAI for deep learning models, potentially limiting its generalizability to other ML algorithms. While providing a survey of frameworks, the research does not delve into the specifics of individual XAI methods or their functionalities.
[11]	2019	The research clarifies key concepts related to model explainability and the various motivations for seeking interpretable ML methods. The study proposes a global taxonomy for XAI approaches, offering a unified classification of different techniques based on common criteria.	With limited detail on frameworks, while providing a taxonomy, the research might not delve into the specifics of individual XAI techniques or their functionalities. The research focuses on identifying limitations and future considerations, but it does not offer specific solutions or address how to address the challenges raised.
[5]	2018	The research provides a comprehensive background on XAI, leveraging the "Five Ws and How" framework (What, Who, When, Why, Where, and How) to cover all key aspects of the field. The research emphasizes the impact of XAI beyond academic research, highlighting its applications across different domains.	The research identifies challenges and open issues in XAI but does not offer specific solutions or address how to address them. The research provides limited technical details while providing a broad overview of XAI approaches, and it does not delve into the technical specifics of each method.

1.5. Our Contribution

This study aims to make significant extended contributions beyond our previous works [12,13] to the field of XAI by addressing a critical need for comprehensive knowledge and informed decision-making when selecting XAI solutions. Several studies highlight the limitations of the current evaluation methods and the need for broader frameworks that encompass various aspects of explanation quality. A study by [4] emphasizes that the current XAI toolkits often cover a limited range of explanation quality aspects. This research guides researchers in selecting evaluation methods that offer a more comprehensive assessment, suggesting cross-evaluation between the frameworks and original metric implementations to identify and address inconsistencies. While [4] does not propose a new framework itself, it emphasizes the need for such advancements. Building upon these limitations, this study will establish a comprehensive knowledge base of XAI frameworks, analyzing them based on various attributes like explanation type, model dependence, and output format. This will provide a clear understanding of their capabilities and potential shortcomings in addressing the diverse explanation quality aspects. Research by [6,8] highlights the need for a stronger focus on empirical evaluation methods to assess how well the XAI explanations meet the needs of human stakeholders and to have a framework that connects the design goals with the evaluation methods for XAI systems, ensuring a more holistic approach. By incorporating these insights, this research will develop a methodical framework for evaluating XAI frameworks, empowering practitioners to make informed decisions based on their specific needs and application context. This will ultimately contribute to improved transparency, trust, and responsible AI development. Finally, the study aims to contribute to the future advancement of XAI by highlighting the current limitations and challenges. This aligns with the observations made in [1,5,11], where these papers emphasize the need for a deeper understanding of the XAI limitations, including the lack of comprehensive overviews of the existing frameworks and the focus on developing explanations without sufficient attention to user interpretability. By addressing these gaps, this research paves the way for identifying promising research directions that can enhance the capabilities and user-centricity of Explainable AI in the future. It is important to note that this paper currently does not present any technical analysis with comparative evaluations as this was not the scope of this article, but it will be addressed in future works.

1.6. Paper Structure

The paper is divided into seven sections as shown in Figure 1. Section 2 provides the background information and explanations on the research topic. Section 3 outlines the research methodologies that will be employed to complete the literature review. Section 4 presents and discusses the main findings of this research, organized according to the research objectives. Section 5 identifies the challenges and provides applicable recommendations for XAI frameworks. Section 6 focuses on the future works. In Section 7, we provide concluding statements.

Figure 1. This figure explains the structure and organization of the paper.

2. Research Background

2.1. What Is Explainable AI?

A developing field in ML and AI called XAI aims to make the decisions made by AI systems understandable for humans. As AI systems' capabilities develop, it is becoming increasingly important to close the gap between the opaque "black-box" character of many AI algorithms and the demands of interpretability and disclosure [5]. XAI is primarily concerned with shedding light on the inner workings of AI models, offering an understanding of how these algorithms make their judgments or predictions [7]. Consider an AI system as a complicated puzzle [14], in which each piece signifies a step in the algorithm's decision-making protocol [15]. By explaining the importance of each piece and how it fits together to produce the larger picture, XAI acts as a guiding light for academics and stakeholders as they navigate through this complex jigsaw [10]. By using a blend of sophisticated algorithms, visualization strategies, and model-independent approaches, XAI enables people to solve the puzzles around AI, promoting confidence, comprehension, and responsibility in the application of AI technology in many fields [14]. Essentially, XAI shapes a future in which humans and computers work together harmoniously, clearly understanding one another and improving the interpretability of AI systems while simultaneously establishing the foundation for informed decision-making [16].

2.2. Traditional AI vs. Explainable AI

Traditional AI excels in achieving intelligent behavior and solving problems within well-defined domains. Techniques like rule-based systems, decision trees, and logic programming often deliver highly accurate and efficient results. These systems boast transparency, interpretability, and ease of maintenance due to their reliance on clear rules and logic. However, traditional AI struggles with adaptability and flexibility. Faced with new situations or complex, nuanced problems outside their predefined rules, these models can become brittle and prone to errors. Their limited ability to learn and adapt to new information further restricts their effectiveness in handling dynamic environments [17–19].

Explainable AI focuses on making the internal workings and decision-making processes of AI models more transparent and interpretable to humans. It utilizes methods like feature importance analysis, LIMEs (Local Interpretable Model-Agnostic Explanations), and SHAPs (SHapley Additive exPlanations) to explain how AI models arrive at their predictions. Some of XAI's strengths are the increased trust and acceptance of AI systems by stakeholders, enabling the identification and mitigation of potential biases in AI models, and improved debugging and model improvement capabilities. However, like traditional AI, XAI also has limitations and weaknesses, such possibly being computationally expensive and resource-intensive, not all AI models can be fully explained, especially

complex deep learning models, and explainability techniques may not always be readily understandable by non-experts [17–19].

In essence, traditional AI prioritizes achieving the desired outcome, while XAI focuses on understanding how that outcome is achieved. While traditional AI remains valuable in specific applications, XAI is becoming increasingly important for building trust, ensuring fairness, and the responsible development of AI systems.

2.3. Frameworks in XAI

XAI frameworks are organized approaches, computations, and technologies intended to make the creation and use of transparent and comprehensible AI systems easier. Recognizing that standard AI models are inherently opaque and frequently function as "black boxes", making it difficult for users to comprehend their decision-making processes, the idea of frameworks in XAI was born [20]. The origins of frameworks in XAI may be found in the early stages of AI research when attaining high prediction accuracy levels was the main goal without necessarily placing a strong priority on interpretability [21]. However, when AI applications started to appear in several industries, such as healthcare, banking, and criminal justice, questions about responsibility, equity, and confidence in AI systems started to surface [22]. The necessity of creating approaches that might illuminate the internal operations of AI models was acknowledged by researchers and practitioners; this would help stakeholders to comprehend and have faith in the choices that these systems make [23]. Frameworks play a crucial role in bridging the gap between advanced AI algorithms and human comprehension, as the introduction to XAI emphasizes [20]. Frameworks facilitate informed decision-making and improve transparency by offering organized methods for interpretability, which enable users to derive useful insights from AI models [24]. Additionally, XAI frameworks provide standardized approaches and tools that can be used in a variety of domains and applications, acting as a catalyst for the advancement of research in the subject [25]. Frameworks in XAI constitute a significant advancement in AI, providing organized methods for resolving the interpretability problem that arises naturally in AI systems [26]. By encouraging openness, responsibility, and confidence in the use of intelligent systems across various fields, frameworks are vital in influencing the direction of AI as the subject develops [27].

2.4. Categories in XAI

Model-Specific or Model-Agnostic: This categorization determines whether the interpretation method is tailored to a specific model or applicable across various ML models. Model-specific methods and tools are designed for a particular model, while model-agnostic methods can be applied universally to any ML model to gain interpretability. Model-agnostic methods cannot access internal model data such as weights and structural details [28].

Intrinsic or Extrinsic (post hoc): This classification distinguishes between inherently interpretable models and those requiring post-training methods to achieve interpretability. Intrinsic models, like decision trees, are inherently interpretable, while extrinsic methods rely on post-training strategies to attain interpretability [28].

Local or Global: This differentiation depends on whether the interpretation method describes a single data record or the behavior of the entire model. Global methods interpret the entire model, whereas local methods focus solely on explaining individual predictions [28].

3. Research Methodology

3.1. Literature Review

Since the idea behind this research involves reviewing and identifying existing papers on XAI frameworks, it is only fitting that we outline the literature search criteria and the necessary process we need to follow to achieve and review quality papers. The processes are highlighted as follows:

3.1.1. Constructing the Search Terms

Words such as XAI, Explainable AI, XAI frameworks, XAI techniques, and XAI challenges were searched in several notable databases. We utilized the XAI framework names in the search to obtain specific results, such as XAI LIME, XAI SHAP, and so on.

3.1.2. Inclusion Criteria

The inclusion criteria were determined based on key terms matched, the quality of the publication, the experimental design with relevant results, and, in some cases, the number of citations obtained. These generally included studies that extensively looked into XAI frameworks.

3.1.3. Exclusion Criteria

The following criteria determined why we excluded certain papers:

- Some studies were misleading and were not relevant to our research objectives.
- There were many papers focused on a single application concept that did not mention much on the XAI framework side.
- The papers did not mention any of the key attributes of the XAI framework.

3.2. Selecting Primary Sources

The selection of primary sources was carried out in two phases, the former being the primary selection and the latter being the final selection. We looked at the title, keywords, and abstract in the primary selection phase to see if the research paper fit our needs. If it did, we would select it as our final selection; if not, we rejected it. The final selection phase includes papers found helpful in the primary selection phase. Here, we extensively read through the entire paper to understand the objectives, generate ideas, understand the experimental setup, and identify the gaps and limitations. Table 2 summarizes the entire literature search and selection criteria.

Table 2. Literature search and selection criteria for XAI frameworks.

Stage	Description	Details
Search Terms	Keywords used for searching	XAI Explainable AI, XAI frameworks, XAI techniques, XAI challenges, specific framework names (e.g., XAI LIME and XAI SHAP)
Databases	Platforms used for searching	IEEE Xplore, Science Direct, MDPI, Google Scholar, ACM Digital Library, Scopus, and arXiv.
Inclusion Criteria	Criteria for Selecting Relevant Papers	• Papers matching key terms • High publication quality • Strong experimental design with relevant results • High citation count (in some cases) • Extensive focus on XAI frameworks
Exclusion Criteria	Reasons for excluding papers	• Misleading content irrelevant to research objectives • Focus on single application concept with limited XAI framework discussion • Lack of mention of key XAI framework attributes • Duplicate ideas or methods mentioned in the paper
Primary Source Selection	Process for selecting key papers	**Phase 1:** • Review the title, keywords, and abstract to assess relevance. • Select papers that align with research needs. • Reject irrelevant papers. **Phase 2:** • Conduct an in-depth review of selected papers. • Analyze objectives, generate ideas, understand the experimental setup, and identify gaps and limitations.

3.3. Framework Selection

We will focus on three main aspects: how popular they are, what areas they are used for, and their technical features.

First, we will gather information on various XAI frameworks from academic publications, open-source repositories, and industry reports. To measure popularity, we will look at factors like how often research papers on a framework are cited, how active the framework's online community is, and how often it is mentioned in industry reports.

Next, we will categorize the frameworks by the problems they are typically used for. This could include areas like image recognition, analyzing text, healthcare applications, finance, or general-purpose frameworks that can be used across many tasks.

Finally, we will analyze the technical details of each framework to understand what kind of explanations it can provide and how it makes those explanations understandable to people. We will consider factors like whether the framework can explain any model or just specific types, what techniques it uses to make explanations clear, and for whom the explanations are designed (data scientists, other experts, or even regular people with no technical background).

By considering these aspects of popularity, application area, and technical features, we can select a representative set of XAI frameworks that capture the field's current state. This will include frameworks that are both popular and unpopular in many different areas, and that use a variety of other techniques to explain AI models in ways that people can understand.

3.4. Attribute Analysis

To provide a comprehensive understanding of prominent XAI frameworks, we will analyze them based on a set of predefined attributes. This analysis empowers researchers and practitioners to select the most suitable framework for their specific application domain. The attributes considered include the following:

- *Year Introduced*: This indicates the framework's maturity and potential level of development.
- *Reference Paper*: The core research paper describing the framework's theoretical foundation and methodology.
- *XAI Category*: This classifies the framework based on its explanation approach (e.g., extrinsic, intrinsic, model-agnostic, model-specific, global, or local).
- *Output Formats*: The types of explanations the framework generates (e.g., arguments, textual explanations, or visualizations).
- *Current Use Cases*: Examples of real-world applications where the framework has been employed.
- *Programming Language*: The primary programming language used to develop the framework.
- *Machine- and Deep Learning Scope*: Whether the framework is designed for traditional ML models or specifically caters to deep learning models.
- Understanding Model Complexity: The framework's ability to explain complex models effectively.
- *Development Model*: Categorization as a commercial or open-source framework.
- *License Type*: Applicable only to open-source frameworks, specifying the license terms under which they are distributed.

By analyzing these predefined attributes, we aim to create a detailed profile for each chosen XAI framework. This profile will highlight the framework's strengths, limitations, and suitability for various application domains.

3.5. Comparative Analysis

By analyzing these predefined attributes, we aim to create a detailed profile for each chosen XAI framework. This profile will highlight the framework's strengths, limitations, and suitability for various application domains.

Next, we will assess each framework's strengths based on the analysis. For example, a framework offering various explanation formats like text and visualizations provides more explanation flexibility. We will then identify limitations based on the attributes. For instance, a framework designed for basic ML models might struggle with complex deep learning models. To fully understand the landscape, we will compare different frameworks across two major attributes that are output format and model complexity. This will involve creating tables or charts to visualize their strengths and weaknesses side-by-side.

