Survey on Explainable AI: Techniques, Challenges and Open Issues

Adel Abusitta[a,*], Miles Q. Li[b], Benjamin C. M. Fung[b]

[a]*Department of Computer and Software Engineering, Polytechnique Montréal, Montréal, QC, Canada*
[b]*School of Information Studies, McGill University, Montréal, QC, Canada*

Abstract

Artificial Intelligence (AI) has become an important component of many software applications. It has reached a point where it can provide complex and critical decisions in our life. However, the success of most AI-powered applications is based on black-box approaches (e.g., deep neural networks), which can create learned models that are able to predict and make decisions. While these advanced models could achieve high accuracy, they are generally unable to explain their decisions (e.g., predictions) to users. As a result, there is a pressing need for explainable machine learning systems in order to be trustworthy by governments, organizations, industries, and users. This paper classifies and compares the main findings in the domain of explainable machine learning and deep learning. We also discuss the application of Explainable AI (XAI) in sensitive domains such as cybersecurity. In addition, we characterize each reviewed article on the basis of the methods and techniques used to achieve XAI. This, in turn, allows us to discern the strengths and limitations of the existing XAI techniques. We finally discuss some substantial challenges and future research directions related to XAI.

Keywords: Explainable artificial intelligence, Machine Learning, Interpretability, Trusted artificial intelligence

*Corresponding author
Email addresses: adel.abu-sitta@polymtl.ca (Adel Abusitta), miles.qi.li@mail.mcgill.ca (Miles Q. Li), ben.fung@mcgill.ca (Benjamin C. M. Fung)

1. Introduction

The latest advances in the Artificial Intelligence (AI) and Machine Learning domains have led to their adoption in many complex and sensitive applications, including those of autonomous cars, 5G connected mobile devices, smart robots, healthcare, AI-powered cybersecurity, etc. (Bega et al., 2019; Miotto et al., 2018; Stilgoe, 2018). Although they achieve promising results in terms of accuracy and reduce the need for human interactions, most machine learning models are not able to provide a clear interpretation about the decisions they make Rabiul Islam et al. (2021) Linardatos et al. (2021). In other words, these approaches are "good" at extracting patterns and models behind data, but they often cannot explain data correlation in a way useful for most humans, and are therefore unable to present interpretable reasons of the prediction or classification to users. As a result, analysts and decision makers are concerned about the reliability of these complex black box systems DND (2021). In order for these systems to be trustworthy Kästner et al. (2021), it is important to integrate *causality* into their design Knight (2021); Janzing et al. (2020). The problem of trust in AI systems is complex and multidimensional since it is largely dependent on interactions between humans and machines (Trunk et al., 2020). Gaining trust in these complex systems is challenging and needs efficient and effective solutions to encourage public and private sectors to accept these systems.

A "perfect" Machine Learning model is one that can produce accurate predictions, and at the same time, describes in detail its predictions (i.e, provides an interpretation) Gunning et al. (2019). There is a common belief among AI researchers and professionals that there is a trade-off between accuracy and interpretability in a machine learning model (Arrieta et al., 2020; Rai, 2020). For example, linear models, such as Naive Bayes and linear/logistic regression have excellent interpretability, but the accuracy of these models is too low when they are used on complex and noisy data (e.g., images and malware data) Arrieta et al. (2020). On the other hand, deep neural networks (DNNs) are more accurate than traditional machine learning models, but the decisions of DNNs are mostly not interpretable to users (Camburu, 2020; Rabiul Islam et al., 2021). Figure 1 illustrates the trade-off between accuracy and interpretability Alonso

et al. (2018); Campagnolo & Sharkey (2021); Arrieta et al. (2020).

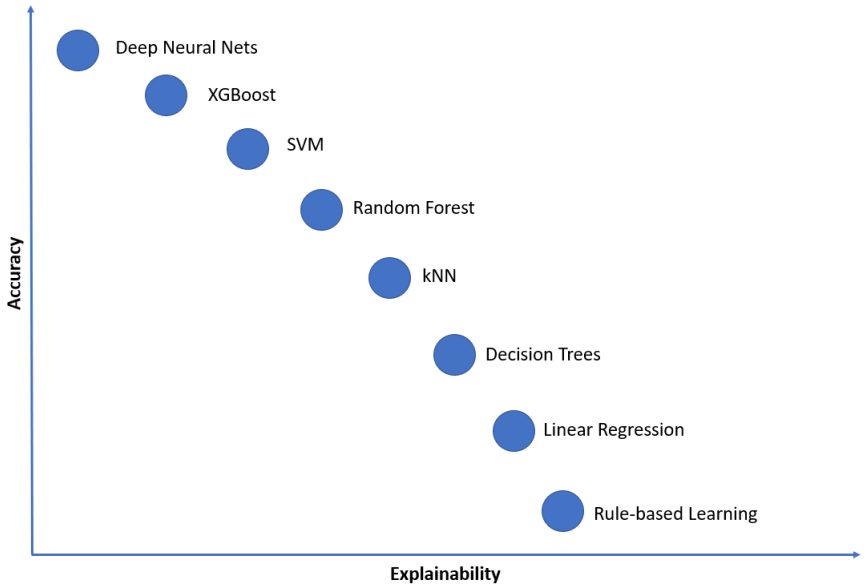

Figure 1: Trade-off between model accuracy and explainability Alonso et al. (2018); Campagnolo & Sharkey (2021); Arrieta et al. (2020)

Explainable artificial intelligence (XAI) can be seen as a set of procedures and techniques that can be applied and/or integrated to/into machine learning models in order to make them explainable Samek & Müller (2019); Arrieta et al. (2020). To this end, XAI-powered applications should not only provide accurate results, but also be able to justify their results in a way that is understandable by humans and especially end users. For example, in the context of XAI-powered cybersecurity, the machine learning technique should not only suggest that a given activity is an attack, but it should also detail what features have been captured to support such a decision. Having XAI allows humans (e.g., employees, users) to understand the decision and output (e.g., prediction) of their intelligent applications (Li et al., 2021b). This, in turn, encourages governments and organizations to adopt these systems in developing sensitive and critical AI-powered applications (e.g., cybersecurity and healthcare) Knight (2021); DND (2021); Arrieta et al. (2020).

Recently, many researchers in AI and other fields have started to find advanced

solutions to understand the behavior of machine learning systems in order to achieve explainable solutions. To this end, new techniques have been proposed. These techniques can be categorized into three types: model-agnostic, model-specific, and model-semi-agnostic techniques Arrieta et al. (2020); Linardatos et al. (2021). The first can be applied to any machine learning model Arrieta et al. (2020); Linardatos et al. (2021). The second type of technique can be only applied to a specific machine learning model such as deep neural networks Arrieta et al. (2020); Linardatos et al. (2021). Finally, the last type of technique can be applied to a **subset** (not all) of machine learning models.

1.1. The Scope

This survey compares and classifies the main and recent findings in XAI. The selection of papers was done based on their importance. The importance is evaluated according to the following factors: publication venues (e.g., impact factor, rank), number of citations, and impact. We focus on these papers that can produce accurate predictions, and at the same time, describe in detail its predictions (i.e., explanation). Additionally, our paper includes XAI techniques that are proposed for data-driven machine learning models (DMLM). Other XAI techniques that are not dedicated to DMLM are not included in this paper because their goal is to cover all AI techniques. The difference between XAI and interpretable machine learning (IML) is not addressed in this paper as well, as it is out of the scope. Other surveys introduced the difference between XAI and IML (e.g., Arrieta et al. (2020)). This survey also aims at summarizing existing solutions and identifying the main issues and challenges in XAI. Specifically, our analysis leads to the recognition of four major problems to address:

- The need to find an acceptable balance between explainability and accuracy in machine learning.
- The efficiency and scalability of XAI methods, as the size of data and number of features need to scale largely in some critical applications.
- The protection of XAI against robust and ongoing adversarial attacks.
- The integration of XAI into existing machine learning platforms to help developers apply XAI efficiently.

We also propose a few guidelines to address these challenges.

1.2. Contribution

The main contributions of this survey include:

- Proposing a new taxonomy for describing and comparing the recent and main findings in XAI.
- Providing a detailed comparative analysis of existing literature on XAI.
- Identifying and presenting open issues and challenges related to XAI, and suggesting guidelines on how to improve XAI solutions to address current and continuing challenges.

1.3. Organization

We organize this literature review as follows. In Section 3 and Section 5, we present the applications of XAI in sensitive domains and propose a new taxonomy for organizing reviewed XAI approaches, respectively. Section 6 characterises reviewed papers according to the proposed taxonomy. In Section 8, we discuss the challenges and current issues related to XAI. Finally, Section 9 concludes the literature review.

2. Related Work

Popular XAI techniques and examples on credit default prediction were presented by Rabiul Islam et al. (2021). They also discussed and analyzed the difference between local and global XAI methods, as explored in studies by Angelov et al. (2021), Belle & Papantonis (2021), and McDermid et al. (2021). Moreover, recommendations about XAI in the context of human-centered AI were provided.

Linardatos et al. (2021) presented a literature review and taxonomy of XAI. They focus on programming implementations of the existing XAI techniques. Furthermore, they presented four categories for XAI techniques: techniques for interpreting black-box models, techniques for building white-box models, techniques for fairness assurance in AI, and finally, techniques for analysing the prediction sensitivity of the ma-

chine learning models. Similarly, Adadi & Berrada (2018) presented a useful background in XAI. They focus on providing detailed information about what is XAI, why we need XAI, and how to use XAI. They also provide a taxonomy about XAI.

Das & Rad (2020) provided a mathematical description of XAI approaches. They also presented a new taxonomy (classification) of XAI techniques. The proposed taxonomy is essentially based on three dimensions: explanation scope, algorithm used, and level of explanation. They also discussed the most important principles adopted in XAI works. Furthermore, they presented a new framework for evaluating XAI algorithms. Similarly, Guidotti et al. (2018) provided a taxonomy of techniques used for explaining black box models.

Recently, Arrieta et al. (2020) presented a new survey of explainable artificial intelligence. They discussed the main concepts of XAI such as understandability, comprehensibility, and interpretability of XAI. They also proposed taxonomies of XAI. In addition, they presented some challenges in responsible AI. Guidotti et al. (2021) introduced principles of XAI. They discussed several techniques used for XAI in the literature and elaborated their usability and applicability in real-world AI applications.

Vilone & Longo (2021) introduced the limitation of XAI formats (formats of explanations). The provided some recommendations on how to create applicable and realistic XAI techniques. Similarly, Langer et al. (2021) provided some recommendations on how to build XAI that satisfy and meet stakeholders' desiderata Langer et al. (2021).