Finally, to highlight use case suitability, we will define representative use cases from various application domains like marketing, healthcare, or finance. By considering the use case requirements (e.g., explaining complex models and generating visualizations), we will identify the XAI frameworks best suited for each scenario. This matching process will leverage the strengths identified earlier. We will then discuss the rationale behind these selections, highlighting the factors that make them the most suitable options.

Through this multi-step evaluation and comparison, we aim to provide a clear picture of XAI framework capabilities and limitations. By showcasing their strengths, weaknesses, and suitability for various use cases, this methodology will empower researchers and practitioners to make informed decisions when selecting an XAI framework for their specific needs.

4. Discussion

In the rapidly evolving landscape of AI, the need for XAI frameworks has become increasingly apparent, as well as a comprehensive examination of the XIA frameworks, with a specific focus on defining and analyzing each framework's essence. The necessity for XAI frameworks has become increasingly apparent in the rapidly advancing field of AI as they provide critical transparency and interpretability to AI systems. The aim is to offer an understanding of their roles in enhancing the transparency and interpretability within AI systems. Through this analysis, readers will gain invaluable insights into the diverse landscape of XAI frameworks, thereby facilitating informed decision-making and implementation strategies in AI development. Table 3 offers a comprehensive comparison of 31 XAI frameworks, categorized by the key attributes that influence their application. It serves as a critical tool for researchers and practitioners in the field of XAI. By comparing these frameworks across nine key characteristics, users can make informed decisions about which framework best suits their specific needs when working with AI models. The table investigates the details beyond just the framework names. It includes the publication year of the research paper introducing each framework, providing insight into its relative maturity. The citations for the papers are also provided, allowing for further exploration of each framework's functionalities.

The remaining attributes focus on the functionalities themselves. These characteristics can be grouped into three main categories:

- Explanation Focus: This distinction clarifies whether the framework provides explanations for the overall model's behavior (global explanations) or focuses on explaining individual predictions (local explanations).
- Model Integration: This categorization highlights whether the framework operates post hoc, analyzing a model after training, or is intrinsically integrated during training.
- Model Compatibility: This attribute specifies whether the framework is model-agnostic, applicable to various AI models, or model-specific, designed for a particular model type (potentially offering deeper insights but limited applicability).

Understanding these attributes is crucial for selecting the most suitable XAI framework. The information provided in the table not only empowers researchers and practitioners but also offers broader benefits for the XAI field itself. The table serves as a valuable benchmark

for the current XAI landscape, showcasing the diversity of the available frameworks and potentially highlighting areas where further development is needed. This comparison can guide the development of new, more comprehensive, or user-friendly approaches to explainability. Furthermore, by comparing different XAI frameworks, the table emphasizes the importance of explainability in AI development and deployment. It highlights the need for transparency and understanding in AI models, which can foster trust and responsible AI practices. The table can also act as a springboard for the collaboration between the researchers and practitioners in the XAI field. Providing a shared understanding of various XAI frameworks can facilitate discussions and encourage the development of new even more effective approaches to explainability.

Table 3. XAI categories on recent frameworks.

		XAI Categories		
Year	**XAI Frameworks**	**Extrinsic (Post Hoc)/Intrinsic**	**Model-Agnostic/Model-Specific**	**Global/Local**
2024	TNTRules [29]	Extrinsic	Agnostic	Global
2022	ClAMPs [30]	Extrinsic	Agnostic	Both
2021	LOREs [31]	Extrinsic	Agnostic	Local
2021	DLIME [32]	Extrinsic	Specific	Global
2021	OAK4XAI [33]	Extrinsic	Specific	Local
2021	TreeSHAPs [34,35]	Extrinsic	Agnostic	Global
2021	DALEX [36]	Extrinsic	Agnostic	Local
2020	CEM [37]	Extrinsic	Agnostic	Local
2020	Alteryx Explainable AI [38]	Extrinsic	Agnostic	Global
2019	GraphLIMEs [39]	Extrinsic	Agnostic	Local
2019	Skater [1]	Extrinsic	Agnostic	Global
2019	CasualSHAPs [40]	Extrinsic	Agnostic	Global
2019	Explain-IT [41]	Extrinsic	Agnostic	Local
2018	Anchors [42]	Extrinsic	Agnostic	Local
2018	Captum [43]	Intrinsic	Agnostic	Global
2018	RISE [44]	Extrinsic	Agnostic	Local
2018	INNvestigate [45,46]	Extrinsic	Specific	Global
2018	interpretML [47]	Extrinsic	Agnostic	Global
2017	SHAPs [48]	Extrinsic	Agnostic	Both
2017	GRAD-CAM [49]	Extrinsic	Specific	Local
2017	Kernel SHAPs [48]	Extrinsic	Agnostic	Global
2017	Integrated Gradients [50,51]	Extrinsic	Agnostic	Global
2017	DeepLIFT [52,53]	Extrinsic	Agnostic	Global
2017	ATTN [54]	Extrinsic	Specific	Local
2017	TCAV [55]	Extrinsic	Agnostic	Local
2016	LIMEs [56]	Extrinsic	Agnostic	Local

Table 3. *Cont.*

		XAI Categories		
Year	**XAI Frameworks**	**Extrinsic (Post Hoc)/Intrinsic**	**Model-Agnostic/Model-Specific**	**Global/Local**
2016	LRP [57]	Extrinsic	Agnostic	Local
2016	What-IF Tool [58]	Extrinsic	Agnostic	Local
2016	AIX360 [59]	Intrinsic	Agnostic	Global
2016	EBM [60]	Extrinsic	Specific	Both
2015	Eli5 [59,61]	Extrinsic	Agnostic	Local

4.1. XAI Frameworks

4.1.1. LIMEs

Local interpretable model-agnostic explanations (LIMEs) introduce a novel approach to delivering explanations at an individual level for predictions. They operate in an extrinsic manner, meaning they provide explanations post hoc without delving into the inner workings of the model. This framework is particularly notable for its model-agnostic nature, capable of explaining predictions across diverse models without needing access to their internal mechanisms. To unravel the contributions of interpretable input to predictions, LIMEs employ a perturbation technique within the input's neighborhood and observe the resulting behavior of the model's predictions. This process involves creating a new dataset comprising perturbed samples and their corresponding predictions from the black-box model. Subsequently, LIMEs utilize this new dataset to train an interpretable model, which is weighted based on the proximity of the sampled instances to the instance under examination [28,56].

4.1.2. GraphLIMEs

GraphLIMEs [39] build upon the foundational concept of LIMEs but extend the application to a specialized type of neural network architecture known as graph neural networks (GNNs). Unlike the traditional models, GNNs are designed to handle non-Euclidean data organized in a graph structure [39], making them well-suited for tasks such as node classification, link prediction, and graph classification. Similar to LIMEs, GraphLIMEs aim to derive an interpretable model, specifically the Hilbert–Schmidt Independence Criterion (HSIC) Lasso model, to explain the individual nodes within the input graph. In the training process of GNNs, non-linear aggregation and combination methods utilize neighboring node features to determine each node's representative embedding. This embedding is crucial for tasks like node classification and graph classification, where sorting nodes or graphs into different classes is essential [39,62].

4.1.3. SHAPs

SHapley Additive exPlanations (SHAPs) comprise a framework designed with a clear objective: to elucidate the prediction of a given instance by evaluating the contribution of each feature to that prediction. Similar to LIMEs, SHAPs operate under a local-based, post hoc, and model-agnostic approach. The technique employed by SHAPs leverages coalitional game theory to compute Shapley values, with each feature value of a data instance acting as a coalition member. These Shapley values provide insights into the fairness of the prediction distribution across different characteristics. Unlike LIMEs, SHAPs do not necessitate the establishment of a local model; instead, they utilize the same function to calculate Shapley values for each dimension [28,48].

4.1.4. Anchors

The Anchors Approach, as outlined in [28], identifies a decisive rule, termed an "anchor", which sufficiently explains the individual predictions generated by any black-box

classification model. When alterations in other feature values have no impact on the prediction, this rule serves as an anchor. Anchors streamline the process by merging reinforcement learning methods with a graph search algorithm, thereby minimizing the number of model calls required. The resultant explanations are articulated as straightforward IF–THEN rules, referred to as anchors. This framework adheres to a local-based, post hoc, and model-agnostic paradigm [28,42].

4.1.5. LOREs

LOcal Rule-based Explanations (LOREs), as introduced in [31], construct an interpretable predictor tailored to a specific black-box instance. This approach employs a decision tree to train the local interpretable predictor using a comprehensive set of artificial cases. Leveraging the decision tree enables the extraction of a localized explanation, comprising a solitary choice rule and an array of counterfactual rules for the reversed decision. Notably, this framework follows a local-based, post hoc, and model-agnostic methodology [28,31].

4.1.6. Grad-CAM

Gradient-weighted Class Activation Mapping (GRAD-CAM), as described in [49], is a method aimed at generating a class-specific heatmap from a single image. It produces a localization map that is discriminative for the class under consideration. This technique capitalizes on the feature maps generated by the final convolutional layer of a CNN. Notably, GRAD-CAM operates under a local-based post hoc framework, although it is model-specific in nature. By utilizing the class-specific gradient information from the final convolutional layer, Grad-CAM produces a coarse localization map highlighting the important regions in an image for classification tasks. This method enhances the transparency of CNN-based models, enabling a better understanding of the image classification processes [18,28,49]

4.1.7. CEM

The Contrastive Explanation Method (CEM) is a technique in the field of XAI that aims to provide explanations for the predictions made by ML models. The CEM, introduced in [37], offers explanations tailored for classification models. Specifically, it identifies the features necessary for predicting the same class as the input instance and determines the minimal alterations required to associate the input instance with a different class. The CEM operates under a local-based post hoc framework while remaining model-agnostic [37].

4.1.8. LRP

Layer-wise Relevance Propagation (LRP) [63] is an explanation method based on propagation, which necessitates access to the model's internals (including topology, weights, activations, etc.). Despite requiring additional information about the model, this enables LRP to streamline and consequently more efficiently address the explanation task. Specifically, LRP does not explain the prediction of a deep neural network in a single step, as model-agnostic methods would do. Instead, it leverages the network structure and redistributes the explanatory factors, referred to as relevance (R), layer by layer. This process initiates from the model's output and propagates onto the input variables (e.g., pixels) [57].

4.1.9. ELI5

ELI5, or Explain Like I'm 5, is an approach in XAI that aims to provide understandable insights into why a model makes specific predictions. It simplifies complex ML models into language that is easy for non-experts to grasp. In Python, ELI5 is implemented through the ELI5 package, which offers a range of tools for visualizing and debugging various ML models in terms of both the intrinsic and extrinsic explainability. It supports model-agnostic and model-specific explanations, providing global and local insights through text and visualizations, catering to various ML tasks in Python. It supports both white-box models like linear regression and decision trees, as well as black-box models such as those

from Keras, XGBoost, and LightGBM. ELI5 works for both regression and classification tasks, making it a versatile tool for interpreting ML models [18,61].

4.1.10. What-If Tool

The What-If Tool (WIT) is a user-friendly feature of Google's TensorBoard web application, aimed at simplifying the debugging and evaluation of ML models. It offers intuitive visualization and exploration capabilities for both classification and regression models, making it accessible to ML researchers, developers, and non-technical stakeholders. The WIT enables the analysis of different what-if scenarios and provides insights into model behavior in real-time. Its three tabs, Data-point Editor, Performance & Fairness, and Feature, offer comprehensive features such as custom visualization, performance evaluation metrics, and dataset feature balance assessment. Overall, the WIT is a versatile and accessible tool for understanding and enhancing ML models, facilitating transparent model analysis [58,64].

4.1.11. AIX360

AI Explainability 360 or AIX 360 stands out as a popular toolkit with the primary objective of offering a straightforward, flexible, and unified programming interface, coupled with an associated software architecture. This architecture addresses the gap in understanding explainability techniques, which is crucial for various stakeholders. The toolkit aims to facilitate communication between data scientists, who may possess knowledge of AI explainability, and algorithm developers. The programming interface closely resembles Python model development tools like sci-kit-learn, utilizing structured classes to expose standard methods to users for explaining the data, model, and prediction components [65].

4.1.12. Skater

Skater is a widely used XAI framework integrated into the Python library, utilized for discerning and interpreting the relationships between data and the features employed in testing models and predictions. The primary goal of Skater is to unveil both the global and local interpretations of black-box models, facilitating a more efficient understanding of the interpreted data. With Skater, users can conduct global interpretations using partial dependence plots and feature importance techniques. Additionally, Skater can gauge how a model's performance changes over time following the deployment in a production environment. The framework possesses the capability to interpret and allows practitioners to assess how the feature interactions vary across different platforms [1,61].

4.1.13. Captum

Captum is a PyTorch-based library designed to enhance model interpretability. It offers insights into models through various pre-built methods included in the Captum library, along with valuable visualization techniques. The interpretability methods provided by Captum are categorized into three groups that assess different aspects of ML model predictions: primary attribution, layer attribution, and neuron attribution. Primary and neuron attribution assess the features present in the data but to varying stages of the model. Primary attribution examines the features in the input data, while neuron attribution evaluates the features within a hidden neuron in the network. The features in an image could represent the specific characteristics of an object that the model is trained to classify. Layer attribution differs from these methods as it evaluates the impact of each neuron within a specific layer [43].

4.1.14. DLIME

The Deterministic Local Interpretable Model-Agnostic Framework (DLIME), proposed in 2023, aims to combine global and regional explanations for diverse ML models. It accomplishes this through a unique approach that leverages game theory concepts and feature attributions. Still in its early stages, the DLIME presents a novel perspective on

Explainable AI, potentially offering a more comprehensive analysis of model behavior. However, further development, testing, and real-world application are crucial to fully understand its effectiveness and future potential. As the field of XAI continues to evolve, staying informed about the DLIME's development can be valuable for those interested in its future applications [32]

4.1.15. EBM

Explainable Boosting Machines (EBMs) represent an interpretable ML method designed to aid in comprehending the relationship between the input and output variables in a prediction process. The EBM algorithm is akin to popular ML models like XGBoost and random forest (RF) as it employs an advanced ensemble learning strategy to train a specialized model. The EBM approach has found successful application and implementation in diverse fields such as geoengineering, environmental assessment, and healthcare. The EBM algorithm provides excellent performance in predicting compressive strength [60].

4.1.16. RISE

The XAI framework RISE (Randomized Input Sampling for Explanation of Black-box Models) focuses on generating explanations, particularly interpretable heatmaps, for image classification tasks. It utilizes a randomized input sampling approach to probe complex deep learning models, often considered "black boxes", and identify the relevant regions within an image contributing to the model's prediction. While initially developed for image classification, the ongoing research in the field of XAI might lead to future developments or alternative methods in both RISE and other frameworks [44].

4.1.17. Kernel SHAPs

The XAI framework, Kernel SHAPs, focuses on explaining the individual predictions for various models with numerical outputs. It utilizes local explanations to delve into how each feature value contributes to a specific prediction, offering insights into the model's decision-making process for a particular instance. Unlike SHAPs, which rely on linear models, Kernel SHAPs employ kernels to handle non-linear feature relationships, making them well-suited for explaining complex models. Notably, they can be applied to diverse model types, including deep learning architectures, and leverage kernels and Shapley values to provide interpretable explanations [48].

4.1.18. Integrated Gradients

Integrated Gradients (IGs) is an XAI framework delving into individual predictions completed by complex models, particularly deep learning models. It sheds light on the model's decision-making process for specific instances by attributing the final prediction to each input feature value. This enables understanding how much each feature contributes to the model's output. IGs leverage the concept of gradients, which indicate how the model's output changes as a specific input feature changes. By calculating the gradients along a path from a baseline (no information) to the actual input, IGs estimate the contribution of each feature. This information can be visualized through heatmaps, where regions with higher attributions contribute more significantly to the prediction. Integrated Gradients offer a valuable approach to understanding the individual predictions in complex models. Notably, its versatility, intuitive visualizations, and interpretability contribute to the approach's popularity in the XAI field [50,51].

4.1.19. TNTRules

TUNE-NOTUNE Rules (TNTRules) is a post hoc rule-based explanation approach for optimization tasks, particularly in the context of Bayesian optimization (BO) for parameter tuning. It aims to generate high-quality explanations through a multiobjective optimization approach [29]. The proposed method aims to address the limitations of the

current Explainable AI methods in optimization tasks, potentially reducing the biases in the tuning process.

4.1.20. ClAMPs

The XAI framework ClAMPs (Cluster Analysis with Multidimensional Prototypes) is an approach for explaining clustering. It generates human-readable rule-based explanations for cluster analysis. These explanations combine the strengths of local explanations with the generality of global ones, allowing for a better understanding of the cluster shapes and distributions. The approach utilizes random, isolation forest, and K-D trees-based approaches. The study participants found that the ClAMPs method enables better descriptions of clusters and helps in understanding clusters well, particularly when applied to artificially generated datasets. This means it is potentially applicable to diverse model types [30].

4.1.21. OAK4XAI

OAK4XAI is an XAI framework that leverages knowledge maps and ontologies. Unlike the existing XAI methods that focus solely on explaining feature contributions, OAK4XAI delves deeper. It incorporates domain-specific knowledge through ontologies (formal descriptions of concepts) and knowledge maps (visual representations of relationships). This approach, exemplified by the Agriculture Computing Ontology (AgriComO), aims to provide not just explanations for results but also a clear understanding of the underlying concepts, algorithms, and values used by the models. By bridging the gap between data analysis and semantic understanding, OAK4XAI empowers users in agriculture to not only trust the models' outputs but also grasp the reasoning behind them [33].

4.1.22. TreeSHAP

TreeSHAPs (Tree SHapley Additive exPlanations) are designed to explain the predictions produced by tree-based algorithms for ML, such as decision trees and random forests. While the model is built on the SHAP architecture, its main purpose is to effectively manage the structure of models that are tree-based. TreeSHAPs effectively navigate decision trees bottom-up to assign contributions to each feature by computing Shapley values, which indicate each feature's contribution to a model's prediction. Because of their high computational efficiency, TreeSHAPs can handle enormous datasets and intricate models with ease. In the end, TreeSHAPs offer comprehensible justifications by assigning feature significance ratings to every input feature. This facilitates the identification of the features that have the most significant influence on the model predictions and promotes confidence and openness in AI systems [34,66].

4.1.23. CasualSHAPs

An approach widely used in XAI to explain the individual predictions made by ML models is called CasualSHAPs. Every feature is assigned a significance value that indicates how much it contributes to the model's prediction. The Shapley value notion, in particular, is the foundation of cooperative game theory and SHAP values. Conversely, CasualSHAPs integrate causal inference methods with SHAP values. Beyond simple correlations, causal inference seeks to comprehend cause-and-effect linkages in data. CausalSHAPs aim to offer interpretations that not only emphasize the significance of characteristics but also uncover the connections between features and projections by integrating causal inference into SHAPs [40].

4.1.24. INNVestigate

INNvestigate is a Python library and framework designed for the interpretability and analysis of neural network models. It stands for "INNermost layer inVestigation And VisualizaTion Engine". This framework provides a suite of tools and methods for

understanding the inner workings of neural networks, particularly deep learning models, in various applications, such as image classification and natural language processing [45].

4.1.25. Explain-IT

Explain-IT sheds light on the often-opaque world of unsupervised learning algorithms. Unlike supervised learning with labeled data, unsupervised learning deals with uncategorized data. Explain-IT tackles this challenge by making the knowledge discovery process transparent. It first transforms the data for exploration, optionally incorporating expert knowledge to guide the search for patterns. Then, it groups similar data points through clustering, revealing hidden structures. Finally, Explain-IT leverages LIMEs, an XAI technique, to understand the factors that influence these groupings. This allows Explain-IT to interpret the content of each cluster, providing local explanations for why certain data points were grouped together. Explain-IT is a comprehensive tool that unveils the secrets of unsupervised learning and makes its outcomes more understandable [41].

4.1.26. DeepLIFT

The Explainable AI (XAI) approach known as DeepLIFT (Deep Learning Important FeaTures) is used to assign the inputs of input characteristics to the results produced by a neural network model [67,68]. By evaluating each neuron's stimulation within the network regarding a reference activation state usually selected as a baseline state like zero or the average activation it functions, DeepLIFT rates each feature according to its significance in terms of the model's predictions by calculating the differential in activations between the present state and the reference condition. By revealing which information properties are significant in shaping the model's judgments, these significance scores improve the ability to understand and comprehend neural network activity. DeepLIFT is very useful in fields like image analysis, financial forecasting, and healthcare diagnostics, where it is essential to understand feature contributions [53,69].

4.1.27. ATTN

In Explainable AI (XAI), attention (ATTN) refers to the use of attention mechanisms, which are prevalent in deep learning models, to improve the understandability of those models rather than a particular approach. These models may focus on significant portions of the input data thanks to attention. Researchers utilize this concentration in XAI to elucidate the logic behind the model. We can learn more about the aspects that most affect the model's output by looking at the sections that attract the most interest. This can enhance the transparency in activities such as healthcare natural language processing (NLP) by emphasizing the sentences that have the most influence on a diagnosis. Although still in its early stages of development, attention-based XAI shows promise and is being actively investigated by researchers as a possible explanation for AI decision-making [70].

4.1.28. Alteryx Explainable AI

XAI is the focus of an integrated set of technologies offered by Alteryx. Alteryx emphasizes usability and accessibility and provides a range of ways to explain ML models that are integrated into the system it operates in. These consist of individual conditional expectation (ICE) plots, partial dependence plots (PDP), and feature significance analyses. Alteryx seeks to provide users with actionable insights into their models' decision-making processes by customizing explanation methodologies based on the kind of model, promoting open communication and trust in AI solutions [38].

4.1.29. InterpretML

An open-source toolkit called InterpretML was created to make XAI jobs more accessible for a variety of ML model types. The wide range of explainability strategies, including LIMEs, SHAPs, PDP, and InterpretML, provide users the freedom to select the approach that best fits their unique requirements. InterpretML enables academics, develop-

ers, and practitioners to break through the mystery surrounding ML models and obtain a better understanding of their behavior by democratizing the access to sophisticated XAI techniques [47].

4.1.30. DALEX

(The Explainable AI Library by DARPA) At the vanguard of Explainable AI research, DALEX emerges as an open-source toolkit that offers a vast array of algorithms tailored to different XAI applications. This extensive collection includes both model-independent techniques and explanations designed specifically for deep learning models. DALEX facilitates transparency, reliability, and trust in AI-driven decision-making procedures by providing a broad variety of tools that allow users to analyze and comprehend the inner workings of complicated ML algorithms [36].

4.1.31. TCAV

TCAV (Testing with Concept Activation Vectors) is specifically designed to comprehend convolutional neural networks' (CNNs') decisions in image classification tasks. TCAV offers a gradient-based explanatory methodology. TCAV determines the underlying causes influencing the final categorization by examining the concepts the model learned and their activations in particular layers. By emphasizing the clarification of acquired ideas, TCAV provides insightful information on CNN decision-making procedures, improving the clarity and comprehension in recognition of image techniques [55].

Table 4 also explores various attributes regarding the 31 XAI frameworks, which might be of value to some researchers and practitioners.

4.2. Comparative Analysis

This section delves into three key attributes of XAI frameworks, providing a comprehensive picture of their capabilities. First, we will leverage data from Tables 3 and 4 to perform a comparative analysis of the output types across various XAI frameworks. This will shed light on how these frameworks explain the inner workings of ML models. Second, we will analyze the model complexity that different XAI frameworks can handle. This will help us to understand which frameworks are best suited for explaining simpler models versus complex ones, particularly deep learning models. Finally, we will explore the diverse application domains (use cases) where XAI frameworks are instrumental in aiding humans in understanding the often-opaque nature of ML models, often referred to as "black boxes". This three-pronged approach will equip readers with a thorough understanding of how XAI frameworks empower us to interpret and gain insights from complex machine learning models. Figure 2 highlights our understanding of the XAI framework that incorporates key attributes, application domains, and the common challenges involved with popular frameworks.

XAI frameworks use various output types to provide insights into an ML model's decision-making process. Table 5 highlights and shows the output formats for different XAI frameworks. Textual explanations offer a user-friendly format for understanding the general logic behind the model's choices. However, textual explanations can become verbose and lack nuance for complex models.

Visualizations provide a more intuitive approach by using charts and graphs to represent the importance of features and their impact on predictions. While visualizations excel at identifying key factors, they might become overwhelming for models with intricate relationships. Feature importance offers a quantitative measure that helps in understanding how different data points contribute to the model's output. However, this type of output might not explain "why" a feature is important, or how the features interact with each other. Finally, arguments provide additional context and justification for the model's predictions. These can include counterfactual explanations (showing how changing an input would change the prediction) or specific examples that support the model's reasoning. The choice of output type depends on the specific use case and the user's needs.

Table 4. Comparative analysis of XAI framework against several attributes.

Year	XAI Framework	Output Format	Use Cases	Programming Language	ML/DL	Model Complexity	Open Source/ Commercial	Licence Type
2024	TNTRule [29]	Textual Rules, Visualizations	Optimization Tasks for Parameter tuning of cyber–physical System	Python	Unsupervised ML	Flexible	Open Source	Apache 2.0
2022	ClAMPs [30]	Human-readable Rules, Visualization	unsupervised problem to human readable rules	Python	UnSupervised ML, scikit-learn, tensorFlow and PyTorch	Flexible	Open Source	Apache 2.0
2021	LOREs [31]	Decision and Counterfactual Rules	Understanding Individual Predictions	Python	Various	Simple to Complex	Open Source	MIT
2021	DLIME [32]	Text, Visualization	Understanding Model, behavior, identifying Bias, debugging models	Python	Primarily TensorFlow models	Flexible	Open Source	Apache 2.0
2021	OAK4XAI [33]	Text, factual grounding highlights	Understanding the reason behind generated outputs, identifying influential features, analyzing model behavior	Python	Various	Simple to Complex	Open Source	N/A
2021	TreeSHAPs [34,35]	Text, Visualizations (force Plots, dependence Plots)	Understanding feature importance, identifying influential features, analyzing model behavior	Python	Decision trees and Random Forests	Flexible	Open Source	MIT

Table 4. *Cont.*

Year	XAI Framework	Output Format	Use Cases	Programming Language	ML/DL	Model Complexity	Open Source/ Commercial	Licence Type
2021	DALEX [36]	Visual Explanations including interactive what-if plots and feature attribution heatmaps	Understanding the reasoning behind specific model predictions, identifying influential factors in individual cases.	Python	TensorFlow, PyTorch and scikit-learn	Simple to Complex	Open Source	Apache 2.0
2020	CEM [37]	Counterfactual Examples	Understanding individual Predictions, identifying Biases	Python	Various	Simple to Complex	Open Source	Apache 2.0
2020	Alteryx Explainable AI [38]	Primarily visual explanation through interactive dashboards within the Alteryx platform	Understanding model behavior, identifying influential features, debugging models, and improving model performance	Python or R	Agnostic or Specific to Model	Varying Complexity from simple linear model to complex deep learning models	Commercial	Proprietary
2019	Explain-IT [41]	Visualization, feature Contributions	Understanding individual Predictions, identifying feature contribution	Python	Unsupervised ML	from simple to Complex	N/A	N/A
2019	GraphLIMEs [39]	Feature contributions, Textual explanations, Visualizations	Understanding individual predictions in graph data	Python	Graph ML	Simple to complex	Open Source	Apache 2.0