More recently, Saranya and Subhashini Saranya & Subhashini (2023) reviewed 91 recent articles on XAI to explore its applications and development across healthcare, manufacturing, transportation, and finance. The review, covering publications from January 2018 to October 2022, was conducted using databases such as Scopus, Web of Science, IEEE Xplore, and PubMed. XAI aims to enhance the transparency, trustworthiness, and accountability of AI systems by providing explanations for their decisions or predictions, addressing concerns over the opaque "black box" nature and cumulative complexity of traditional AI models. The findings serve as a roadmap for further research, emphasizing the need for continued efforts to improve AI model explainability in high-stakes applications. Similarly, Schwalbe et al. Schwalbe & Finzel (2023)

present a unified taxonomy of XAI methods, integrating terminologies, concepts, and approaches from over 50 prominent surveys. Their aim is to assist researchers and practitioners in understanding, comparing, and selecting appropriate XAI methods based on specific use-case requirements. While providing a comprehensive reference, the paper's limitation lies in its potential oversimplification of complex, nuanced distinctions between methods due to the broad unification of diverse terminologies and concepts.

Another survey by Saeed et al. Saeed & Omlin (2023), also published recently, presents a systematic meta-survey of XAI, focusing on the challenges and future research directions in two key areas: general challenges and those specific to the machine learning life cycle phases (design, development, and deployment). It consolidates scattered insights from various reviews, offering a structured overview that aims to advance XAI's transparency and adoption in critical domains. While the paper provides a valuable guide for future exploration, its limitation lies in the potential for overlooked nuances due to the broad categorization of challenges and directions, which may oversimplify complex, context-specific issues in XAI.

In the context of the Internet of Things (IoT) domain, K"ok et al. Kök et al. (2023) provided a systematic review of recent studies on Explainable AI (XAI) in IoT, addressing the need for transparency, interpretability, and responsibility in AI/ML models. The review classifies studies by methodology and application area, highlights challenges and open issues, and suggests future research directions. However, a limitation of the review is its broad classification, which may overlook nuanced distinctions and context-specific challenges within different IoT applications.

2.1. Summarization of the existing survey papers on XAI and their differences with the proposed survey

In this section, we present a table (Table 1) summarizing existing works on surveys of XAI. The table provides an overview of the Main Focus, Methodologies Covered, Key Findings, and Gaps Identified in each survey paper. Additionally, we analyze the scope of our research and highlight how our proposed survey distinguishes itself from existing literature.

In this paper, we propose a novel taxonomy of XAI techniques, expanding upon

Table 1: Summary of papers conducting surveys on XAI

Paper	Main Focus	Methodologies Covered	Key Findings	Gaps Identified
Rabiul et al. (2021)	Survey of XAI approaches, focusing on overcoming lack of explainability in AI "black box" systems.	Case study analysis (credit default prediction), local and global perspectives of explainability, insights on quantifying explainability.	Competitive advantages, merits, and demerits of XAI methods; future research directions for responsible and human-centered AI systems.	Open challenges in XAI, need for responsible and human-centric AI systems.
Angelov et al. (2021)	Analytical review of current state-of-the-art in XAI, particularly in machine learning and deep learning.	Historical introduction, taxonomy of XAI, challenges based on National Institute of Standards' principles, review and analysis of recent methods.	Challenges in achieving explainability in AI systems; trends in XAI methods; future research directions.	Need for methods aligned with principles of transparency and trustworthiness.
McDermid et al. (2021)	Overview of technical and ethical dimensions of Artificial XAI.	Overview of XAI methods, alignment with stakeholder purposes, integration into AI development life cycle.	Importance of integrating XAI into AI development life cycle; framework for accountability and ethical decision-making.	Further integration and application of XAI methods to enhance transparency and interpretability in AI systems.
Belle et al. (2021)	Principles and practices of Explainable Machine Learning (ML), focusing on understanding decisions and enhancing trust in AI systems.	Survey of data-driven methods (ML, pattern recognition), distillation of results and observations from literature.	Impact of ML in various domains; concerns about model drawbacks and biases; survey aimed at educating industry practitioners and data scientists.	Gap in awareness among practitioners about emerging academic approaches; need for better tools and practices in applying XAI methods.
Kok et al. (2023)	Review of XAI models and applications within the IoT domain.	Systematic review categorizing studies based on methodology and application areas of XAI in IoT.	Applications and challenges of XAI in IoT; future directions and open issues for research in integrating XAI into IoT applications.	Need for defining and integrating XAI into IoT applications; guidance for future research in XAI within IoT contexts.
Saeed et al. (2023)	Systematic meta-survey of challenges and future opportunities in XAI.	Meta-survey of challenges and research directions in XAI, organized around general and lifecycle phases (design, development, deployment).	Consolidated challenges and research directions in XAI; guide for future exploration and development in the field.	Organized approach needed to address challenges in XAI; specific research needs across different phases of AI lifecycle.
Schwalbe et al. (2023)	Comprehensive taxonomy of XAI methods based on existing surveys and taxonomies in literature.	Structured literature analysis, meta-study of over 50 surveys on XAI methods, merging terminologies and concepts into unified taxonomy.	Unified taxonomy of XAI methods; practical tool for selecting methods based on use-case contexts; foundations for future research.	Need for a unified taxonomy to aid in selecting appropriate XAI methods; future research directions in context-sensitive XAI applications.
Saranya et al. (2023)	Systematic review of recent developments and trends in XAI models and applications.	Systematic literature review of 91 articles on XAI, focusing on methodologies, applications in healthcare, manufacturing, transportation, finance.	Developments and applications of XAI in various fields; need for transparency and accountability in AI systems; roadmap for future XAI research.	Need for increased transparency and accountability in AI systems; future directions for research in XAI applications across different sectors.
Vilone et al. (2021)	Classification of XAI methods based on their output formats.	Systematic review and hierarchical classification of XAI methods, focusing on explanation formats.	Practical classification system for selecting explanation formats; challenges in current explanation formats; future research directions in improving XAI evaluation.	Need for improved explanation formats; challenges in selecting appropriate XAI methods for specific problems.
Das et al. (2020)	Opportunities and challenges in XAI: A survey.	Taxonomy and categorization of XAI techniques, mathematical summaries of seminal work, evaluation of explanation maps generated by XAI algorithms.	Various XAI techniques and their applications; limitations and future directions for improving XAI evaluation and trustworthiness.	Complexity and "black box" nature of AI models; limitations in current XAI approaches for critical domains like healthcare.
Arrieta et al. (2020)	Concepts, taxonomies, opportunities, and challenges in XAI for responsible AI.	Literature review, taxonomy of XAI contributions, historical timeline of landmark studies, evaluation of explanation maps from XAI algorithms.	Overview of XAI concepts, taxonomies, and challenges; emphasis on responsible AI deployment; future directions for advancing XAI in critical domains.	Challenges in integrating XAI into critical domains; future prospects for responsible and accountable AI systems.
Guidotti et al. (2018)	Methods for explaining black box models in decision support systems.	Classification of problems, definitions of interpretability, approaches for providing explanations for black box systems.	Categorization of approaches to open black box models; perspectives on interpretability and explanation in various application domains.	Need for comprehensive solutions for interpreting black box models; ongoing research questions in interpretability and explanation.
Linardatos et al. (2021)	Review of machine learning interpretability methods within the field of XAI.	Literature review, taxonomy of interpretability methods, links to programming implementations.	Various methods for explaining and interpreting machine learning models; complexity and "black box" issues in advanced AI models.	Difficulty in adopting AI in sensitive domains due to lack of interpretability; scientific interest in developing XAI methods for practical deployment.

recent advancements and emphasizing model-semi-agnostic approaches. Our study encompasses a comprehensive review of recent literature, highlighting new methodologies in XAI. We conduct a comparative analysis of state-of-the-art XAI methods within this taxonomy. Additionally, we address key challenges in XAI and propose practical guidelines to mitigate these issues.

Our analysis identifies four critical challenges that must be addressed in the advancement of XAI:

Firstly, achieving a delicate balance between explainability and accuracy is paramount in machine learning. This equilibrium ensures that AI models not only deliver precise outcomes but also provide transparent explanations of their decisions.

Secondly, the efficiency and scalability of XAI methods are essential, especially as datasets and the complexity of applications grow. Scalable methods are necessary to handle large volumes of data and diverse feature sets effectively.

Thirdly, safeguarding XAI against robust and persistent adversarial attacks is crucial for maintaining the reliability and trustworthiness of AI systems. Ensuring that explanations remain robust in the face of adversarial attempts is imperative.

Lastly, integrating XAI seamlessly into existing machine learning frameworks is essential for widespread adoption. This integration streamlines the application of XAI techniques, making them accessible and practical for developers across various domains. Addressing these challenges will pave the way for more transparent, reliable, and effective AI systems in critical applications.

3. XAI in AI-powered applications

The use of machine learning models in many critical applications such as cybersecurity and healthcare has led to growing concerns about the possibility of trusting these models Knight (2021); DND (2021). While promising results (i.e., accuracy) can be achieved using advanced machine learning models such as deep neural networks, these models cannot explain their results Arrieta et al. (2020); Linardatos et al. (2021). In other words, these approaches can extract useful patterns and models behind complex malware and cyber-attacks, but they often cannot explain data correlation in a

way useful for most humans. As a result, people, decision makers, and employees are concerned about the reliability of these "black box" systems when applied to critical decisions related to cybersecurity, especially in important sectors (Rai, 2020). In this section, we focus on the cybersecurity domain, as it is considered one of the most sensitive domains Usman et al. (2019).

For example, as can be seen in Figure 2 (Today), the organization trains a deep neural network on a dataset that consists of malware (or malicious software). Malware is a program written by an attacker to harm computer programs' users, companies, and critical IT infrastructures. It can corrupt benign programs and operating systems, block network connections, steal sensitive and private information from users and organizations, and demand ransom after encrypting critical files on a computer.

The objective is to build an AI-powered malware detection system (Figure 2 - Today). Since some malware are complex and can be obfuscated by the attackers using anti-analysis techniques (Abusitta et al., 2021), the application of deep neural networks is useful to build a robust AI-powered malware detection system (Abusitta et al., 2021). The term "robust" means that it is able to extract effective and robust features that are able to detect malware under changing and obfuscated environments (Abusitta et al., 2021). The output of the training process is a learned model that is able to inform the end user whether a suspicious file is malware or not. Note that the end user does not know what the system did inside before providing the user with the decisions. As a result, the end user may be concerned about the correctness of the decisions, especially if the system is not able to answer important questions such as: Why did you say that? How did you know? Why not something else? How can I trust you?