Table 4. *Cont.*

Year	XAI Framework	Output Format	Use Cases	Programming Language	ML/DL	Model Complexity	Open Source/ Commercial	Licence Type
2019	Skater [1]	Text, Visualizations	Understanding Model behavior, identifying bias, debugging models	Python	Framework Agnostic	Flexible	Open Source	Apache 2.0
2019	CausalSHAPs [40]	Feature Importance with Visualizations	Understanding causal relationships between features and predictions, identifying features with the strongest causal effect	Python	Primarily Focused on tree-based models	Flexible	Open Source	Apache 2.0
2018	Anchors [42]	Feature Importance , Anchor set	Understanding individual predictions, debugging models	Python	ML	Simple to Complex	Open Source	Apache 2.0
2018	Captum [43]	Text, Visualizations (integrated with various libraries)	Understanding Model behavior, identifying bias, debugging models, counterfactual explanations	Python	Framework Agnostic, various PyTorch Models	Flexible	Open Source	Apache 2.0
2018	RISE [44]	Visualizations (Saliency maps)	Understanding local feature importance in image classification models	Python	Primarily focused on Image classification models	Flexible	Open Source	MIT
2018	INNvestigate [45,46]	Text, Visualizations (integrated with Matplotlib)	Understanding model behavior, identifying important features, debugging models	Python	Keras models	Flexible	Open Source	Apache 2.0

Table 4. *Cont.*

Year	XAI Framework	Output Format	Use Cases	Programming Language	ML/DL	Model Complexity	Open Source/ Commercial	Licence Type
2018	InterpretML [47]	Textual explanations, and visualizations (including feature importance plots, partial dependence plots, and SHAP values)	Understanding model behavior, identifying influential features, debugging models, and improving model performance	Python	Agnostic to specific frameworks	Can handle models of varying complexity	Open Source	Apache 2.0
2017	SHAP [48]	Feature importance, Shapley values	Understanding individual predictions, Debugging models	Python	Various	Simple to complex	Open Source	MIT
2017	Grad-Cam [49]	Heatmaps highlighting important regions	Understanding individual predictions in CNNs	Python	CNNs	Simple to complex	Open Source	Varies
2017	Kernel SHAP [48]	Text, Visualizations (force plots, dependence plots)	Understanding feature importance, identifying bias, debugging models	Python	Framework agnostic	Flexible	Open Source	Apache 2.0
2017	Integrated Gradients [50]	Saliency maps or attribution scores	Understanding feature importance, identifying influential factors in model predictions, debugging models.	Varies	Varies	Varies	Open Source	Varies

Table 4. *Cont.*

Year	XAI Framework	Output Format	Use Cases	Programming Language	ML/DL	Model Complexity	Open Source/ Commercial	Licence Type
2017	DeepLIFT [67,68]	Textual explanations, visualizations (heatmaps for saliency scores)	Understanding feature importance, identifying influential regions in input data, debugging models	Python	TensorFlow, PyTorch, and Keras	Designed for complex deep learning models	Open Source	Apache 2.0
2017	ATTN [54]	Visualizations (heatmaps highlighting attention weights)	Understanding how transformers attend to different parts of the input sequence, identifying which parts are most influential for specific predictions, debugging NLP models	Python	TensorFlow and PyTorch	Primarily designed for complex deep learning models like transformers	Open Source	Varies
2017	TCAV [55]	Visualizations (heatmaps highlighting concept activation regions)	Understanding which image regions contribute to specific model predictions, identifying biases in image classification models	Python	TensorFlow, PyTorch, and sci-kit-learn	Designed for complex deep learning model image classification	Open Source	Apache 2.0
2016	LIMEs [56]	Arguments, Text, Visualizations	Understanding individual predictions, debugging models, Identifying biases, Image Classification, Text Classification, Tabular Data Analysis	Python	TensorFlow, PyTorch, and Keras	Simple to complex	Open Source	BSD 3-Clause

Table 4. *Cont.*

Year	XAI Framework	Output Format	Use Cases	Programming Language	ML/DL	Model Complexity	Open Source/ Commercial	Licence Type
2016	LRP [57]	Relevance scores	Understanding individual predictions in deep neural networks	Python	Deep neural networks	Simple to complex	Open Source	Varies
2016	What-IF Tool [58]	Feature contributions, Visualizations	Exploring hypothetical scenarios, Debugging models	Python/ TensorFlow/ JavaScript	Various	Simple to complex	Discontinued	N/A
2016	AIX360 [65]	Text, Visualizations	Understanding model behavior, identifying bias, debugging models	Python	Framework agnostic	Flexible	Open Source	Apache 2.0
2016	EBMs [60]	Text, Visualizations (partial dependence plots, feature importance)	Understanding model behavior, identifying feature interactions, debugging models	Python	Gradient boosting models	Flexible	Open Source	Varies
2015	ELI5 [71]	Textual explanations, Arguments, Visualizations	Understanding individual predictions, Image Classification, Text Classification, Tabular Data Analysis	Python	Various	Simple to complex	Open Source	Varies

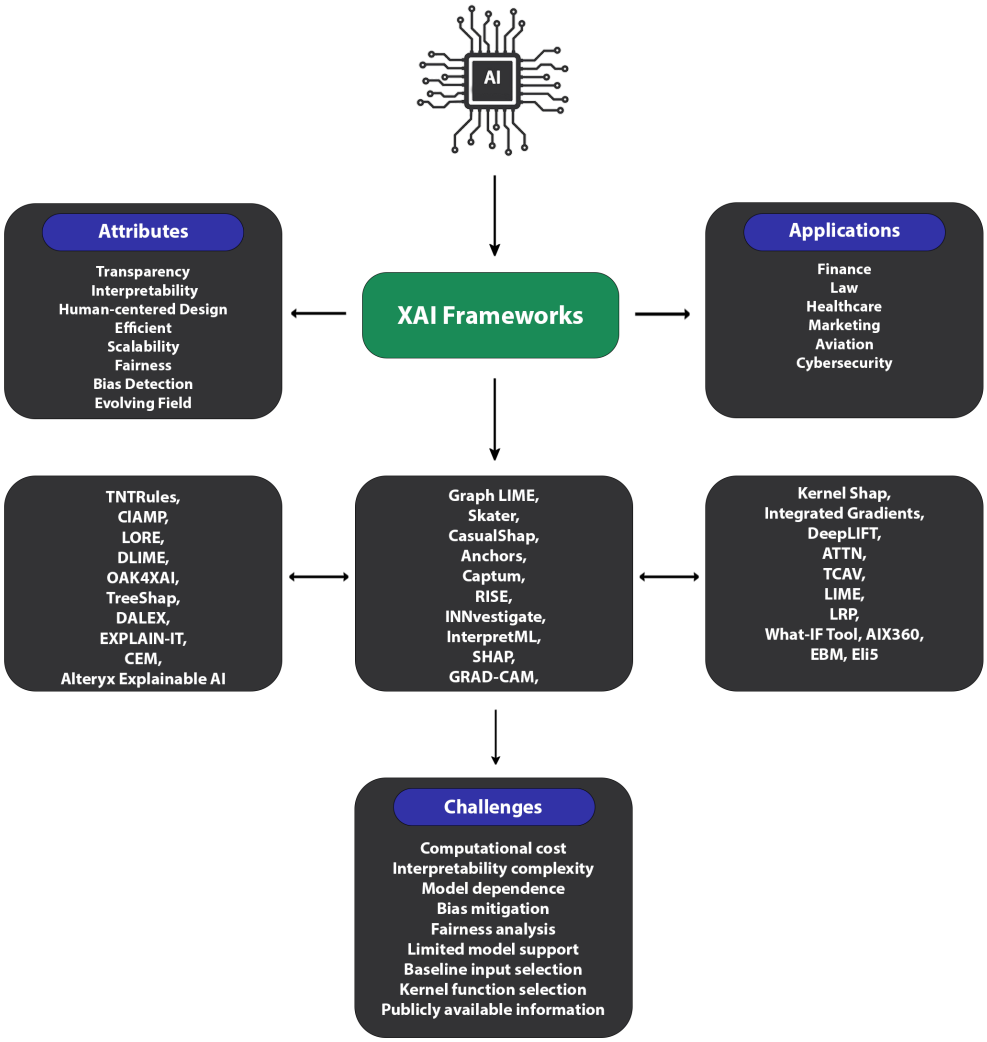

Figure 2. A standardized XAI framework and its components.

Table 5. A comparative analysis of XAI framework output formats.

XAI Framework	Text	Visualization	Feature Importance	Arguments
TNTRules	✓	✓		
CIAMP	✓	✓		
LOREs	✓			✓
DLIME	✓	✓		
OAK4XAI	✓	✓		
TreeSHAPs	✓	✓		
DALEX	✓		✓	
CEM				✓
Alteryx		✓		
GraphLIMEs	✓	✓	✓	
Skater	✓	✓		

Table 5. *Cont.*

XAI Framework	Text	Visualization	Feature Importance	Arguments
CasualSHAPs	✓	✓		
Explain-IT		✓	✓	
Anchors			✓	✓
Captum	✓	✓		
RISE		✓		
INNvestigate	✓			
InterpretML	✓	✓	✓	
SHAPs			✓	✓
GRAD-CAM				✓
Kernel SHAPs	✓	✓		
Integrated Gradients				✓
DeepLIFT	✓	✓		
ATTN		✓		
TCAV		✓		
LIMEs	✓	✓		
LRP				✓
What-IF Tool		✓	✓	
AIX360	✓	✓		
EBM	✓	✓	✓	
Eli5	✓	✓		✓

The information in Table 6 enables a more precise analysis of how XAI frameworks handle model complexity. The table categorizes these frameworks into flexible, simple to complex, and complex.

Flexible frameworks are likely versatile tools, adept at explaining models across a spectrum of complexity, from basic to intricate. These frameworks might offer a comprehensive range of explanation techniques (text, visualization, feature importance, or arguments) that can be tailored to the specific model being analyzed. The frameworks categorized as simple to complex might be the most effective for explaining models that range from relatively straightforward to moderately complex. These frameworks could balance interpretability and the ability to handle some level of model intricacy. Finally, complex frameworks are likely designed to tackle the challenge of explaining highly complex models, particularly deep learning models with numerous layers and non-linear relationships between features and predictions. These frameworks might employ more advanced techniques specifically suited to provide insights into these intricate models.

It is important to remember that the specific capabilities of each XAI framework can vary. To obtain a complete picture, it is always recommended to identify the strengths and weaknesses of each framework, particularly their suitability for different model complexities. We will be looking into this in Section 5.

In various domains, such as finance, healthcare, cybersecurity, law, marketing, and aviation, the adoption of XAI frameworks plays a pivotal role in enhancing the transparency, interpretability, and decision-making processes. XAI frameworks provide valuable insights into model predictions, aiding practitioners in understanding the underlying logic and factors influencing the outcomes.

Table 6. A comparative analysis of model complexity of various XAI frameworks.

XAI Framework	Flexible	Simple to Complex	Complex
	TNTRules	LOREs	OAK4XAI
	CIAMP	DALEX	InterpretML
	DLIME	CEM	DeepLIFT
	TreeSHAPs	Alteryx	ATTN
	Skater	GraphLIMEs	TCAV
	CausalSHAPs	Explain-IT	
	Captum	Anchors	
	RISE	SHAPs	
	INNvestigate	GRAD-CAM	
	Kernel SHAPs	Integrated Gradients	
	AIX360	LIMEs	
	EBM	LRP	
		What-if	
		Eli5	

4.2.1. Finance

In the finance sector, where accurate risk assessment and regulatory compliance are critical for decision-making, the choice of suitable XAI frameworks holds immense importance. Among the notable frameworks, SHAPs, Explain-IT, and Integrated Gradients stand out for their tailored solutions to the unique challenges encountered in financial analysis. SHAPs' capability to elucidate the contribution of individual features to predictions aligns perfectly with the needs of risk assessment, offering clarity on the factors influencing portfolio performance and financial risk. The DLIME's speed and scalability are particularly advantageous for managing extensive and dynamic financial datasets, ensuring efficient interpretation without compromising accuracy. Meanwhile, Integrated Gradients' visualization of feature attributions aids in understanding the underlying logic behind model predictions, facilitating transparent decision-making processes.

Extensive research on AI applications in finance [72–74] underscores the growing significance of technological advancements in financial technology (FinTech). While [75] provides a comprehensive overview of the AI solutions in finance, the analysis in [73] identifies thematic clusters, and [74] explores financial intelligence, highlighting its role as the third stage of technological advancement in finance. These studies collectively emphasize the importance of XAI in addressing the challenges related to risk management, stock market analysis, portfolio optimization, anti-money laundering, and electronic financial transaction classification within the finance sector. Each of these frameworks, with its distinct methodologies and strengths, empowers financial analysts and institutions with actionable insights, facilitating informed decision-making and regulatory compliance [76].

4.2.2. Healthcare

In the healthcare sector, where patient safety and regulatory compliance are paramount, the selection of appropriate XAI frameworks is critical. LIMEs, GraphLIMEs, and DeepLIFT stand out as the top contenders, each offering tailored solutions to address the unique challenges of healthcare decision-making. LIMEs' ability to provide explanations at an individual level facilitates the interpretation of model predictions for specific patients, aiding healthcare professionals in making informed decisions regarding diagnosis and treatment. GraphLIMEs' extension to graph neural networks enables the interpretation of complex medical data, enhancing the transparency and reliability of AI-assisted medical diagnoses. Additionally, DeepLIFT's feature importance scores provide valuable insights

into the significance of medical indicators, facilitating the development of accurate and reliable diagnostic models. By leveraging these frameworks, healthcare practitioners can enhance patient care and safety, ensuring compliance with the regulatory standards and best practices.

XAI has been increasingly integrated into the healthcare practices in medicine, where AI has greatly improved medical image analysis. The recent efforts have focused on combining AI's accuracy with enhanced model interpretability. Ref. [77] emphasizes the importance of transparent AI in healthcare decision-making and observed over 200 papers employing XAI in deep learning-based medical image analysis, highlighting the prevalence of visual explanations, such as saliency maps [78], over textual and example-based explanations.

Ref. [79] investigated the application of XAI in psychiatry and mental health, emphasizing the heightened need for explainability and understandability due to the complex probabilistic relationships among syndromes, outcomes, disorders, and symptoms. They proposed the TIFU (Transparency and Interpretability For Understandability) framework, which underscores the importance of making models comprehensible to users through transparency and interpretability. Their study highlights XAI's key roles in the prediction and discovery within mental health, stressing the necessity for understandability in clinical settings where the stakes are high. The authors advocate for AI tools that support clinicians without adding unnecessary complexity.

4.2.3. Cybersecurity

In the domain of cybersecurity within IT, where ensuring system integrity and protecting against cyber threats is paramount, the strategic selection of XAI frameworks is pivotal. Captum, ELI5, and Skater, along with the Anchors Approach, Layer-wise Relevance Propagation, and ClAMPs, emerge as the standout solutions, each offering distinct advantages tailored to the diverse needs of cybersecurity professionals. Captum's advanced attribution methods provide detailed insights into the model behavior, aiding in the identification and mitigation of vulnerabilities within complex IT systems. ELI5's intuitive interface facilitates the effective communication of model insights to non-experts, fostering collaboration and informed decision-making across IT teams. Additionally, Skater's comprehensive interpretations enable cybersecurity professionals to understand feature interactions and optimize the system security with precision. The Anchors Approach's concise IF–THEN rules and Layer-wise Relevance Propagation's relevant insights offer further clarity in terms of threat detection and mitigation. Furthermore, ClAMP's model-agnostic explanations empower cybersecurity professionals to comprehend diverse cybersecurity models and identify potential weaknesses or vulnerabilities, enhancing the overall security resilience. By harnessing the capabilities of these frameworks, cybersecurity practitioners can bolster their organization's defenses, detect threats, and safeguard sensitive information against cyber-attacks.

In [41,80], the authors concentrate on the examination of encrypted traffic, particularly to enhance the accurate detection of DoH (DNS Over HTTPS) attacks. They incorporate Explainable AI techniques using SHAPs, enabling the visualization of individual feature contributions regarding the model's classification decisions. Similarly, EXPLAIN-IT [41] tackles the YouTube video quality classification issue within encrypted traffic scenarios. The methodology deals with unlabeled data, creating meaningful clusters, and providing explanations of the clustering outcomes to end-users. They utilize LIMEs to interpret clusters, employing a local-based strategy. Similarly, ROULETTE [81] focuses on network traffic, specifically employing attention coupled with a multi-output deep learning strategy to better distinguish between the categories of network intrusions. For post hoc explanations, they employ visual explanation maps generated through Grad-CAM.

In [82], a two-stage ML-based Wireless Network Intrusion Detection System (WNIDS) is deployed to enhance the detection of impersonation and injection attacks within a Wi-Fi network. XAI techniques are integrated to provide insights into the decisions made

by the initial ML model, particularly regarding instances predicted as impersonation or injection attacks. By employing SHAPs, the features that exert a significant influence on these predictions are identified. Remarkably, this set of features closely aligns with those pinpointed by the feature selection method utilized for the second-stage ML model.

4.2.4. Legal

In the legal domain, where clarity and transparency are essential for justifiable decisions, the selection of suitable XAI frameworks is critical. The Anchors Approach, CEM, and LOREs emerge as the top contenders, offering tailored solutions to meet the unique challenges faced in legal reasoning and decision-making. The Anchors Approach's succinct IF–THEN rules provide concise explanations for individual predictions, aligning seamlessly with the legal precedents and decision-making frameworks. The CEM's focus on identifying the necessary features and minimal alterations for class prediction offers a systematic approach to understanding the legal outcomes and implications. Additionally, LOREs' localized explanations and counterfactual rules contribute to nuanced interpretations of legal decisions, aiding in the exploration of alternative scenarios and legal arguments. By leveraging these frameworks, legal practitioners can navigate complex legal landscapes with clarity and confidence, ensuring fair and just outcomes. Ref. [83] investigates the application of XAI in the legal domain, an area of interest within the AI and law community. It highlights the gap in the user experience studies concerning XAI methods and the overall concept of explainability. The study evaluates the effectiveness of various explainability methods (Grad-CAM, LIMEs, and SHAPs) in explaining the predictions for legal text classification, with legal professionals assessing the accuracy of these explanations. Additionally, the respondents provide insights into the desired qualities of AI legal decision systems and their general understanding of XAI. This research serves as a preliminary study to explore lawyers' perspectives on AI and XAI, paving the way for more in-depth investigations in the field.

Ref. [84] discusses the role of XAI in ensuring legal compliance, particularly in the context of automated decision-making (ADM) systems governed by regulations such as the GDPR. It introduces the "Explanation Dialogues" study, aiming to understand how legal experts perceive and assess the explanations generated by ADM systems. By focusing on GDPR provisions and expert interviews, the research sheds light on how XAI can facilitate transparency and accountability in legal processes. This interdisciplinary approach underscores the importance of integrating XAI principles within regulatory frameworks to enhance compliance and uphold legal standards.

4.2.5. Marketing

In the marketing sector, where understanding consumer behavior and predicting market trends is crucial for business success, the selection of appropriate XAI frameworks is essential. SHAPs, TreeSHAPs, and Kernel SHAPs emerge as the top contenders, each offering unique strengths to address the diverse needs of marketers. SHAPs' ability to elucidate the contribution of individual features to predictions enables marketers to identify the key factors influencing consumer behavior and tailor their marketing strategies accordingly. TreeSHAPs' capability to handle tree-based models makes it well-suited for interpreting the decision trees commonly used in marketing analytics, providing valuable insights into customer segmentation and targeted advertising. Additionally, Kernel SHAPs' ability to handle non-linear feature relationships enhances their suitability for explaining complex models, enabling marketers to develop more accurate and effective marketing campaigns. By leveraging these frameworks, marketers can gain actionable insights into consumer behavior, optimize marketing strategies, and drive business growth and success.

In a recent study [85] on XAI in marketing, the consumer preferences for explanation attributes were investigated using choice-based conjoint analysis. By integrating marketing theories, the study offers guidance on designing XAI algorithms in marketing strategies, showcasing the potential for future exploration in understanding the impact of explanation

types on consumer behavior and business outcomes to examine the consumer preferences for explanation attributes in XAI within marketing contexts. Computer vision techniques offer valuable insights, but interpretability is crucial for consumer trust. Moreover, ref. [86] explores how XAI methods address this, categorizing them based on transparency and interpretability. It suggests future research to enhance model interpretability, including studying the economic value of augmented reality and testing strategies to mitigate algorithmic bias.

4.2.6. Aviation

In the aviation sector, where safety, efficiency, and reliability are paramount, the selection of appropriate XAI frameworks is crucial. Among the standout frameworks, Captum, SHAPs, GraphLIMEs, Integrated Gradients, and ELI5 offer tailored solutions to address the unique challenges and requirements of aviation operations. Captum's primary attribution and neuron attribution methods provide granular insights into the behavior of the AI models used in aviation systems, aiding in understanding flight safety, route optimization, and aircraft maintenance. SHAPs' ability to elucidate the contribution of individual features to predictions is valuable for understanding complex aviation systems, such as weather conditions, aircraft performance, and air traffic patterns. GraphLIMEs' extension to graph neural networks enables the interpretable analysis of air traffic networks, flight trajectories, and airport operations, facilitating predictions related to flight delays, route planning, and airspace management. The Integrated Gradients approach to attributing the model's output to each input feature aids in understanding the critical parameters affecting flight safety, efficiency, and performance. Finally, ELI5's user-friendly interface and simplicity facilitate collaboration and decision-making among aviation stakeholders, providing clear and understandable explanations of AI-driven processes and predictions in aviation systems. Leveraging these XAI frameworks empowers aviation professionals to enhance the safety, efficiency, and reliability in aviation operations, ultimately ensuring the smooth and secure functioning of the aviation industry.

In aviation lies its ability to provide transparent and interpretable insights into complex AI models, particularly in critical decision-making scenarios. XAI ensures that human users, including pilots, air traffic controllers, maintainers, and aircraft manufacturers, can understand and trust the decisions made by AI systems. By employing a comprehensive XAI taxonomy and explanation framework, ref. [87] aims to enhance the decision-making processes in aerospace by integrating human-centric and scientific explanations. The related work discussed in this paper focuses on the implementation of XAI in aerospace, emphasizing its application in supporting operators such as pilots and air traffic controllers in making informed decisions based on meaningful information within time constraints. The increasing utilization of intelligent Decision Support Tools in Air Traffic Management (ATM) necessitates the assurance of their trustworthiness and safety, particularly with the integration of non-traditional entities like AI and ML. Despite the extensive research on these technologies, a gap exists between the research outcomes and practical implementation due to regulatory and practical challenges, including the need for transparency and explainability. To address these challenges, a novel framework is proposed to instill trust in AI-based automated solutions, drawing from the current guidelines and end-user feedback [88]. Recommendations are provided to facilitate the adoption of AI- and ML-based solutions in ATM, leveraging the framework to bridge the gap between research and implementation.

Figure 3 illustrates the popular application domains and the respective XAI framework that is currently being utilized.

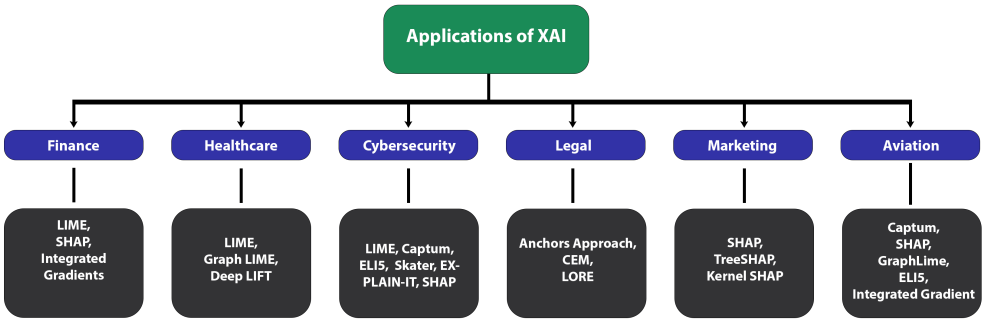

Figure 3. Application domains for different XAI frameworks.

5. Challenges

XAI is characterized by a variety of frameworks, each with unique strengths and limitations (see Table 7). LIMEs provide a straightforward approach, understandable explanations, and is compatible with a variety of models, although it may find it difficult to completely capture complicated model behavior and may be computationally expensive for large datasets [56]. Although computationally demanding, especially for models with large feature sets, SHAPs offer both local and global explanations with a foundation in game theory [89]. Although they depend on rule creation and might not cover every decision boundary, anchors are excellent at producing precise, intelligible rules for humans [42].

Despite their reliance on an external rule generator and potential scalability issues with large datasets, LOREs provide quick results by effectively explaining the decisions using pre-existing rules [31]. Despite being primarily intended for convolutional neural networks (CNNs), Grad-CAM effectively visualizes heatmaps for image classification tasks [49]. Although they provide smooth feature attribution representations, Integrated Gradients can be computationally costly and require the selection of a baseline input [50]. Although it depends on selecting the right reference class, the CEM provides class-specific explanations that may be understood using saliency maps [37]. Notwithstanding the perhaps large computing costs and intricate interpretations, LRP offers thorough attribution ratings for every input characteristic [90,91].

The What-If Tool is a tool that facilitates the interactive study of model behavior through hypothetical situations. However, it is limited to certain model types and requires expertise to develop meaningful explanations [58]. Although AIX360 is primarily concerned with fairness analysis, it does include several tools for identifying and reducing the bias in models [59,65]. It may not offer comprehensive explanations for specific forecasts. Despite its limited model support, Skater allows for counterfactual explanations, offering perspectives on "what-if" circumstances [92]. Although it may require expertise with its API, Captum provides a range of explainability strategies in a single Python module [93].

The DLIME provides a deterministic substitute appropriate for real-time applications, albeit being less interpretable than the original LIMEs and restricted to tabular data [94]. Despite its primary focus on picture classification problems, TCAV produces interpretable visualizations utilizing idea activation vectors [10]. Explainable Boosting Machines (EBMs) are restricted to decision tree-based models, although they help with feature significance analysis [95,96]. Despite focusing mostly on image classification challenges, RISE finds significant picture areas to produce interpretable heatmaps [44]. Although Kernel SHAPs rely on computationally costly kernel functions, it provides flexibility in managing complicated feature interactions [97,98].

ClAMPs and TNTRules seek to explain model predictions; nevertheless, some constraints and difficulties might not be easily accessible [30]. Although it depends on locating optimum anchors, which may be computationally costly, OAK4XAI seeks to develop succinct and understandable anchors using game theory techniques [33]. Although explanations for highly complicated decision trees may become difficult, TreeSHAPs provide clarity on how tree-based algorithms arrive at predictions [99]. By providing academics

and practitioners a wide range of tools for comprehending and evaluating ML models, these frameworks work together to advance XAI.