The aforementioned concerns have led the research community to think about a new AI-powered application, called an "explainable" AI-powered application. As can be seen in Figure 2 (Tomorrow), the system is not only able to say whether a suspicious file is malware or not, it also gives interpretation about why such a decision has been made. In fact, to gain trust in AI-driven systems, it is important to integrate "causality" into AI Knight (2021); DND (2021). The problem of trust in these autonomous systems is complex and multidimensional since it is largely dependent on interaction between humans and machines. Gaining trustworthiness in these complex systems is

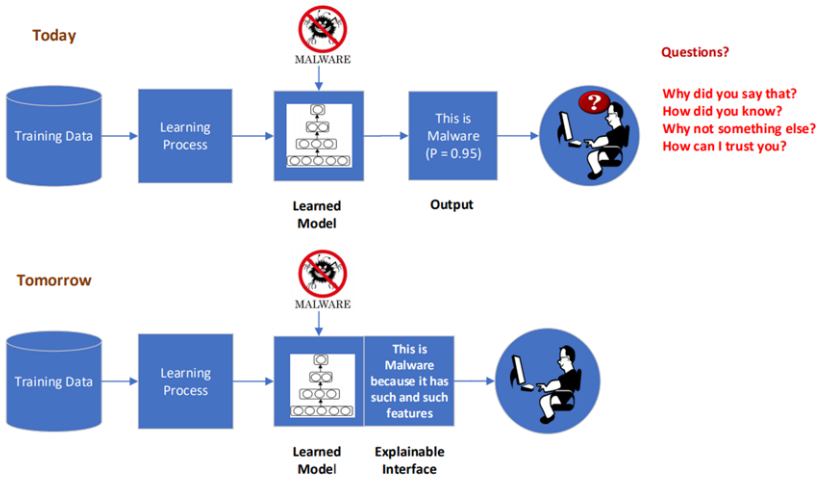

Figure 2: An example of Explainable AI in Cybersecurity: AI-powered Malware Analysis

problematic and requires solutions to encourage acceptance by private and public sectors.

3.1. Other Applications

XAI holds significant promise in the realm of theft detection within machine learning applications. By employing XAI techniques, such systems can offer insights into the decision-making processes of the models, which are pivotal for comprehending and enhancing their efficacy. XAI facilitates the analysis of feature importance, enabling the identification of pivotal variables driving theft indicators. Moreover, it enables the interpretation of complex models, aiding analysts in understanding the rationale behind model predictions. Through anomaly detection explanation, XAI elucidates why certain behaviors are deemed suspicious, offering valuable context for investigations. Additionally, counterfactual explanations generated by XAI shed light on how alterations in input variables could impact theft detection outcomes, fostering proactive prevention strategies. In human-in-the-loop systems, XAI fosters collaboration between analysts and models, promoting iterative refinement and validation of predictions. Furthermore, XAI ensures regulatory compliance by providing transparent explanations for model decisions, aligning with data privacy and accountability mandates. By detecting and

mitigating training data biases, XAI enhances fairness and effectiveness in theft detection, contributing to more reliable outcomes. In essence, XAI serves as a cornerstone in fortifying the transparency, accountability, and efficiency of theft detection systems, empowering analysts to comprehend, validate, and optimize machine learning models in combating theft and fraud.

XAI can also offer valuable tools for enhancing price prediction models in various domains. By employing XAI techniques, analysts gain insights into the factors driving price predictions, enabling more informed decision-making. One significant advantage is feature importance analysis, which helps prioritize relevant variables and optimize data collection efforts. Moreover, XAI facilitates the interpretation of complex models, shedding light on the relationships between input features and predicted prices. With explanations for individual predictions, users can understand the rationale behind model outputs, aiding risk management and strategic planning. XAI also aids in bias detection, ensuring fairer predictions by identifying and mitigating biases in training data. Scenario analysis becomes feasible through XAI's capability to generate counterfactual explanations, allowing analysts to explore the impact of variable changes on price forecasts. Furthermore, human-in-the-loop systems benefit from XAI's interpretability, fostering collaboration between analysts and models for continuous refinement. Lastly, XAI supports regulatory compliance by providing transparent explanations for model decisions, aligning with data privacy and accountability requirements. In essence, XAI empowers analysts to develop more transparent, reliable, and effective price prediction models across various industries, facilitating better decision-making and risk management.

4. Importance, Classification, and Target Audience

This section presents the importance of our approach in surveying and classifying recent findings in Explainable AI (XAI), explains why this classification is crucial, and identifies the target audience who will benefit from it. Our comprehensive survey, which introduces a new taxonomy for XAI techniques, helps organize and synthesize critical information, facilitating easier navigation and understanding of the field. By

addressing key challenges such as balancing explainability and accuracy, efficiency, security, and integration, our work provides valuable guidelines for future research and practical applications. This classification is essential for researchers, practitioners, policy makers, and educators, offering them a structured framework to advance the development and implementation of effective and transparent AI systems.

4.1. Importance of the Approach

The paper's approach of providing a thorough survey and classification of the main and recent findings in XAI is crucial because it focuses on the significance of these findings based on factors like publication venues, citation counts, and impact. This ensures that the survey includes highly influential research, giving readers a curated overview of the most critical developments in XAI. By systematically classifying and comparing these findings, the paper helps organize the vast array of research in this field, making it easier for researchers and practitioners to navigate the literature.

Introducing a new taxonomy for describing and comparing XAI techniques is another vital aspect of this paper. This taxonomy helps in understanding the relationships and differences between various XAI methods, facilitating a more structured and comprehensive view of the field. Such a systematic classification is essential for making sense of the diverse approaches and innovations in XAI, and it aids in identifying gaps and opportunities for further research.

The paper's focus on data-driven machine learning models (DMLM) is particularly relevant as these models are extensively used in real-world applications. By concentrating on XAI techniques specifically designed for DMLM, the paper narrows its scope to the most practical and impactful applications of XAI, ensuring that the findings are directly applicable to contemporary machine learning practices.

Additionally, the paper identifies four major challenges in XAI: balancing explainability and accuracy, ensuring efficiency and scalability, protecting against adversarial attacks, and integrating XAI into existing platforms. Recognizing these challenges and proposing guidelines to address them provides valuable direction for future research and development in XAI, helping to advance the field towards more practical and robust solutions.

Providing a detailed comparative analysis of state-of-the-art XAI methods based on the proposed taxonomy is crucial for understanding the strengths and weaknesses of different approaches. This analysis helps in identifying which methods are best suited for specific applications, thereby aiding researchers and practitioners in making informed decisions about the tools and techniques they use.

4.2. Importance of This Classification

The new taxonomy and comparative analysis presented in this paper serve as a roadmap for future research in XAI. By highlighting underexplored areas and promising directions, the paper guides researchers towards new and innovative solutions. This classification also provides a clear understanding of the current state of XAI, helping to identify gaps and opportunities for further advancements.

For practitioners in the field of machine learning and AI, this classification offers a comprehensive overview of available XAI techniques and their applicability. This information is essential for selecting appropriate methods to enhance the transparency and interpretability of machine learning models, especially in critical applications where understanding the model's decisions is crucial.

Establishing a common taxonomy helps to standardize the terminology and concepts in XAI. This standardization is vital for clear communication and collaboration among researchers and practitioners, facilitating a more coherent and unified development of the field. Such a common framework also aids in the education and dissemination of knowledge about XAI.

Addressing the challenges of explainability, efficiency, security, and integration contributes to making XAI techniques more robust and practical. This enhances the trust and adoption of AI systems in critical applications where interpretability is paramount, such as healthcare, finance, and autonomous systems. By improving these aspects, the paper helps ensure that AI systems are both effective and explainable, thereby increasing their acceptance and use.

4.3. Audience

Researchers in the field of AI and machine learning benefit greatly from the comprehensive survey, new taxonomy, and identification of open issues provided by this

paper. This information guides their investigations and contributes to the advancement of the field, helping them to focus on the most critical challenges and opportunities in XAI.

Practitioners and developers in AI and machine learning gain valuable insights from this paper into the most effective XAI techniques and how to apply them. The guidelines provided can help them integrate XAI methods into their existing workflows, improving the transparency and accountability of their models. This practical guidance is essential for enhancing the usability and effectiveness of AI systems in various applications.

Policy makers and regulators concerned with the ethical and transparent use of AI can use the findings and guidelines from this paper to inform policies and standards for AI deployment. Ensuring that AI systems are both effective and explainable is crucial for maintaining public trust and compliance with ethical standards, and this paper provides the necessary insights to support these goals.

Educators and students in AI and machine learning can use this survey as a valuable learning resource. The structured presentation of information aids in teaching and understanding the current state of XAI, its challenges, and future directions. This resource is essential for developing a new generation of researchers and practitioners who are well-versed in the principles and practices of explainable AI.

5. Taxonomy of Explainable Machine Learning

We present in this section the proposed taxonomy of explainable AI. Figure 3 shows that XAI techniques can be divided into two categories: interpretable models and post-hoc interpretations. The post-hoc interpretation itself is divided into three categories: model-agnostic, model-specific, and model-semi-agnostic techniques.

5.1. Interpretable models

A machine learning model is considered interpretable (or transparent) if its predictions can be understood easily Vellido et al. (2012); Guidotti & Ruggieri (2019). In the literature, there are several interpretable models such as linear/logistic regression

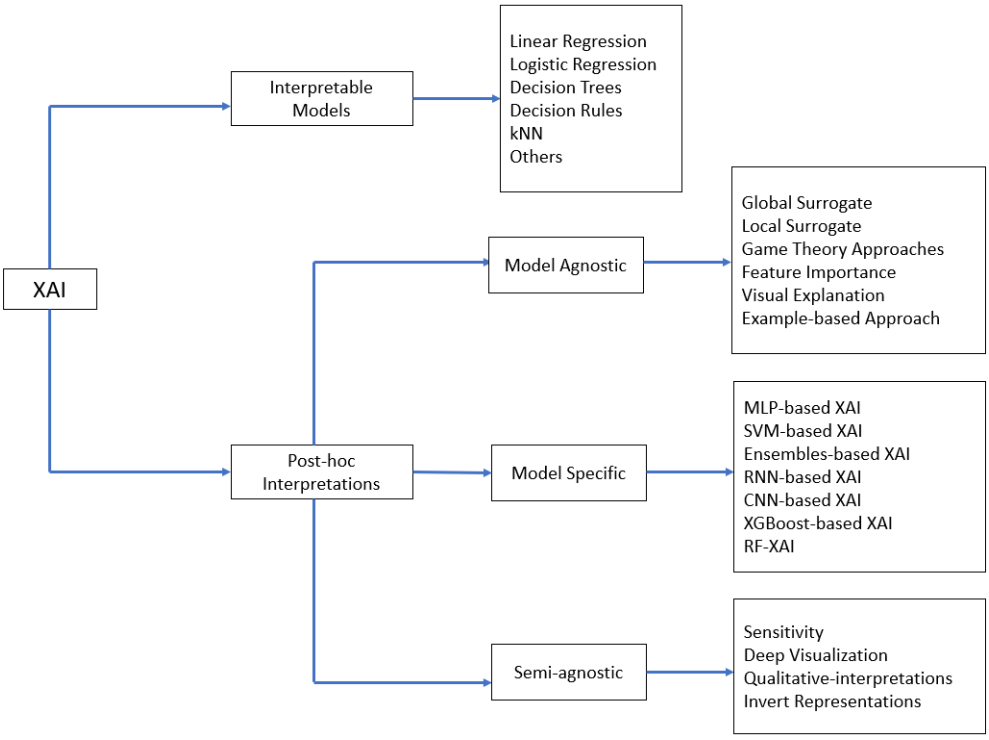

Figure 3: The proposed taxonomy

(Hoffrage & Gigerenzer, 1998; Bursac et al., 2008; Jaccard & Jaccard, 2001; Peng et al., 2002), decision trees (DT) (Laurent & Rivest, 1976; Utgoff, 1989; Quinlan, 1986; Maimon & Rokach, 2014; Rovnyak et al., 1994; Nefeslioglu et al., 2010), and rule-based learning (Johansson et al., 2004a; Quinlan, 1987). In order for a machine learning model to be interpretable, at least one of the following three properties have to be achieved: Simulatability, Decomposability, and Algorithmic Transparency (Lipton, 2018). Below we describe these properties.

Simulatability. Simulatability is the property that someone (i.e., a human) can go through the machine learning model's steps and understand them Peng et al. (2022); Lage et al. (2019). Therefore, a human can check whether each model's step is reasonable to her/him or not. Logistic/linear models, decision trees, rule based models, and Bayesian models can be considered interpretable models based on the above definition. The computation that these models need in order to make decisions and/or

to predict something is easy to understand and interpret. For example, in the linear model, there is a direct mapping between two correlated variables (dependent and independent variables). The linear relationship between these variables makes this model understandable and hence interpretable. Moreover, a decision tree with a few nodes is understandable if someone traverses the tree.

One should note that if the machine learning model is not simulatable, it would be very difficult to understand, even if it is simple. For example, a huge number of simple rules would not allow a human to guess its predictions Vaishak & Ioannis (2021).

Decomposability. Decomposability is testing interpretability with respect to a model's sub-components Slack et al. (2019); Lu et al. (2019). For example, consider a DT, whose nodes correspond to understood factors. The prediction of a DT can be easily explained in terms of what decisions are made at the tree's various nodes.

Algorithmic Transparency. A machine learning algorithm is considered explainable if its decisions are explainable by humans Peng et al. (2022); Ahmad et al. (2018). In other words, it consists of preferred properties that can be understood by a human. To this end, the algorithm should rely on mathematical techniques and tools to provide proof and to make the model easy to understand Peng et al. (2022). Examples of algorithmic transparency are logistic/linear models and Bayesian models.

5.2. Post-hoc Interpretation

In the previous section, we have seen that the simplest way to obtain explainable machine learning is to use simple and/or linear models such as linear/logistic regression and decision trees. These models have good interpretability, but the accuracy obtained using these models is too low, especially when applied on complicated data Arrieta et al. (2020); Samek et al. (2017).

As a result, there is a pressing need for post-hoc interpretation, which are methods used to provide interpretation to complex models such as deep neural networks. We classify post-hoc interpretation into three types: model-agnostic, model-specific, and model-semi-agnostic techniques.

5.2.1. Model-agnostic techniques

Model-agnostic techniques are techniques that can be used with any machine learning models. In particular, these techniques only consider the inputs and outputs of machine learning models and try to understand the relation between them. Below we describe the main model-agnostic techniques in the literature.

Permutation Feature Importance. The permutation feature importance method, as elucidated in several studies (Datta et al., 2016; Adler et al., 2018; Koh & Liang, 2017), offers a robust approach to assessing the significance of features in machine learning models. The fundamental concept involves determining the importance of a feature by measuring the increase in prediction error of the machine learning model when the feature's values are randomly permuted (Fisher et al., 2019).

In essence, the rationale is straightforward: if shuffling the values of a feature results in a notable increase in the model's error, it indicates that the model heavily relies on that feature for accurate predictions (Fisher et al., 2019). Conversely, if there's minimal change in the model's error upon shuffling the feature's values, it suggests that the feature is less influential in the predictive process and may be deemed "unimportant."

Moreover, the determination of feature importance can also be approached through metrics like 'Gain,' a method commonly utilized in prominent machine learning libraries such as 'scikit-learn' for various decision tree-based algorithms. This approach is also the default method employed to compute feature importance in frameworks like XGBoost Abu-Rmileh (2019). By quantifying the gain or loss in predictive performance associated with each feature, practitioners gain insights into the relative significance of different features within their models, thereby facilitating informed decision-making in feature selection and model refinement processes.

The main advantage of the permutation feature importance approach is that it gives good and reasonable interpretations. In fact, it makes sense that the feature importance is related to the increase in a machine learning model's error (Molnar, 2020; Adadi & Berrada, 2018). However, the main disadvantage of the permutation feature importance approach is that there is no clear vision about the kind of data that should be used. In other words, it is unclear whether we should use test data or training data to determine

the feature importance. Another disadvantage of this approach is that in most cases, features are mathematically correlated, which makes the permutation feature importance biased by unrealistic data instances (Rudin, 2019; Molnar, 2020). In addition, it is a time-consuming process to permute the whole set of features and run the model many times.

It is worth mentioning here that Rudin (2019) urges to stop explaining black box machine learning models for high stakes decisions and use inherently interpretable models instead, because the black box model may cause great harm to society Rudin (2019). We do not fully agree with this message because the use of inherently interpretable models tend to produce inaccurate decisions, especially if applied to complex systems (e.g., healthcare, criminal justice). Moreover, there are not enough practical results (evidences) that show interpretable models could replace black box models in critical applications. Having inaccurate results would also cause great harm to society.

Surrogate Models. This XAI approach exploits an interpretable model (e.g., Decision Tree) to interpret an uninterpretable black-box model. To achieve this, an interpretable model should be trained to approximate the decisions provided by the black box model. After training, we measure how close the interpretable machine learning model (surrogate model) is to the black-box model.

The surrogate model approach has been adopted by several works (e.g., (Craven, 1996; Ribeiro et al., 2016b; Su et al., 2015; Che et al., 2016)). The overall steps to achieve a surrogate model are as follows. First, select a dataset D, which has been adopted for training the black box. Secondly, obtain the prediction of the model (i.e., black-box model) using the selected data D. Thirdly, choose any type of interpretable models (e.g., DT). Fourthly, train the interpretable model on the same data D and find its prediction. Finally, the interpretable model becomes the surrogater that can be used to explain the original model Danilevsky et al. (2020); Islam et al. (2021). Figure 4 shows an example of the surrogate model. In Figure 4, the decision tree is trained to approximate the decisions provided by the deep neural networks

One should note that the surrogate model can be applied globally (as shown above) or locally (i.e., local surrogate model) (Danilevsky et al., 2020; Islam et al., 2021; Mol-

nar, 2020). In the local surrogate model, the objective is to discern why the AI model made a certain or specific prediction. The steps used to achieve this are as follows. First, select the instances (points) for which you would like to have an interpretation of the black-box prediction. Secondly, add noise to your data and obtain the predictions of the black box using the new points. Thirdly, find the weight of the new samples based on their closeness to the instances. Fourthly, train the interpretable model on the data with the variations. Finally, interpret the results (predictions) by explaining the generated local model (Danilevsky et al., 2020; Islam et al., 2021; Molnar, 2020). There are two advantages of using the surrogate model: flexibility and diversity in explanation. In fact, any interpretable model can be adopted as a surrogate model Seungjun (2022), so it provides users with flexibility. Besides, multiple interpretable models can be used to explain the black-box model, which leads to diversity in explanation. The diversity in explanation allows us to better understand black-box models Seungjun (2022).

The disadvantage of this kind of method is that the surrogate models cannot perfectly approximate the black-box model. Their expressive ability is more limited than the black-box models, which are hard to explain by themselves.

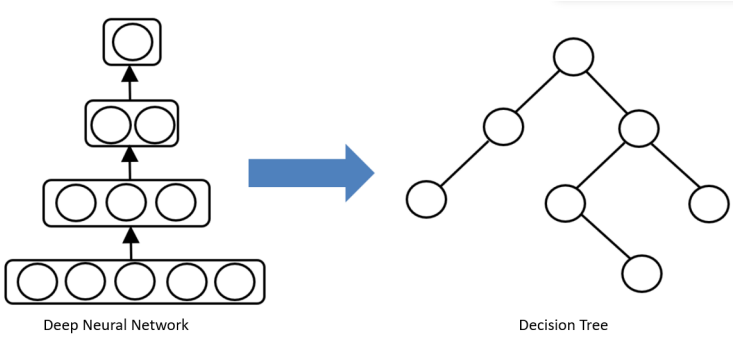

Figure 4: The surrogate model

Game theoretic-based model. The idea of a game theoretic-based explainable model is to consider each feature value as a player in a game (e.g., (Lundberg & Lee, 2017; Amoukou et al., 2021; Messalas et al., 2019). In other words, a machine learning prediction can be interpreted by considering each value in the feature as a player and

the prediction as the payout. To this end, the Shapley value concept is adopted (Winter, 2002; Roth, 1988), which is a method used in the literature of cooperative game theory that calculates the (weighted) average of all the marginal contributions of players (i.e., feature values) to all possible coalitions.

The game theoretic-based model provides a mechanism to measure the importance of each feature with respect to a given prediction. Specifically, it explains the prediction of an input x by computing individual feature contributions towards that output prediction. This can be done by formulating the feature values as players in a coalition game Tallón-Ballesteros & Chen (2020); Fryer et al. (2021). Next, the Shapley values are computed for all feature values. The feature that has the highest Shapley value is considered the most important feature and vice versa Fryer et al. (2021).

The main advantage of the game theoretic-based model is that it provides fair distribution among the feature values, that is, the explanation is not biased against any feature Fryer et al. (2021); Molnar (2020). Moreover, since this approach is based on game theory, it provides a solid theoretical guarantee Lundberg et al. (2020). However, the main limitation of this approach is that it requires a lot of computation time. This is due to the fact that calculating the Shapley values is intractable if we have a huge number of features (Roth, 1988).