Several techniques exist to understand how ML models arrive at their predictions. Some methods, like CasualSHAP [100] and INNvestigate [45], delve into the causal relationships between features and outcomes but require additional data and can be computationally expensive. While they provide deeper insights than just correlation, interpreting their inner workings might be challenging. DeepLIFT [67,68] and ATTN [54] focus on explaining the specific aspects of models, like individual neurons or attention mechanisms, but may not be suitable for all tasks [101]. DALEX allows for "what-if" scenarios to see how input changes affect the predictions, but it might have limited documentation [36]. ELI5 offers clear explanations for specific models but has limited applicability [71]. GraphLIMEs excel at explaining models where feature interactions are important but can be computationally heavy [39]. The DLIME [32] and InterpretML [47,102] provide various explanation techniques but require additional setup or might have limitations regarding the model types or user-friendliness. Finally, Alteryx Explainable AI offers a commercial toolkit with visual and interactive tools for interpretability [38].

Table 7. XAI framework challenges and strengths comparison.

XAI Framework	Key Limitations/Challenges	Primary Strengths
LIMEs [56]	May not fully reflect complex model behavior, computationally expensive for large datasets	Simple, intuitive explanations, work with various models
SHAPs [48]	Computationally expensive, explanations can become complex for large feature sets	Grounded in game theory, offers global and local insights
Anchors [42]	Reliance on rule generation may not cover all decision boundaries	Provides high-precision, human-understandable rules
LOREs [31]	Reliance on an external rule generator; may not scale well with large datasets	Can explain decisions using existing rules, offering fast results
Grad-CAM [49]	Primarily for convolutional neural networks (CNNs); they may not work well with other architectures	Efficiently visualizes heatmaps for image classification
Integrated Gradients [50]	Relies on choosing a baseline input, can be computationally expensive	Offers smooth visualizations of feature attributions
CEM [51]	On choosing an appropriate reference class, which may not be intuitive for complex models	Offers class-specific explanations, interpretable through saliency maps
LRP [57]	Computational costs can be high for complex models, explanations can be complex to interpret	Provides comprehensive attribution scores for each input feature
What-If Tool [58]	Requires expert knowledge to design effective explanations, limited to specific model types	Allows interactive exploration of model behavior through hypothetical scenarios
AIX360 [59]	Primarily focused on fairness analysis; may not provide detailed explanations for individual predictions	Offers various tools for detecting and mitigating bias in models
Skater [1]	Limited model support, requires knowledge of counterfactual reasoning	Enables counterfactual explanations, providing insights into "what-if" scenarios
Captum [43]	Requires familiarity with the library's API, not as beginner-friendly as other options	Offers a variety of explainability techniques in a unified Python library

Table 7. *Cont.*

XAI Framework	Key Limitations/Challenges	Primary Strengths
Explain-IT [41]	The clustering step through supervised learning may result in introducing bias due to the application of a specific model	provides improved insights into unsupervised data analysis results, enhancing the interpretability of clustering outcomes
TCAV [55]	Primarily for image classification; may not be applicable to other domains	Generates interpretable visualizations through concept activation vectors
EBMs [60]	Limited to decision tree-based models, can be computationally expensive for large datasets	Offers inherent interpretability through decision trees, facilitates feature importance analysis
RISE [44]	Primarily for image classification tasks; may not generalize well to other domains	Generates interpretable heatmaps by identifying relevant regions of an image
Kernel SHAPs [34,48]	Relies on kernel functions, which can be computationally expensive and require careful selection	Offers flexibility in handling complex relationships between features through kernels
ClAMPs [30]	Limited publicly available information about specific limitations and challenges	Aims to explain model predictions using Bayesian hierarchical clustering
OAK4XAI [33]	Relies on finding optimal anchors, which can be computationally expensive	Aims to identify concise and interpretable anchors using game theory concepts
TreeSHAPs [34]	Primarily focused on tree-based models, limiting its usage with other model architectures. Explanations for very complex decision trees can become intricate and challenging to interpret.	Offers greater clarity in understanding how tree-based models arrive at predictions.
CasualSHAPs [40]	Requires access to additional data and assumptions about causal relationships, which might not always be readily available or reliable. Can be computationally expensive, especially for large datasets.	Provides potential insights into which features drive outcomes, going beyond mere correlation.
INNvestigate [45,46]	While aiming for interpretability, the internal workings of the framework itself might not be readily understandable for non-experts.	Explanations for intricate deep learning models might still require some expertise to grasp fully. Makes the decision-making process of "black-box" deep learning models more transparent.
DeepLIFT [67,68]	Hyperparameters can significantly impact the explanation results, requiring careful tuning. May not be suitable for all types of deep learning models or tasks.	Pinpoints the importance of individual neurons within a neural network, offering insights into its internal workings.
ATTN [54]	Primarily focused on explaining attention mechanisms, limiting its use for broader model interpretability. Understanding attention weights and their impact on the model's output can be challenging for non-experts.	Helps decipher how attention mechanisms are used in deep learning models, especially for NLP and computer vision tasks.
DALEX [36]	Compared to some more established frameworks, DALEX might have less readily available documentation and resources. DALEX might still be under development, and its capabilities and limitations might evolve.	Facilitates "what-if" scenarios for deep learning models, enhancing understanding of how input changes might impact predictions.
ELI5 [71]	Limited to sci-kit-learn and XGBoost models. Might not generalize to all ML models	Provides clear, human-readable explanations for decision trees, linear models, and gradient-boosting models

Table 7. *Cont.*

XAI Framework	Key Limitations/Challenges	Primary Strengths
GraphLIMEs [39]	Computationally intensive, not ideal for real-time applications. Explains tabular data predictions by modeling feature relationships as graphs.	Especially useful when feature interactions are important
DLIME [32]	Relatively new framework with less mature documentation and community support	Aims to combine global and local explanations using game theory and feature attributions for diverse model types
InterpretML [47,102]	Requires additional installation and integration to work. Might not be as user-friendly for beginners	Offers a diverse set of interpretability techniques as part of a larger ML toolkit
Alteryx Explainable AI [38]	Commercial offering with potential cost implications	Provides various visual and interactive tools for model interpretability within the Alteryx platform
TNTRules [29]	The black-box nature may limit the collaborative tuning process, potentially introducing biases in decision-making	aiming to generate high-quality explanations through multiobjective optimization

5.1. Applicable Recommendations

The increasing pervasiveness of ML models in our lives necessitates a deeper understanding of their decision-making processes. This is where XAI frameworks come in, with an aim to shed light on the "black box" nature of these algorithms. However, the current XAI frameworks face various limitations that hinder their effectiveness. This study proposes a series of recommendations to improve these frameworks and unlock their full potential.

Computational Efficiency is Key: A significant hurdle for many XAI frameworks is their computational burden, especially when dealing with large datasets or complex models. Research into more efficient algorithms and hardware acceleration techniques is crucial. Frameworks like LIMEs, SHAPs, LOREs, and EBMs could benefit greatly from such advancements.

Interpretability Over-Complexity: While some frameworks excel at technical explanations, the user-friendliness often suffers. Efforts to simplify the explanations or provide clear visualizations would significantly enhance the usability of frameworks like iNNvestigate, LRP, and Kernel SHAP, making them more accessible to non-experts.

Expanding the XAI Toolbox: Many frameworks are limited to specific model types. Increased flexibility to handle a wider range of architectures would broaden their practical applications. Frameworks like Anchors, CEM, TCAV, EBMs, RISE, and TreeSHAPs could benefit from such expansion.

Scalability for the Real World: The effectiveness of some frameworks diminishes with large datasets. Finding ways to scale frameworks like LOREs and GraphLIMEs efficiently is essential for real-world applications.

User-Centric Design: Frameworks like Captum and InterpretML require additional setup and knowledge for implementation. Developing more user-friendly interfaces, along with comprehensive documentation and tutorials, would significantly lower the barrier to entry for beginners.

Beyond the Basics: Several frameworks offer specific functionalities. For example, AIX360 excels at fairness analysis but lacks individual prediction explanations. Integrating these functionalities would provide a more holistic understanding of model behavior. Similarly, frameworks like Skater could benefit from expanding the model support and user education on counterfactual reasoning.

Collaboration is Key: For frameworks like ClAMPs and DALEX, with limited publicly available information, fostering community involvement and research could lead to significant advancements.

Balancing Power and Ease of Use: Frameworks with powerful functionalities, like OAK4XAI and DeepLIFT, often require careful hyperparameter tuning. Developing user interfaces to guide the parameter selection and exploring predefined options would improve their accessibility without sacrificing accuracy.

Generalizability Beyond the Niche: Frameworks like ATTN and TCAV focus on the specific aspects of models, limiting their broader applicability. Investigating ways to expand their functionalities or develop complementary tools for broader interpretability would be beneficial.

Open Source for Open Minds: While Alteryx Explainable AI provides valuable functionalities, a freemium model or open-source alternatives could encourage wider adoption and foster community development.

In conclusion, the quest for transparency in ML necessitates continuous improvement in XAI frameworks. By focusing on the computational efficiency, user-friendly explanations, model and application diversity, user-centric design, and fostering community involvement, we can unlock the full potential of XAI and build trust in the intelligent systems that shape our world.

5.2. New Proposed XAI Framework

In this section, we propose a novel framework, termed eXplainable AI Evaluator (XAIE), designed to comprehensively evaluate and compare the existing XAI techniques. Currently, limitations exist in objectively assessing the strengths and weaknesses of various XAI methods. XAIE addresses this gap by establishing a standardized approach for cross-framework comparisons among diverse tasks and ML models.

Key Features:

Unified Benchmarking Suite: We propose the development of a comprehensive suite of benchmarks encompassing various XAI evaluation aspects. These benchmarks will assess the factors critical to XAI effectiveness, including the following:

- Fidelity: Measures how accurately the explanation reflects the true decision-making process of the underlying AI model.
- Fairness: Evaluates whether the explanation highlights or perpetuates any biases present within the model.
- Efficiency: Analyzes the computational cost associated with generating explanations using the XAI method.
- User-Centricity: Assesses the understandability and interpretability of the explanations for users with varying levels of technical expertise.

Standardized Explanation Representation: To facilitate meaningful comparisons, XAIE will define a common format for the explanations generated by different XAI techniques. This standardized format will allow for consistent evaluation and analysis across diverse frameworks.

Modular Design: To accommodate the evolving nature of the XAI field, the framework will be designed with a modular architecture. This modularity enables the seamless integration of new XAI methods and the incorporation of novel evaluation metrics as the research progresses.

Workflow: The XAIE framework will follow a user-driven workflow for XAI evaluation and comparison:

User Input:

- Users will specify the XAI methods they wish to compare.
- Details regarding the target AI model and its designated task (e.g., image classification or loan approval prediction) will also be provided by the user.

Evaluation Process:

- The XAIE framework will execute each selected XAI method on the target model, prompting the generation of explanations.
- The generated explanations will then be evaluated against the chosen benchmarks from the comprehensive suite.

Output:

- Upon completion, the framework will deliver a comprehensive report comparing the XAI methods based on the evaluation results. This report will include the following:
- A detailed breakdown of the strengths and weaknesses of each XAI method for the specific model and task.
- Visual representations of explanation fidelity and user-friendliness.
- Recommendations for the most suitable XAI method based on user priorities (e.g., prioritizing explainability over efficiency).

Figure 4 illustrates the flow chart that explains the processes involved in the new proposed XAIE model.

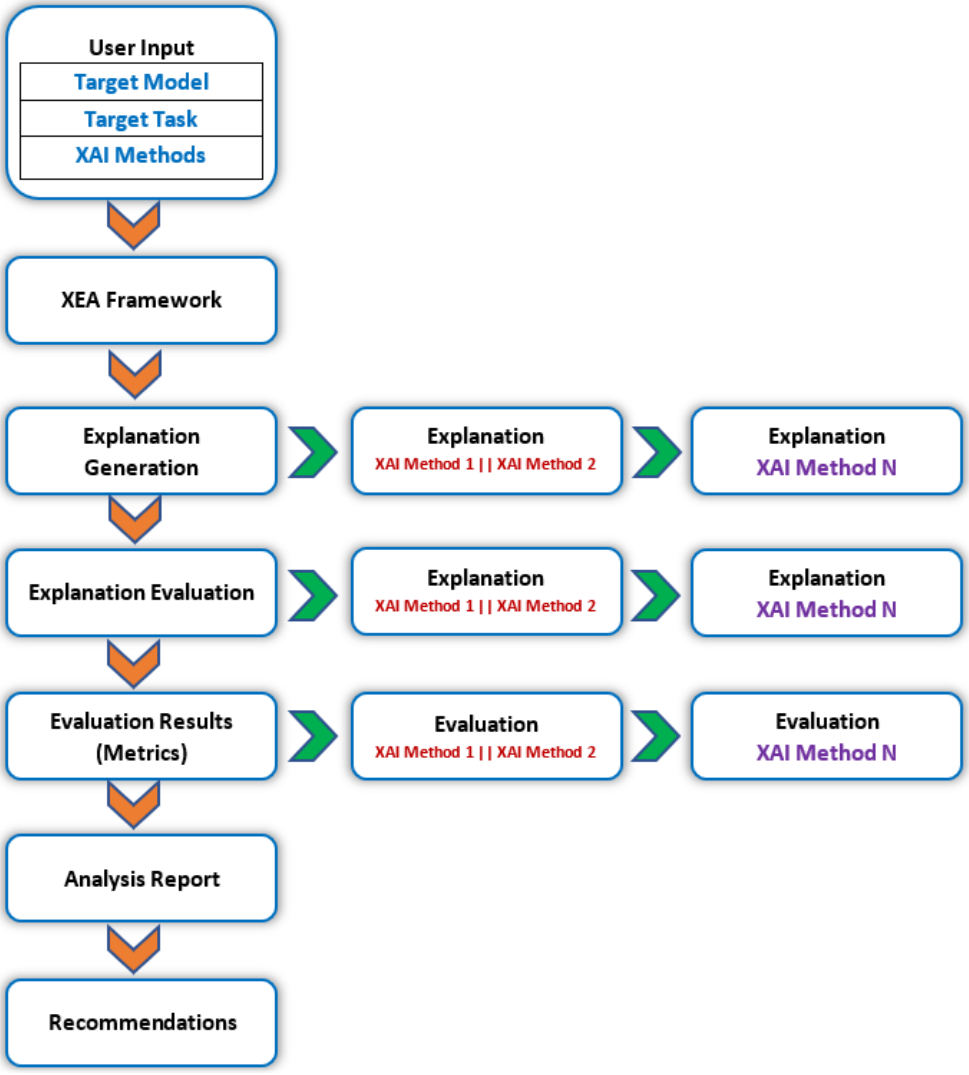

Figure 4. The workflow model of XAIE framework.