Visualization approach. The visualization approach adopts visualization techniques to achieve explainable machine learning. In the literature, there are two famous visualization-based methods: partial dependence plot and accumulated local effect plot. The partial dependence plot is used to show the marginal effect of features (one or two features) on the machine learning model's outputs Friedman (2001). On the other hand, the accumulated local effect plot shows how features affect the model's outputs on average Apley & Zhu (2020). Recently, Goldstein et al. (2015) proposed a visualization tool based on Individual Conditional Expectation (ICE) plots, which can be used to visualize the model learned by any supervised machine learning algorithm. Cortez & Embrechts (2011) designed a set of visualization methods based on Global Sensitivity Analysis, which provides an explanation of a black-box machine learning model. Similarly, Cortez & Embrechts (2013) proposed three visualization techniques based on

data-based, Monte-Carlo, and cluster-based Sensitivity Analysis. They also proposed a new visualization method based on Average Absolute Deviation (Zhang, 2016). While visualization techniques could be used to achieve an understandable explanation of machine learning through the uncovering of heterogeneous relationships of features, they are generally unable to display (or visualize) all features and/or the relationship between features.

Example-based approach. The example-based approach provides examples that can be used to gain insights into the machine learning predictions. For instance, Kanehira & Harada (2019) show examples that reflect the capability of the learned model in distinguishing instances, even if the instances are slightly different. Koh & Liang (2017) have shown the ability to identify instances (from the training data) that had significant influence on the model prediction. The main advantage of the example-based approach is that it gives a clear and reasonable interpretation. This is due to the fact that it helps us construct logical (or mental) models of the AI models using examples (Molnar, 2020; Cai et al., 2019). However, the main limitation of this approach is that in some cases, we may have multiple inconsistent interpretations (Cai et al., 2019; van der Waa et al., 2021; Molnar, 2020).

5.2.2. Model-specific techniques

In the previous section, we presented model-agnostic approaches for interpretation, which can be used to explain any machine learning model. We have shown that these approaches provide flexibility to the AI developers by letting them only focus on the input and output of the learned model. In the literature, there are other techniques that are designed only for specific machine learning models (model-specific techniques). For example, Chaves et al. (2005) created a new approach that allows us to produce fuzzy rules for the support vectors of the learned model of the SVM (Jakkula, 2006). Similarly, Barakat & Bradley (2007) and Barakat & Diederich (2008) proposed using rule-based models instead of fuzzy rules. These models allow them to create rule-based models from the support vectors of the model that has been trained.

Yu et al. (2021) proposed an XAI method for random forest. They introduced a framework to overcome the feature importance bias caused by small datasets. To

this end, they propose to use feature interaction networks, reveals hidden interactional factors, and boosts model interpretability Yu et al. (2021).

Zeiler et al. (2011), Zeiler et al. (2010), and Zeiler & Fergus (2014) proposed a new network called "DeconvNet". The purpose of this network is to reconstruct the maximum activation at each layer output in a Convolutional Neural Network (CNN) (O'Shea & Nash, 2015). Having these reconstructions enables us to form a useful idea about which parts of the data (e.g., images) generated that activation. In other words, it gives an idea about the parts of data that have an overall effect.

Bach et al. (2015) proposed a visualization technique to vistualize the contribution of each pixel of the image to the prediction. To this end, they adopted a Layer-wise Relevance Propagation method, which is used to find points that are close to the predicted point. Among other approaches, Zhang et al. (2018) proposed using a loss function for each CNN's filter. This is useful to allow each filter to learn a specific component and thus enhance the network explainability. Their results show that the learned patterns are much more explainable compared to the vanilla CNN.

In the context of recurrent neural networks (RNNs) (Medsker & Jain, 2001; Choi et al., 2016) proposed a REverse Time AttentIoN (RETAIN) framework, which is used to detect influential previous patterns through the adoption of a two-level attention model. Similarly, Wisdom et al. (2016) integrated the Sequential Iterative Soft-Thresholding Algorithm (SISTA) into the RNN to build an interpretable RNN. The SISTA helps model the sequence of connected observations with the other sequence called Sparse Latent Vectors (SLV). This, in turn, makes the RNNs much more interpretable. Similarly, Krakovna & Doshi-Velez (2016) integrate Hidden Markov Model (HMM) into the RNN to achieve a trade-off between accuracy (obtained from RNN) and interpretability (obtained from HMM).

A visualization method for building explainable RNNs is proposed by Karpathy et al. (2015). The proposed framework is based on n-grams that are used to distinguish explainable cells within GRUs (Gated Recurrent Units) and Long Short-Term Memory (LSTM).

For Multilayer Perceptrons (MLPs), Shrikumar et al. (2016) proposed a new framework called DeepLIFT, which is used to compute the difference between a neuron's

activation and its predefined reference activation. By comparing the difference, the contribution scores of each neuron can be assigned. In the context of the simplification approach, Che et al. (2016) proposed a new method that is used to obtain an explainable model using gradient trees (Che et al., 2016). Similarly, Thiagarajan et al. (2016) present a new approach for building features hierarchically. This new format simplifies the MLP model and thus makes it more interpretable. (Montavon et al., 2017) used the approach of feature relevance by decomposing the MLP's prediction into corresponding inputs. To this end, they assume that each MLP's neuro is a decomposable object (Montavon et al., 2017).

Other specific-based techniques in the literature use a concept-based approach. The goal of the concept-based approach is to quantify the importance of features that are defined by a user's concept (high-level concept). For example, Kim et al. (2018) proposed Concept Activation Vectors (CAVs), which are used to explain a neural network's state in terms of human concepts (Kim et al., 2018). In other words, they look into the neural network's high-dimensional internal state. Recently, Akula et al. (2020) proposed Conceptual and Counterfactual Explanations (CoCoX), which is a model used to explain decisions generated by a CNN. CoCoX interprets decisions made by a CNN using fault-lines, which provide explanations that are easy to understand by humans.

In the context of cybersecurity, Li et al. (2021b) proposed an Interpretable MAlware Detector (I-MAD), which provides interpretability while at the same time achieving state-of-the-art performance on malware detection. Its top layer classification module is called an interpretable feed-forward neural network (IFFNN). It has the expressive ability of a multi-layer feed-forward neural network and the interpretability of logistic regression through providing an impact value of each feature related to the classification result Li et al. (2021a).

In summary, the main advantage of model-agnostic XAI over model-specific XAI is that it gives flexibility to machine learning developers. This is due to the fact that it can be applied on any model. However, model-specific XAI has an absolute advantage in giving high performance when applied to a specific model.

5.2.3. Model-semi-agnostic techniques

The model-semi-agnostic techniques are techniques that can be applied to some classes (families) of machine learning models. Most of these techniques are used to provide interpretations to different deep learning architectures. For example, Sundararajan et al. (2017) proposed an axiomatic approach (sensitivity) for attributing the prediction of a deep neural network. This approach can be applied on any deep network architecture. The proposed approach is based on integrated gradients that attribute the prediction of a deep network to its inputs.

Another model-semi-agnostic technique is proposed by Erhan et al. (2009). They propose a new framework for obtaining useful qualitative interpretations of high-level features, which are represented by deep networks. To do this, they compared many methods applied on Deep Belief Networks (Hinton et al., 2006; Hinton, 2009) and Stacked Denoising Autoencoders (Vincent et al., 2010) that were trained on a set of datasets. Their results show that interpretation would be possible at the unit level.

Mahendran & Vedaldi (2015) proposed a new framework for understanding deep data representations by inverting them. To this end, they proposed a general method to invert representations. They also studied the inverse of recent CNNs' image representation. Their results show that the visualisations shed light on the data represented at each layer. Similarly, Dosovitskiy & Brox (2016) show that inverting a deep network trained on a large dataset (e.g., ImageNet) provides many useful insights into the characteristics of the representation of features learned during the training. More specifically, they show that some features, such as colors, can be reconstructed in higher layers by applying activation.

Yosinski et al. (2015) propose two tools for understanding deep networks through deep visualization. The first tool is used to visualize the activation generated on each network's layer. They have found that by observing the change of activation according to user input, it would be possible to perceive how the deep network works. The second tool allows the visualization of features at each network's layer using regularized optimization. They presented some regularization techniques that can be combined to generate more interpretable visualizations.

6. Characterization of Surveyed Papers

This section characterizes each reviewed paper. Table 2 provides information about the approach and category used for each paper and provides more details about the main procedure(s) used for each paper.

Table 2: Comparison Summary.

Begin of Table			
Work(s)	**Category**	**Approach**	**Description**
Letham et al. (2015) Kim et al. (2014)	Transparent	Specific	Uses Bayesian model for interpretation.
Caruana et al. (2015)	Transparent	Specific	Uses generalized linear model for interpretation.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Ribeiro et al. (2016a) Aung et al. (2007) Tan et al. (2018)Johansson et al. (2004a) Johansson et al. (2004b)Konig et al. (2008) Lakkaraju et al. (2017) Mishra et al. (2017) Su et al. (2015) Ribeiro et al. (2016b)	Posthoc	Model-agnostic	Uses interpretable models to explain black-box model.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Craven (1996) Domingos (1998)Zhou & Hooker (2016) Thiagarajan et al. (2016) Johansson & Niklasson (2009) Craven (1996)Bastani et al. (2017)	Posthoc	Model-agnostic	Uses Decision Tree to explain black-box model.
Che et al. (2016)Hooker (2004)	Posthoc	Model-agnostic	Uses interpretable models to simplify black-box models.
Datta et al. (2016) Adler et al. (2018) Koh & Liang (2017)	Posthoc	Model-agnostic	Uses influence function to trace the prediction of a given model back to the training data. Thus, identifies the main training points.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Cortez & Embrechts (2011) Cortez & Embrechts (2013)	Posthoc	Model-agnostic	Uses sensitive analysis (Sensitivity) to measure the importance of features.
Lundberg & Lee (2017) Strumbelj & Kononenko (2010) Chen et al. (2021)	Posthoc	Model-agnostic	Uses the concept of game theory (Shapley values) to calculate the marginal contribution of features.
Fong & Vedaldi (2017) Dabkowski & Gal (2017)	Posthoc	Model-agnostic	Uses image saliency method to identify unique features
Henelius et al. (2017) Henelius et al. (2014)	Posthoc	Model-agnostic	Uses feature interaction method to explain machine learning models.
Krause et al. (2016)	Posthoc	Model-agnostic	Uses local explanations to explain machine learning models.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Lundberg & Lee (2017) Baehrens et al. (2010)	Posthoc	Model-agnostic	Uses rule-based learning to understand why the model made a certain prediction (local explanation).
Ribeiro et al. (2018)	Posthoc	Model-agnostic	Uses rules to understand why the model made a certain prediction (local explanation).
Robnik-Šikonja & Kononenko (2008)	Posthoc	Model-agnostic	Uses visualization methods to understand the prediction of black-box models
Martens & Provost (2014) Chen et al. (2017)	Posthoc	Model-agnostic	Uses interpretable models to understand why the model made a certain prediction (local explanation).
Che et al. (2016) Lundberg & Lee (2017) Goldstein et al. (2015) Casalicchio et al. (2018)	Posthoc	Model-agnostic	Uses conditional and Shapley plots for visual explanation.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Fong & Vedaldi (2017)Cortez & Embrechts (2011)Cortez & Embrechts (2013)	Posthoc	Model-agnostic	Uses Sensitive Analysis for visual explanation.
Dabkowski & Gal (2017)	Posthoc	Model-agnostic	Uses Salience Analysis for visual explanation.
Xu et al. (2018) Henelius et al. (2017) Apley & Zhu (2020) Hoffrage & Gigerenzer (1998) Robnik-Šikonja & Kononenko (2008)	Posthoc	Model-agnostic	Uses visualization techniques to find important features.
Tan et al. (2020) Deng (2019) Hara & Hayashi (2016)	Posthoc	Model-agnostic	Uses Decision Tree to explain ensembles and Multiple Classifier Systems (MCS).