XAIE offers several advantages. Firstly, it promotes standardization in XAI evaluation. This consistency enables a more unified understanding of how effective different XAI techniques are. Secondly, XAIE empowers users to make informed decisions. By providing a platform for comparing various XAI methods, the users can select the most suitable technique for their specific needs and applications. Finally, XAIE acts as a valuable tool for

researchers. The framework facilitates the comparison and improvement of the existing XAI methods, ultimately accelerating the advancements in the field. However, developing XAIE also presents some challenges. Reaching a consensus on a common explanation format and a comprehensive set of evaluation metrics can be difficult due to the inherent diversity among the XAI methods. Additionally, running multiple XAI techniques on complex models can be computationally expensive. This necessitates exploring optimization techniques to ensure that the framework executes efficiently. Looking towards the future, the development of XAIE presents exciting opportunities. We aim to establish collaborations with XAI researchers to refine the proposed benchmarking suite and achieve a broader consensus on the evaluation metrics. Furthermore, we plan to integrate the framework with the existing XAI libraries and tools, promoting user-friendliness and widespread adoption. Finally, we will explore techniques for optimizing the framework's computational efficiency to handle complex AI models. By addressing these challenges and pursuing further advancements, XAIE has the potential to become a cornerstone for rigorous XAI evaluation, ultimately fostering the development and deployment of trustworthy and transparent AI systems.

6. Future Works

The field of XAI is constantly evolving, with new research and development efforts emerging at a rapid pace. Here are some key areas where the future work holds immense promise:

Human-Centered Explainability: Moving beyond technical explanations, future XAI frameworks should prioritize explanations tailored to human comprehension. This may involve incorporating cognitive science principles and user feedback to design explanations that are not only accurate but also resonate with the intended audience.

Explainable AI for Emerging Technologies: As AI ventures into new frontiers, like explainable reinforcement learning and interpretable deep fakes, the XAI frameworks need to adapt and evolve. Developing specialized tools for these domains will be crucial for ensuring the transparency and trust in these cutting-edge technologies.

Explainable AI for Real-Time Applications: The current limitations of many XAI frameworks make them unsuitable for real-time applications. Research into more efficient algorithms and hardware-specific implementations will be essential to bridge this gap and enable the real-time explainability for time-sensitive scenarios. The development of XAI with ProcessGPT [103] can improve the human comprehension of and support for the business processes.

Integration with Model Development Pipelines: Currently, XAI is often considered an afterthought in the ML workflow. The future advancements should integrate explainability considerations right from the model development stage. This would enable the proactive design of interpretable models and streamline the overall process.

Standardization and Benchmarking: The lack of standardized evaluation metrics for XAI frameworks makes it difficult to compare their effectiveness. Establishing standardized benchmarks and fostering collaboration between researchers will be essential for objectively evaluating and improving the XAI techniques.

By addressing these areas, the future of XAI holds the potential to transform our relationship with intelligent systems. We can move from a world of opaque algorithms to one where users can understand and trust the decisions made by AI, fostering the responsible development and deployment of these powerful technologies.

7. Conclusions

The quest for transparency in AI necessitates a comprehensive understanding of Explainable AI (XAI) frameworks. This paper has provided a detailed analysis of various XAI solutions, highlighting their strengths and limitations. While XAI offers several strengths, including the increased trust and acceptance of AI systems by stakeholders, the ability to identify and mitigate the potential biases in AI models, and improved debugging and model improvement capabilities, it also faces limitations and weaknesses. These

include its computational expensiveness and resource intensiveness, its inherent difficulty in fully explaining all AI models (particularly complex deep learning models), and the potential for its explainability techniques to be challenging for non-experts to understand. Moreover, we established a taxonomy for the XAI frameworks based on their key attributes and explored a diverse range of techniques, empowering researchers and practitioners to navigate the XAI landscape. Furthermore, the paper proposed a framework called XAIE for evaluating XAI solutions, enabling informed decision-making when selecting appropriate tools for specific application contexts. This fosters the responsible development and deployment of AI models by promoting user trust and understanding. Looking ahead, several key challenges remain. The computational efficiency needs to be improved, particularly for complex models and large datasets. User-friendliness should be prioritized to make XAI frameworks more accessible to a broader audience. Additionally, the XAI techniques need to adapt and evolve to handle the emerging AI frontiers, like explainable reinforcement learning and interpretable deep fakes. By addressing these challenges and fostering collaboration among researchers and developers, the future of XAI holds immense promise. We can move towards a future where AI models are not just powerful but also transparent and trustworthy. This will enable responsible AI development and unlock the full potential of AI for the benefit of society.

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work, the author(s) used Gemini and Grammarly to fix the grammatical errors and typos. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Author Contributions: Conceptualization, N.A.S. and R.R.C.; data collection, N.A.S., R.R.C., Z.B., and A.H.; validation, N.A.S.; formal analysis, N.A.S.; investigation, N.A.S.; resources, N.A.S.; writing—original draft preparation, N.A.S.; writing—review and editing, N.A.S., R.R.C., Z.B., A.B.M.S.A., A.H., and A.B.; visualization, N.A.S., R.R.C., and A.H.; supervision, A.B.M.S.A. and A.B.; project administration, N.A.S., A.B.M.S.A., and A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.* **2023**, *55*, 194:1–194:33. [CrossRef]
2. Palacio, S.; Lucieri, A.; Munir, M.; Ahmed, S.; Hees, J.; Dengel, A. Xai handbook: Towards a unified framework for explainable AI. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3766–3775.
3. Holzinger, A.; Saranti, A.; Molnar, C.; Biecek, P.; Samek, W. Explainable AI Methods—A Brief Overview. In *xxAI—Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*; Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.R., Samek, W., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 13–38. [CrossRef]
4. Le, P.Q.; Nauta, M.; Nguyen, V.B.; Pathak, S.; Schlötterer, J.; Seifert, C. Benchmarking eXplainable AI—A Survey on Available Toolkits and Open Challenges. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macau, China, 19–25 August 2023; pp. 6665–6673. [CrossRef]
5. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
6. Langer, M.; Oster, D.; Speith, T.; Hermanns, H.; Kästner, L.; Schmidt, E.; Sesing, A.; Baum, K. What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell.* **2021**, *296*, 103473. [CrossRef]
7. Liao, V.; Varshney, K. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv* **2021**, arXiv:2110.10790.
8. Mohseni, S.; Zarei, N.; Ragan, E.D. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv* **2020**, arXiv:1811.11839.

9. Hu, Z.F.; Kuflik, T.; Mocanu, I.G.; Najafian, S.; Shulner Tal, A. Recent Studies of XAI-Review. In Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, New York, NY, USA, 21–25 June 2021; UMAP '21; pp. 421–431. [CrossRef]
10. Das, A.; Rad, P. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv* **2020**, arXiv:2006.11371.
11. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Ser, J.D.; 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]
12. Hanif, A.; Beheshti, A.; Benatallah, B.; Zhang, X.; Habiba; Foo, E.; Shabani, N.; Shahabikargar, M. A Comprehensive Survey of Explainable Artificial Intelligence (XAI) Methods: Exploring Transparency and Interpretability. In *Web Information Systems Engineering—WISE 2023*; Zhang, F., Wang, H., Barhamgi, M., Chen, L., Zhou, R., Eds.; Springer Nature: Berlin/Heidelberg, Germany, 2023; pp. 915–925. [CrossRef]
13. Hanif, A.; Zhang, X.; Wood, S. A Survey on Explainable Artificial Intelligence Techniques and Challenges. In Proceedings of the 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), Gold Coast, Australia, 25–29 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 81–89. [CrossRef]
14. Salimzadeh, S.; He, G.; Gadiraju, U. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, New York, NY, USA, 26–29 June 2023; UMAP '23; pp. 215–227. [CrossRef]
15. Grosan, C.; Abraham, A. *Intelligent Systems*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 17.
16. Rong, Y.; Leemann, T.; Nguyen, T.T.; Fiedler, L.; Qian, P.; Unhelkar, V.; Seidel, T.; Kasneci, G.; Kasneci, E. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2104–2122. [CrossRef] [PubMed]
17. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable artificial intelligence: A survey. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; pp. 0210–0215. [CrossRef]
18. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef]
19. Duval, A. *Explainable Artificial Intelligence (XAI)*; MA4K9 Scholarly Report; Mathematics Institute, The University of Warwick: Coventry, UK, 2019; Volume 4.
20. Kim, M.Y.; Atakishiyev, S.; Babiker, H.K.B.; Farruque, N.; Goebel, R.; Zaïane, O.R.; Motallebi, M.H.; Rabelo, J.; Syed, T.; Yao, H.; et al. A Multi-Component Framework for the Analysis and Design of Explainable Artificial Intelligence. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 900–921. [CrossRef]
21. John-Mathews, J.M. Some critical and ethical perspectives on the empirical turn of AI interpretability. *Technol. Forecast. Soc. Chang.* **2022**, *174*, 121209. [CrossRef]
22. Dwivedi, Y.K.; Hughes, L.; Ismagilova, E.; Aarts, G.; Coombs, C.; Crick, T.; Duan, Y.; Dwivedi, R.; Edwards, J.; Eirug, A.; et al. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *Int. J. Inf. Manag.* **2021**, *57*, 101994. [CrossRef]
23. Ribes, D.; Henchoz, N.; Portier, H.; Defayes, L.; Phan, T.T.; Gatica-Perez, D.; Sonderegger, A. Trust Indicators and Explainable AI: A Study on User Perceptions. In *Human-Computer Interaction—INTERACT 2021*; Ardito, C., Lanzilotti, R., Malizia, A., Petrie, H., Piccinno, A., Desolda, G., Inkpen, K., Eds.; Springer: Cham, Switzerland, 2021; pp. 662–671. [CrossRef]
24. Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.A.; Mooney, C. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl. Sci.* **2021**, *11*, 5088. [CrossRef]
25. Ehsan, U.; Saha, K.; Choudhury, M.; Riedl, M. Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI. *Proc. Acm -Hum.-Comput. Interact.* **2023**. [CrossRef]
26. Lopes, P.; Silva, E.; Braga, C.; Oliveira, T.; Rosado, L. XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Appl. Sci.* **2022**, *12*, 9423. [CrossRef]
27. Naiseh, M.; Simkute, A.; Zieni, B.; Jiang, N.; Ali, R. C-XAI: A conceptual framework for designing XAI tools that support trust calibration. *J. Responsible Technol.* **2024**, *17*, 100076. [CrossRef]
28. Capuano, N.; Fenza, G.; Loia, V.; Stanzione, C. Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access* **2022**, *10*, 93575–93600. [CrossRef]
29. Chakraborty, T.; Seifert, C.; Wirth, C. Explainable Bayesian Optimization. *arXiv* **2024**, arXiv:2401.13334.
30. Bobek, S.; Kuk, M.; Szelążek, M.; Nalepa, G.J. Enhancing Cluster Analysis with Explainable AI and Multidimensional Cluster Prototypes. *IEEE Access* **2022**, *10*, 101556–101574. [CrossRef]
31. Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F.; Giannotti, F. Local Rule-Based Explanations of Black Box Decision Systems. *arXiv* **2018**, arXiv:1805.10820.
32. Zafar, M.R.; Khan, N. Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 525–541. [CrossRef]
33. Ngo, Q.H.; Kechadi, T.; Le-Khac, N.A. OAK4XAI: Model Towards Out-of-Box eXplainable Artificial Intelligence for Digital Agriculture. In *Artificial Intelligence XXXIX*; Bramer, M., Stahl, F., Eds.; Springer: Cham, Switzerland, 2022; pp. 238–251. [CrossRef]

34. 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]
35. Radhakrishnan, A.; Beaglehole, D.; Pandit, P.; Belkin, M. Mechanism of feature learning in deep fully connected networks and kernel machines that recursively learn features. *arXiv* **2023**, arXiv:2212.13881. [CrossRef].
36. Biecek, P. DALEX: Explainers for Complex Predictive Models in R. *J. Mach. Learn. Res.* **2018**, *19*, 1–5.
37. Dhurandhar, A.; Chen, P.Y.; Luss, R.; Tu, C.C.; Ting, P.; Shanmugam, K.; Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2–8 December 2018; NIPS'18; pp. 590–601.
38. Alteryx. The Essential Guide to Explainable AI (XAI). Available online: https://www.alteryx.com/resources/whitepaper/essential-guide-to-explainable-ai (accessed on 1 April 2024).
39. Huang, Q.; Yamada, M.; Tian, Y.; Singh, D.; Yin, D.; Chang, Y. GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks. *arXiv* **2020**, arXiv:2001.06216. [CrossRef].
40. Heskes, T.; Bucur, I.G.; Sijben, E.; Claassen, T. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 6–12 December 2020; NIPS '20; pp. 4778–4789.
41. Morichetta, A.; Casas, P.; Mellia, M. EXPLAIN-IT: Towards Explainable AI for Unsupervised Network Traffic Analysis. In Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks, New York, NY, USA, 9 December 2019; Big-DAMA '19; pp. 22–28. [CrossRef]
42. Ribeiro, M.T.; Singh, S.; Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [CrossRef]
43. Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; et al. Captum: A unified and generic model interpretability library for PyTorch. *arXiv* **2020**, arXiv:2009.07896. [CrossRef].
44. Petsiuk, V.; Das, A.; Saenko, K. RISE: Randomized Input Sampling for Explanation of Black-box Models. *arXiv* **2018**, arXiv:1806.07421.
45. Alber, M.; Lapuschkin, S.; Seegerer, P.; Hägele, M.; Schütt, K.T.; Montavon, G.; Samek, W.; Müller, K.R.; Dähne, S.; Kindermans, P.J. iNNvestigate Neural Networks! *J. Mach. Learn. Res.* **2019**, *20*, 1–8.
46. Garcea, F.; Famouri, S.; Valentino, D.; Morra, L.; Lamberti, F. iNNvestigate-GUI - Explaining Neural Networks Through an Interactive Visualization Tool. In *Artificial Neural Networks in Pattern Recognition*; Schilling, F.P., Stadelmann, T., Eds.; Springer: Cham, Switzerland, 2020; pp. 291–303. [CrossRef]
47. Nori, H.; Jenkins, S.; Koch, P.; Caruana, R. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv* **2019**, arXiv:1909.09223.
48. Lundberg, S.M.; Allen, P.G.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
49. 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.* **2017**, *128*, 336–359. [CrossRef]
50. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. *arXiv* **2017**, arXiv:1703.01365. [CrossRef].
51. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. SmoothGrad: Removing noise by adding noise. *arXiv* **2017**, arXiv:1706.03825.
52. Zhang, Z.; Cui, P.; Zhu, W. Deep Learning on Graphs: A Survey. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 249–270. [CrossRef]
53. Miković, R.; Arsić, B.; Gligorijević, Đ. Importance of social capital for knowledge acquisition—DeepLIFT learning from international development projects. *Inf. Process. Manag.* **2024**, *61*, 103694. [CrossRef]
54. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
55. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; sayres, R. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In Proceedings of the International Conference on Machine Learning. PMLR, Stockholm, Sweden, 10–15 July 2018; Volume 6, pp. 2668–2677, ISSN 2640-3498.
56. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Volume 13–17-August-2016, KDD '16; pp. 1135–1144. [CrossRef]
57. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef]
58. Wexler, J.; Pushkarna, M.; Bolukbasi, T.; Wattenberg, M.; Viegas, F.; Wilson, J. The what-if tool: Interactive probing of machine learning models. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 56–65. [CrossRef]
59. Arya, V.; Bellamy, R.K.E.; Chen, P.Y.; Dhurandhar, A.; Hind, M.; Hoffman, S.C.; Houde, S.; Liao, Q.V.; Luss, R.; Mojsilović, A.; et al. AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models. *J. Mach. Learn. Res.* **2020**, *21*, 1–6.
60. Liu, G.; Sun, B. Concrete compressive strength prediction using an explainable boosting machine model. *Case Stud. Constr. Mater.* **2023**, *18*, e01845. [CrossRef]

61. Kawakura, S.; Hirafuji, M.; Ninomiya, S.; Shibasaki, R. Adaptations of Explainable Artificial Intelligence (XAI) to Agricultural Data Models with ELI5, PDPbox, and Skater using Diverse Agricultural Worker Data. *Eur. J. Artif. Intell. Mach. Learn.* **2022**, *1*, 27–34. [CrossRef]
62. Asif, N.A.; Sarker, Y.; Chakrabortty, R.K.; Ryan, M.J.; Ahamed, M.H.; Saha, D.K.; Badal, F.R.; Das, S.K.; Ali, M.F.; Moyeen, S.I.; et al. Graph Neural Network: A Comprehensive Review on Non-Euclidean Space. *IEEE Access* **2021**, *9*, 60588–60606. [CrossRef]
63. Binder, A.; Montavon, G.; Bach, S.; Müller, K.R.; Samek, W. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. *arXiv* **2016**, arXiv:1604.00825. [CrossRef].
64. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016.
65. Arya, V.; Bellamy, R.K.E.; Chen, P.Y.; Dhurandhar, A.; Hind, M.; Hoffman, S.C.; Houde, S.; Liao, Q.V.; Luss, R.; Mojsilović, A.; et al. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv* **2019**, arXiv:1909.03012.
66. Klaise, J.; Looveren, A.V.; Vacanti, G.; Coca, A. Alibi Explain: Algorithms for Explaining Machine Learning Models. *J. Mach. Learn. Res.* **2021**, *22*, 1–7.
67. Shrikumar, A.; Greenside, P.; Shcherbina, A.; Kundaje, A. Not just a black box: Learning important features through propagating activation differences. *arXiv* **2016**, arXiv:1605.01713.
68. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. In Proceedings of the International Conference on Machine Learning. PMLR, Sydney, NSW, Australia, 6–11 August 2017; pp. 3145–3153.
69. Li, J.; Zhang, C.; Zhou, J.T.; Fu, H.; Xia, S.; Hu, Q. Deep-LIFT: Deep Label-Specific Feature Learning for Image Annotation. *IEEE Trans. Cybern.* **2022**, *52*, 7732–7741. [CrossRef] [PubMed]
70. Soydaner, D. Attention mechanism in neural networks: Where it comes and where it goes. *Neural Comput. Appl.* **2022**, *34*, 13371–13385. [CrossRef]
71. Fan, A.; Jernite, Y.; Perez, E.; Grangier, D.; Weston, J.; Auli, M. ELI5: Long Form Question Answering. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Korhonen, A., Traum, D., Màrquez, L., Eds.; pp. 3558–3567. [CrossRef]
72. Cao, L. *AI in Finance: A Review*. 2020. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3647625 (accessed on 1 April 2024). [CrossRef]
73. Goodell, J.W.; Kumar, S.; Lim, W.M.; Pattnaik, D. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *J. Behav. Exp. Financ.* **2021**, *32*, 100577. [CrossRef]
74. Zheng, X.L.; Zhu, M.Y.; Li, Q.B.; Chen, C.C.; Tan, Y.C. FinBrain: When finance meets AI 2.0. *Front. Inf. Technol. Electron. Eng.* **2019**, *20*, 914–924. [CrossRef]
75. Cao, N. Explainable Artificial Intelligence for Customer Churning Prediction in Banking. In Proceedings of the 2nd International Conference on Human-Centered Artificial Intelligence (Computing4Human 2021), Danang, Vietnam, 28–29 October 2021; pp. 159–167.
76. Weber, P.; Carl, K.V.; Hinz, O. Applications of Explainable Artificial Intelligence in Finance—A systematic review of Finance, Information Systems, and Computer Science literature. *Manag. Rev. Q.* **2023**, *74*, 867–907. [CrossRef]
77. van der Velden, B.H.M.; Kuijf, H.J.; Gilhuijs, K.G.A.; Viergever, M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* **2022**, *79*, 102470. [CrossRef]
78. Simonyan, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Proceedings of the Workshop at International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
79. Joyce, D.W.; Kormilitzin, A.; Smith, K.A.; Cipriani, A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *NPJ Digit. Med.* **2023**, *6*, 6. [CrossRef] [PubMed]
80. Zebin, T.; Rezvy, S.; Luo, Y. An Explainable AI-Based Intrusion Detection System for DNS Over HTTPS (DoH) Attacks. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 2339–2349. [CrossRef]
81. Andresini, G.; Appice, A.; Caforio, F.P.; Malerba, D.; Vessio, G. ROULETTE: A neural attention multi-output model for explainable Network Intrusion Detection. *Expert Syst. Appl.* **2022**, *201*, 117144. [CrossRef]
82. A. Reyes, A.; D. Vaca, F.; Castro Aguayo, G.A.; Niyaz, Q.; Devabhaktuni, V. A Machine Learning Based Two-Stage Wi-Fi Network Intrusion Detection System. *Electronics* **2020**, *9*, 1689. [CrossRef]
83. Górski, Ł.; Ramakrishna, S. Explainable artificial intelligence, lawyer's perspective. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, New York, NY, USA, 21–25 June 2021; ICAIL '21; pp. 60–68. [CrossRef]
84. Bringas Colmenarejo, A.; Beretta, A.; Ruggieri, S.; Turini, F.; Law, S. The Explanation Dialogues: Understanding How Legal Experts Reason About XAI Methods. In Proceedings of the European Workshop on Algorithmic Fairness: Proceedings of the 2nd European Workshop on Algorithmic Fairness, Winterthur, Switzerland, 7–9 June 2023.
85. Ramon, Y.; Vermeire, T.; Toubia, O.; Martens, D.; Evgeniou, T. Understanding Consumer Preferences for Explanations Generated by XAI Algorithms. *arXiv* **2021**, arXiv:2107.02624. [CrossRef].
86. Feng, X.F.; Zhang, S.; Srinivasan, K. *Marketing Through the Machine's Eyes: Image Analytics and Interpretability*; Emerald Publishing Limited: England, UK, 2022.

87. Sutthithatip, S.; Perinpanayagam, S.; Aslam, S.; Wileman, A. Explainable AI in Aerospace for Enhanced System Performance. In Proceedings of the 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 3–7 October 2021; pp. 1–7, ISSN 2155-7209. [CrossRef]
88. Hernandez, C.S.; Ayo, S.; Panagiotakopoulos, D. An Explainable Artificial Intelligence (xAI) Framework for Improving Trust in Automated ATM Tools. In Proceedings of the 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 3–7 October 2021; pp. 1–10, ISSN 2155-7209. [CrossRef]
89. Vowels, M.J. Trying to Outrun Causality with Machine Learning: Limitations of Model Explainability Techniques for Identifying Predictive Variables. *arXiv* **2022**, arXiv:2202.09875. [CrossRef].
90. Samek, W.; Wiegand, T.; Müller, K.R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv* **2017**, arXiv:1708.08296. QID: Q38135445.
91. Letzgus, S.; Wagner, P.; Lederer, J.; Samek, W.; Muller, K.R.; Montavon, G. Toward Explainable Artificial Intelligence for Regression Models: A methodological perspective. *IEEE Signal Process. Mag.* **2022**, *39*, 40–58. [CrossRef]
92. Ribeiro, J.; Silva, R.; Cardoso, L.; Alves, R. Does Dataset Complexity Matters for Model Explainers? In Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 15–18 December 2021; pp. 5257–5265. [CrossRef]
93. Helgstrand, C.J.; Hultin, N. *Comparing Human Reasoning and Explainable AI*; 2022. Available online: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1667042&dswid=-8097 (accessed on 1 April 2024).
94. Dieber, J.; Kirrane, S. Why model why? Assessing the strengths and limitations of LIME. *arXiv* **2020**, arXiv:2012.00093. [CrossRef].
95. El-Mihoub, T.A.; Nolle, L.; Stahl, F. Explainable Boosting Machines for Network Intrusion Detection with Features Reduction. In Proceedings of the Artificial Intelligence XXXIX, Cambridge, UK, 13–15 December 2021; Bramer, M., Stahl, F., Eds.; Springer: Cham, Switzerland, 2022; pp. 280–294. [CrossRef]
96. Jayasundara, S.; Indika, A.; Herath, D. Interpretable Student Performance Prediction Using Explainable Boosting Machine for Multi-Class Classification. In Proceedings of the 2022 2nd International Conference on Advanced Research in Computing (ICARC), Belihuloya, Sri Lanka, 23–24 February 2022; pp. 391–396. [CrossRef]
97. Roshan, K.; Zafar, A. Utilizing XAI Technique to Improve Autoencoder based Model for Computer Network Anomaly Detection with Shapley Additive Explanation(SHAP). *Int. J. Comput. Netw. Commun.* **2021**, *13*, 109–128. [CrossRef]
98. Roshan, K.; Zafar, A. Using Kernel SHAP XAI Method to Optimize the Network Anomaly Detection Model. In Proceedings of the 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 23–25 March 2022; pp. 74–80. [CrossRef]
99. Sahakyan, M.; Aung, Z.; Rahwan, T. Explainable Artificial Intelligence for Tabular Data: A Survey. *IEEE Access* **2021**, *9*, 135392–135422. [CrossRef]
100. Hickling, T.; Zenati, A.; Aouf, N.; Spencer, P. Explainability in Deep Reinforcement Learning: A Review into Current Methods and Applications. *ACM Comput. Surv.* **2023**, *56*, 125:1–125:35. [CrossRef]
101. Šimić, I.; Sabol, V.; Veas, E. XAI Methods for Neural Time Series Classification: A Brief Review. *arXiv* **2021**, arXiv:2108.08009. [CrossRef].
102. Agarwal, N.; Das, S. Interpretable Machine Learning Tools: A Survey. In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia, 1–4 December 2020; pp. 1528–1534. [CrossRef]
103. Beheshti, A.; Yang, J.; Sheng, Q.Z.; Benatallah, B.; Casati, F.; Dustdar, S.; Nezhad, H.R.M.; Zhang, X.; Xue, S. ProcessGPT: Transforming Business Process Management with Generative Artificial Intelligence. In Proceedings of the 2023 IEEE International Conference on Web Services (ICWS), Chicago, IL, USA, 2–8 July 2023; pp. 731–739. [CrossRef]