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Palczewska et al. (2014) Welling et al. (2016) Tolomei et al. (2017) Auret & Aldrich (2012)	Posthoc	Model-specific	Uses feature importance to explain ensembles and MCS.
Welling et al. (2016) Auret & Aldrich (2012) Rajani & Mooney (2018a) Rajani & Mooney (2018b)	Posthoc	Model-specific	Uses Variable Importance to explain ensembles and MCS.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Barakat & Diederich (2008) Barakat & Bradley (2007) Chaves et al. (2005) Fu et al. (2004) Zhang et al. (2005) Navia-Vázquez & Parrado-Hernández (2006) Nunez et al. (2006) Chen et al. (2007) Núñez et al. (2002)	Posthoc	Model-specific	Uses Rule-based learning to explain SVM.
Sollich (2002) Sollich (1999)	Posthoc	Model-specific	Uses probabilistic models to explain SVM.
Haasdonk (2005)	Posthoc	Model-specific	Uses a surrogate model based on geometric interpretation to explain SVM.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Gaonkar et al. (2015) Landecker et al. (2013) Rosenbaum et al. (2011)	Posthoc	Model-specific	Uses Feature Contribution to explain SVM.
Üstün et al. (2007) Jakulin et al. (2005)	Posthoc	Model-specific	Uses Internal Visualization to explain SVM.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Augasta & Kathirvalavakumar (2012) Zhou et al. (2003)Craven & Shavlik (1994) Fu (1994) Towell & Shavlik (1993) Thrun (1995) Setiono & Leow (2000)Taha & Ghosh (1999) Tsukimoto (2000) Arbatli & Akin (1997)	Posthoc	Model-specific	Uses Rule-based learning to explain Multi-Layer Neural Networks.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Craven (1996) Che et al. (2016) Wu et al. (2018) Frosst & Hinton (2017) Krishnan et al. (1999) Thiagarajan et al. (2016) Zilke et al. (2016) Schmitz et al. (1999) Hinton et al. (2015) Adebayo et al. (2018) Papernot & McDaniel (2018)	Posthoc	Model-specific	Uses Decision Tree to explain Multi-Layer Neural Networks.
Montavon et al. (2017)	Posthoc	Model-specific	Uses deep Taylor decomposition to explain Multi-Layer Neural Networks.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Li et al. (2021b)	Posthoc	Model-specific	Uses a specific neural network architecture to make it interpretable as logistic regression.
Kindermans et al. (2017) Shrikumar et al. (2016) Féraud & Clérot (2002) Shrikumar et al. (2017) Féraud & Clérot (2002)	Posthoc	Model-specific	Uses feature importance to explain Multi-Layer Neural Networks.
Sundararajan et al. (2017) Krishnan & Wu (2017)	Posthoc	Model-specific	Uses Sensitive Analysis to explain Multi-Layer Neural Networks.
Kim et al. (2018) Dong et al. (2017)	Posthoc	Model-specific	Uses activation clusters (concept-based approach) to explain Multi-Layer Neural Networks.
Lei et al. (2016) Li et al. (2015)	Posthoc	Model-specific	Uses Caption Generation to explain Multi-Layer Neural Networks.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Papernot & McDaniel (2018)	Posthoc	Model-specific	Uses Salience Analysis for visual explanation of Multi-Layer Neural Networks.
Tan et al. (2015) Rieger et al. (2020) Zhang et al. (2019)	Posthoc	Model-specific	Uses Architecture Modification to explain Multi-Layer Neural Networks.
Bach et al. (2015)	Posthoc	Model-specific	Uses Decision Tree to explain Convolutional Neural Networks (CNN).
Nguyen et al. (2016) Samek et al. (2017) Bach et al. (2015)	Posthoc	Model-specific	Uses activations to explain CNN.
Akula et al. (2020)	Posthoc	Model-specific	Uses concept-based approach to explain CNN.
Nguyen et al. (2016)	Posthoc	Model-specific	Uses Feature Extraction to explain CNN.
Sundararajan et al. (2017)	Posthoc	Model-semi-agnostic	Uses Axiomatic attribution to explain the family of deep neural networks.

Continuation of Table 2			
Work(s)	**Category**	**Approach**	**Description**
Erhan et al. (2009)	Posthoc	Model-semi-agnostic	Uses visualization techniques to explain the family of deep neural networks.
Mahendran & Vedaldi (2015)	Posthoc	Model-semi-agnostic	Uses Inverting Method (IM) to explain the family of deep neural networks.
Yosinski et al. (2015)	Posthoc	Model-semi-agnostic	Uses Visualization techniques to explain the family of deep neural networks.
Bonifazi et al. (2024)	Posthoc	Model agnostic	Leveraging network theory for supporting XAI on classifiers.
End of Table			

7. Metrics for XAI

XAI aims to create models whose decisions can be understood and trusted by humans. Evaluating the effectiveness of explanations provided by XAI models is crucial for their deployment in real-world applications. Several metrics have been proposed to assess the quality of explanations, which can be broadly categorized into objective and subjective metrics.

7.1. Objective Metrics

Objective metrics are quantitative measures that do not rely on human judgment. These metrics often focus on the fidelity, completeness, and accuracy of the explana-

tions.

Fidelity:. Fidelity measures how accurately the explanation reflects the behavior of the model. An explanation with high fidelity accurately represents the decision-making process of the model Ribeiro et al. (2016a). Fidelity can be quantified using metrics such as cosine similarity or mean squared error between the predictions of the original model and the simplified model.

Completeness:. Completeness evaluates whether the explanation includes all the relevant information used by the model to make a decision Ribeiro et al. (2016a).

Accuracy:. Explanation accuracy measures how correct the explanation is in describing the underlying model Ribeiro et al. (2016a).

7.2. Subjective Metrics

Subjective metrics involve human judgment and measure the perceived quality of explanations from the user's perspective. These metrics include interpretability, trust, and usability.

Interpretability:. Interpretability assesses how easily a human can understand the explanation Lipton (2016).

Trust:. Trust measures the degree to which users believe and rely on the explanations provided by the model Doshi-Velez & Kim (2017).

Usability:. Usability evaluates how practical and helpful the explanations are for users in making decisions or gaining insights Doshi-Velez & Kim (2017).

7.3. Hybrid Metrics

Some metrics combine both objective and subjective elements to provide a comprehensive evaluation of explanations. For instance, Doshi-Velez and Kim (2017) propose a framework that includes both human-grounded evaluation and application-grounded evaluation, merging quantitative and qualitative aspects to assess explanations Doshi-Velez & Kim (2017).

8. Challenges/Issues and Improvement in XAI

The works presented in this article are somehow effective in achieving XAI. However, there are still some challenges that need further investigation. In this section, we have identified the following four research challenges/issues. The selection of these challenges is based on 1) our gaps analysis of the existing XAI techniques; 2) our discussion and work with our industry partners.

8.1. The balance between explainablity and accuracy

The world is facing a historical shift toward automating almost everything in our life Abusitta et al. (2019). As a result, there is a pressing need to find robust explainable machine learning algorithms. In fact, most of the effective XAI methods proposed in the literature are ”not built-in,” meaning the XAI method can be applied only after the machine learning algorithm has already made the decision Došilović et al. (2018). One should note that it is always better to have a machine learning algorithm that is interpretable in nature (like a linear model) and is able to provide high accuracy (like a deep neural network). Finding the balance between accuracy and explainability is challenging. To achieve that, we argue researchers to find a unified framework that integrates interpretable models, which guarantee the explainability, and deep learning models, which somehow guarantee high accuracy. Although surrogate models are somehow able to do that, these models, however, are just trying to approximate the decision of deep learning models using interpretable models to enhance the explainability. This makes the resultant accuracy closer to interpretable models (e.g., decision tree) than deep learning models.

8.2. Integration of XAI into Integrated Development Environments for machine learning

It is important for machine learning developers to integrate XAI techniques (e.g., surrogate models) into their AI projects. This can be achieved by integrating XAI techniques into the Integrated Development Environments (IDEs) for machine learning (e.g., RStudio). Machine learning developers then can use these techniques by calling and importing XAI-related libraries into their work space. An IDE should also provide

developers with the ability to evaluate to what extent their learned models are able to explain decisions, with recommendations on how to better achieve explainability without loss of accuracy. To achieve that, we suggest designing XAI-related APIs that can be called to evaluate the explainability of the learned model (e.g., deep neural network) using a surrogate model. The called API should evaluate each interpretable model (e.g., decision tree) by training it and approximate the decisions provided by the learned model (see surrogate models in Section 5.2.1). After training, the API can measure how close the interpretable model is to the learned model. By doing that for every interpretable model, the API would be able to find the best interpretable model that can be used to explain the learned model, without loss of accuracy.

8.3. Adversarial attacks on XAI

Adversarial attacks are actions that can be performed by attackers to fool the machine learning models into making erroneous decisions de Mello (2020); Chakraborty et al. (2018). For example, an attacker can generate adversarial examples which are inputs designed to cause the learned model to make a mistake (an incorrect decision). Adversarial examples can also fool explainable machine learning models by making them produce incorrect and inconsistent explanations (Camburu, 2020; Kuppa & Le-Khac, 2020). Adversarial attacks on XAI can be defined as actions that can be performed by attackers to fool the machine learning models into making erroneous explanations. For example, assume there is a deep neural network model that is trained to classify animals. The model is also trained to give explanations of its classifications (e.g., this is a cat because it has fur, claws, and whiskers). An attacker can generate adversarial examples that make the deep neural network model generate a mistake not only in the classification but also in the explanation (e.g., this is a fish because it has fur, claws, and whiskers!). Recently, Dombrowski et al. (2019) showed that machine learning models can be manipulated (producing un-understandable explanations) using perturbed images. Specifically, they experimentally showed the possibility of manipulating explanations by adding perceptible perturbations to the input that maintain the neural network model's output unchanged. Similarly, Heo et al. (2019) propose a new approach to fooling XAI by fine-tuning the model to undermine its explainability. While sev-

eral mechanisms have been proposed to protect the machine learning models against adversarial examples (e.g., adversarial training (Song et al., 2018) and model aggregation Rieger & Hansen (2020)), these solutions are generally not mature enough to preserve the balance between accuracy and explainability. Therefore, we suggest that the solution for defending against adversarial attacks on XAI should not only protect the model against adversarial attacks, but also preserve or enhance the overall accuracy. To achieve that, we suggest conducting research on finding the best unified framework that integrates adversarial training, which is one of the famous approaches in the literature used for mitigating adversarial attacks, into interpretable models and deep learning models.

8.4. Efficiency and scalability

A practical explainable machine learning can help machine learning developers interpret predictions on-the-fly by considering the scalability of XAI methods. It is important to note that scalability is an important factor, especially when we adopt model-agnostic XAI methods, as they are usually expensive and time-consuming (e.g., game-theory based XAI). These methods should be scalable as the number of features in the database may scale up in many applications. In a practical AI application, an explainable AI tool's scalability and efficiency should be evaluated using a large database that consists of a large number of features in order to measure both accuracy and latency. To achieve scalable model-agnostic XAI, we suggest that heuristic algorithms are used (e.g., linear approximation Fatima et al. (2008)) instead of using the original expensive algorithms. For example, in game theory-based explainable AI, the main challenge in adopting the Shapley value to real-world XAI systems is its computational time which increases exponentially with the number of features van Campen et al. (2018). The approximation method can approximate the value of the Shapley value and produce a value that is very close to it Fatima et al. (2008); van Campen et al. (2018).

8.5. Enhancing XAI using Large Language Models

Large Language Models (LLMs) hold immense potential to revolutionize the eXplainable Artificial Intelligence (XAI) field by enhancing interpretability, transparency,

and trust in AI systems Zhao et al. (2024). By generating human-readable explanations for their outputs, LLMs enable users to understand the rationale behind AI decisions, fostering trust and accountability. Moreover, they can provide domain-specific explanations tailored to diverse user needs, from medical diagnostics to educational contexts, improving usability and acceptance. Additionally, LLMs contribute to ethical considerations by surfacing biases and ethical implications, empowering stakeholders to mitigate risks and promote fairness. Through continuous training and iteration, LLMs evolve to provide more informative explanations, supporting informed decision-making and regulatory efforts for ethical AI usage. In essence, the impact of LLMs on XAI is profound, offering opportunities to enhance transparency, accountability, and user empowerment in AI systems across various domains. Overall, the impact of LLMs on the XAI field is multifaceted, offering opportunities to enhance transparency, trust, and accountability in AI systems across various domains while also addressing ethical considerations and empowering users. However, it's essential to continue research and development efforts to maximize the benefits of LLMs while mitigating potential risks and challenges.

8.6. Complexity and Interpretability Trade-off

One of the fundamental challenges in XAI lies in striking the right balance between providing understandable explanations and maintaining the fidelity of complex models. This dilemma arises because, often, the more accurate and complex a model is, the harder it becomes to explain its decisions in a simple and interpretable manner. This trade-off can limit the adoption and utility of XAI techniques, particularly in domains where interpretability is crucial, such as healthcare or finance. If the explanations provided are too simplistic, users might not trust them, while overly complex explanations may not be comprehensible to non-experts. To address this challenge, researchers are exploring various approaches, such as designing hybrid models that combine the transparency of simpler models with the accuracy of complex ones or developing post-hoc explanation methods that translate complex model behavior into more understandable forms.

8.7. Black Box Interpretability

Another significant challenge in XAI is presented by techniques that offer only black-box interpretations. In such cases, the explanation provided does not directly reveal how the model arrived at its decision. Instead, these methods may rely on surrogate models or feature importance scores that are not easily interpretable by humans. Black-box interpretations can erode trust in AI systems, as users may be skeptical of decisions they cannot understand or explain. Moreover, in regulated industries or applications where transparency is required, black-box interpretations may not meet legal or ethical standards. To address this challenge, efforts are underway to develop XAI methods that provide more transparent explanations, such as generating human-understandable rules or visualizations that elucidate the decision-making process of the model. Additionally, improving model transparency and documentation can enhance trust in black-box systems.

8.8. Robustness and Stability

Many XAI methods lack robustness and stability, meaning that small perturbations in the input data or changes in model parameters can lead to significantly different explanations. This inconsistency undermines the reliability of explanations and hinders their usefulness in decision-making processes. Unstable explanations can lead to confusion and mistrust among users, especially in high-stakes applications where decision confidence is paramount. Moreover, the inability to rely on consistent explanations can hinder the deployment of XAI systems in real-world scenarios. To improve robustness and stability, researchers are investigating methods to quantify the uncertainty of explanations and assess their sensitivity to input variations. Techniques such as sensitivity analysis and uncertainty estimation can help identify and mitigate sources of instability in XAI methods, enhancing their reliability and usability.

9. Conclusion

This paper presents a comprehensive survey of key publications that have significantly advanced the field of XAI. Our work delivers three primary contributions, each

emphasizing critical insights and findings from our research. Firstly, we introduce a novel taxonomy for XAI techniques, offering a structured framework that facilitates a deeper understanding and comparison of the latest advancements in XAI. This taxonomy not only organizes existing knowledge but also highlights relationships and differences among various methods, promoting a clearer perspective on the state-of-the-art in XAI. Secondly, we perform a detailed comparative analysis of current XAI methodologies. Using our proposed taxonomy, we evaluate these techniques to identify their respective strengths and weaknesses. This analysis provides valuable guidance for researchers and practitioners in selecting the most suitable XAI methods for their specific needs and applications. Finally, we identify and discuss several emerging challenges within the XAI landscape. These include the critical need to balance model accuracy with interpretability, the imperative to enhance the efficiency and scalability of XAI methods for handling large-scale datasets, and the urgent requirement to protect XAI systems from adversarial attacks. We also underscore the importance of integrating XAI capabilities into existing machine learning platforms, empowering developers to seamlessly incorporate explainability into their models. By addressing these challenges and offering practical guidelines, our survey aims to propel the advancement and widespread adoption of XAI. We envision our contributions facilitating more transparent, trustworthy, and effective AI systems across various domains and applications, ultimately fostering greater trust and understanding of AI technologies.

Acknowledgments

This research is supported in part by the Canadian DND Innovation for Defence Excellence and Security (W7714-217794/001/SV1), NSERC Discovery Grants (RGPIN-2018-03872), and Canada Research Chairs Program (950-230623). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Abu-Rmileh, A. (2019). *The Multiple faces of 'Feature importance' in XGBoost*. https://towardsdatascience.com/be-careful-when-interpreting-your-features-importance-in-xgboost-6e16132588e7.

Abusitta, A., Aïmeur, E., & Wahab, O. A. (2019). Generative adversarial networks for mitigating biases in machine learning systems. *arXiv preprint arXiv:1905.09972*, .

Abusitta, A., Li, M. Q., & Fung, B. C. (2021). Malware classification and composition analysis: A survey of recent developments. *Journal of Information Security and Applications*, *59*, 102828.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, *6*, 52138–52160.

Adebayo, J., Gilmer, J., Goodfellow, I., & Kim, B. (2018). Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307*, .

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, *54*, 95–122.

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 559–560).

Akula, A., Wang, S., & Zhu, S.-C. (2020). Cocox: Generating conceptual and counterfactual explanations via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 2594–2601). volume 34.

Alonso, J. M., Ramos-Soto, A., Castiello, C., & Mencar, C. (2018). Explainable ai beer style classifier. In *SICSA ReaLX*.

Amoukou, S. I., Brunel, N. J., & Salaün, T. (2021). The shapley value of coalition of variables provides better explanations. *arXiv preprint arXiv:2103.13342*, .

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*, e1424.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*, 1059–1086.

Arbatli, A. D., & Akin, H. L. (1997). Rule extraction from trained neural networks using genetic algorithms. *Nonlinear Analysis: Theory, Methods & Applications*, *30*, 1639–1648.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115.

Augasta, M. G., & Kathirvalavakumar, T. (2012). Reverse engineering the neural networks for rule extraction in classification problems. *Neural processing letters*, *35*, 131–150.

Aung, M. H., Lisboa, P. G., Etchells, T. A., Testa, A. C., Van Calster, B., Van Huffel, S., Valentin, L., & Timmerman, D. (2007). Comparing analytical decision support models through boolean rule extraction: A case study of ovarian tumour malignancy. In *International Symposium on Neural Networks* (pp. 1177–1186). Springer.

Auret, L., & Aldrich, C. (2012). Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering*, *35*, 27–42.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*, e0130140.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K.-R. (2010). How to explain individual classification decisions. *The Journal of Machine Learning Research*, *11*, 1803–1831.

Barakat, N., & Diederich, J. (2008). Eclectic rule-extraction from support vector machines. *International Journal of Computer and Information Engineering*, *2*, 1672–1675.

Barakat, N. H., & Bradley, A. P. (2007). Rule extraction from support vector machines: A sequential covering approach. *IEEE Transactions on Knowledge and Data Engineering*, *19*, 729–741.

Bastani, O., Kim, C., & Bastani, H. (2017). Interpretability via model extraction. *arXiv preprint arXiv:1706.09773*, .

Bega, D., Gramaglia, M., Banchs, A., Sciancalepore, V., & Costa-Pérez, X. (2019). A machine learning approach to 5g infrastructure market optimization. *IEEE Transactions on Mobile Computing*, *19*, 498–512.

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in big Data*, (p. 39).

Bonifazi, G., Cauteruccio, F., Corradini, E., Marchetti, M., Terracina, G., Ursino, D., & Virgili, L. (2024). A model-agnostic, network theory-based framework for supporting xai on classifiers. *Expert Systems with Applications*, *241*, 122588.

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source code for biology and medicine*, *3*, 1–8.

Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 258–262).

Camburu, O.-M. (2020). Explaining deep neural networks. *arXiv preprint arXiv:2010.01496*, .

Campagnolo, G. M., & Sharkey, E. (2021). Algorithmic encounters: an interactional approach to the ai accuracy vs interpretability trade-off, .

van Campen, T., Hamers, H., Husslage, B., & Lindelauf, R. (2018). A new approximation method for the shapley value applied to the wtc 9/11 terrorist attack. *Social Network Analysis and Mining*, *8*, 1–12.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721–1730).

Casalicchio, G., Molnar, C., & Bischl, B. (2018). Visualizing the feature importance for black box models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 655–670). Springer.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, .

Chaves, A. C., Vellasco, M. M., & Tanscheit, R. (2005). Fuzzy rule extraction from support vector machines. In *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)* (pp. 6–pp). IEEE.

Che, Z., Purushotham, S., Khemani, R., & Liu, Y. (2016). Interpretable deep models for icu outcome prediction. In *AMIA annual symposium proceedings* (p. 371). American Medical Informatics Association volume 2016.

Chen, D., Fraiberger, S. P., Moakler, R., & Provost, F. (2017). Enhancing transparency and control when drawing data-driven inferences about individuals. *Big data*, *5*, 197–212.

Chen, H., Lundberg, S., & Lee, S.-I. (2021). Explaining models by propagating shapley values of local components. In *Explainable AI in Healthcare and Medicine* (pp. 261–270). Springer.

Chen, Z., Li, J., & Wei, L. (2007). A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. *Artificial Intelligence in Medicine*, *41*, 161–175.

Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *arXiv preprint arXiv:1608.05745*, .

Cortez, P., & Embrechts, M. J. (2011). Opening black box data mining models using sensitivity analysis. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (pp. 341–348). IEEE.

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*, 1–17.

Craven, M. W. (1996). *Extracting comprehensible models from trained neural networks*. Technical Report University of Wisconsin-Madison Department of Computer Sciences.

Craven, M. W., & Shavlik, J. W. (1994). Using sampling and queries to extract rules from trained neural networks. In *Machine learning proceedings 1994* (pp. 37–45). Elsevier.

Dabkowski, P., & Gal, Y. (2017). Real time image saliency for black box classifiers. *arXiv preprint arXiv:1705.07857*, .

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*, .

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, .

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)* (pp. 598–617). IEEE.

Deng, H. (2019). Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics*, *7*, 277–287.

DND (2021). *Autonomous Systems for Defence and Security: Trust and Barriers to Adoption*. https://www.canada.ca/en/department-national-defence/programs/defence-ideas/current-opportunities/innovation-network-opportunities.html.

Dombrowski, A.-K., Alber, M., Anders, C. J., Ackermann, M., Müller, K.-R., & Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. *arXiv preprint arXiv:1906.07983*, .

Domingos, P. (1998). Knowledge discovery via multiple models. *Intelligent Data Analysis*, *2*, 187–202.

Dong, Y., Su, H., Zhu, J., & Zhang, B. (2017). Improving interpretability of deep neural networks with semantic information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4306–4314).

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. In *arXiv preprint arXiv:1702.08608*.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210–0215). IEEE.

Dosovitskiy, A., & Brox, T. (2016). Inverting visual representations with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4829–4837).

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, *1341*, 1.

Fatima, S. S., Wooldridge, M., & Jennings, N. R. (2008). A linear approximation method for the shapley value. *Artificial Intelligence*, *172*, 1673–1699.

Féraud, R., & Clérot, F. (2002). A methodology to explain neural network classification. *Neural networks*, *15*, 237–246.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, *20*, 1–81.

Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3429–3437).

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, (pp. 1189–1232).

Frosst, N., & Hinton, G. (2017). Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, .

Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*, *9*, 144352–144360.

Fu, L. (1994). Rule generation from neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, *24*, 1114–1124.

Fu, X., Ong, C., Keerthi, S., Hung, G. G., & Goh, L. (2004). Extracting the knowledge embedded in support vector machines. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)* (pp. 291–296). IEEE volume 1.

Gaonkar, B., Shinohara, R. T., Davatzikos, C., Initiative, A. D. N. et al. (2015). Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. *Medical image analysis*, *24*, 190–204.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*, 44–65.

Guidotti, R., Monreale, A., Pedreschi, D., & Giannotti, F. (2021). Principles of explainable artificial intelligence. In *Explainable AI Within the Digital Transformation and Cyber Physical Systems* (pp. 9–31). Springer.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*, 1–42.

Guidotti, R., & Ruggieri, S. (2019). On the stability of interpretable models. In *2019 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). Xai—explainable artificial intelligence. *Science Robotics*, *4*, eaay7120.

Haasdonk, B. (2005). Feature space interpretation of svms with indefinite kernels. *IEEE Transactions on pattern analysis and machine intelligence*, *27*, 482–492.

Hara, S., & Hayashi, K. (2016). Making tree ensembles interpretable. *arXiv preprint arXiv:1606.05390*, .

Henelius, A., Puolamäki, K., Boström, H., Asker, L., & Papapetrou, P. (2014). A peek into the black box: exploring classifiers by randomization. *Data mining and knowledge discovery*, *28*, 1503–1529.

Henelius, A., Puolamäki, K., & Ukkonen, A. (2017). Interpreting classifiers through attribute interactions in datasets. *arXiv preprint arXiv:1707.07576*, .

Heo, J., Joo, S., & Moon, T. (2019). Fooling neural network interpretations via adversarial model manipulation. *Advances in Neural Information Processing Systems*, *32*, 2925–2936.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, .

Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, *4*, 5947.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*, 1527–1554.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic medicine*, *73*, 538–540.

Hooker, G. (2004). Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 575–580).

Islam, S. R., Eberle, W., Ghafoor, S. K., & Ahmed, M. (2021). Explainable artificial intelligence approaches: A survey. *arXiv preprint arXiv:2101.09429*, .

Jaccard, J., & Jaccard, J. (2001). *Interaction effects in logistic regression*. 135. Sage.

Jakkula, V. (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University*, *37*.

Jakulin, A., Možina, M., Demšar, J., Bratko, I., & Zupan, B. (2005). Nomograms for visualizing support vector machines. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 108–117).

Janzing, D., Minorics, L., & Blöbaum, P. (2020). Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics* (pp. 2907–2916). PMLR.

Johansson, U., König, R., & Niklasson, L. (2004a). The truth is in there-rule extraction from opaque models using genetic programming. In *FLAIRS Conference* (pp. 658–663). Miami Beach, FL.

Johansson, U., & Niklasson, L. (2009). Evolving decision trees using oracle guides. In *2009 IEEE Symposium on Computational Intelligence and Data Mining* (pp. 238–244). IEEE.

Johansson, U., Niklasson, L., & König, R. (2004b). Accuracy vs. comprehensibility in data mining models. In *Proceedings of the seventh international conference on information fusion* (pp. 295–300). Citeseer volume 1.

Kanehira, A., & Harada, T. (2019). Learning to explain with complemental examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8603–8611).

Karpathy, A., Johnson, J., & Fei-Fei, L. (2015). Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, .

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)* (pp. 169–175). IEEE.

Kim, B., Rudin, C., & Shah, J. A. (2014). The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in neural information processing systems* (pp. 1952–1960).

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning* (pp. 2668–2677). PMLR.

Kindermans, P.-J., Schütt, K. T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., & Dähne, S. (2017). Learning how to explain neural networks: Patternnet and patternattribution. *arXiv preprint arXiv:1705.05598*, .

Knight, W. (2021). *An AI Pioneer Wants His Algorithms to Understand the 'Why'*. `https://www.wired.com/story/ai-pioneer-algorithms-understand-why/`.

Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *International Conference on Machine Learning* (pp. 1885–1894). PMLR.

Kök, I., Okay, F. Y., Muyanlı, Ö., & Özdemir, S. (2023). Explainable artificial intelligence (xai) for internet of things: a survey. *IEEE Internet of Things Journal*, .

Konig, R., Johansson, U., & Niklasson, L. (2008). G-rex: A versatile framework for evolutionary data mining. In *2008 IEEE International Conference on Data Mining Workshops* (pp. 971–974). IEEE.

Krakovna, V., & Doshi-Velez, F. (2016). Increasing the interpretability of recurrent neural networks using hidden markov models. *arXiv preprint arXiv:1606.05320*, .

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5686–5697).

Krishnan, R., Sivakumar, G., & Bhattacharya, P. (1999). Extracting decision trees from trained neural networks. *Pattern recognition*, *32*.

Krishnan, S., & Wu, E. (2017). Palm: Machine learning explanations for iterative debugging. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics* (pp. 1–6).

Kuppa, A., & Le-Khac, N.-A. (2020). Black box attacks on explainable artificial intelligence (xai) methods in cyber security. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., & Doshi-Velez, F. (2019). Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (pp. 59–67). volume 7.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2017). Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, .

Landecker, W., Thomure, M. D., Bettencourt, L. M., Mitchell, M., Kenyon, G. T., & Brumby, S. P. (2013). Interpreting individual classifications of hierarchical networks. In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (pp. 32–38). IEEE.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, *296*, 103473.

Laurent, H., & Rivest, R. L. (1976). Constructing optimal binary decision trees is np-complete. *Information processing letters*, *5*, 15–17.

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, .

Letham, B., Rudin, C., McCormick, T. H., Madigan, D. et al. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, *9*, 1350–1371.

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2015). Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*, .

Li, M. Q., Fung, B., & Abusitta, A. (2021a). On the effectiveness of interpretable feedforward neural network. *arXiv preprint arXiv:2111.02303*, .

Li, M. Q., Fung, B. C. M., Charland, P., & Ding, S. H. H. (2021b). I-MAD: Interpretable malware detector using Galaxy Transformers. *Computers & Security (COSE)*, *108*, 1–15.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*, 18.

Lipton, Z. C. (2016). The mythos of model interpretability. *Queue*, *16*, 31–57.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*, 31–57.

Lu, J., Lee, D., Kim, T. W., & Danks, D. (2019). Good explanation for algorithmic transparency. *Available at SSRN 3503603*, .

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, .

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, *2*, 56–67.

Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5188–5196).

Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications* volume 81. World scientific.

Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. *Mis Quarterly*, *38*, 73–100.

McDermid, J. A., Jia, Y., Porter, Z., & Habli, I. (2021). Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A*, *379*, 20200363.

Medsker, L. R., & Jain, L. (2001). Recurrent neural networks. *Design and Applications*, *5*, 64–67.

de Mello, F. L. (2020). A survey on machine learning adversarial attacks. *Journal of Information Security and Cryptography (Enigma)*, *7*, 1–7.

Messalas, A., Kanellopoulos, Y., & Makris, C. (2019). Model-agnostic interpretability with shapley values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1–7). IEEE.

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, *19*, 1236–1246.

Mishra, S., Sturm, B. L., & Dixon, S. (2017). Local interpretable model-agnostic explanations for music content analysis. In *ISMIR* (pp. 537–543).

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, *65*, 211–222.

Navia-Vázquez, A., & Parrado-Hernández, E. (2006). Support vector machine interpretation. *Neurocomputing*, *69*, 1754–1759.

Nefeslioglu, H., Sezer, E., Gokceoglu, C., Bozkir, A., & Duman, T. (2010). Assessment of landslide susceptibility by decision trees in the metropolitan area of istanbul, turkey. *Mathematical Problems in Engineering*, *2010*.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *arXiv preprint arXiv:1605.09304*, .

Núñez, H., Angulo, C., & Català, A. (2002). Support vector machines with symbolic interpretation. In *VII Brazilian Symposium on Neural Networks, 2002. SBRN 2002. Proceedings.* (pp. 142–147). IEEE.

Nunez, H., Angulo, C., & Catala, A. (2006). Rule-based learning systems for support vector machines. *Neural Processing Letters*, *24*, 1–18.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, .

Palczewska, A., Palczewski, J., Robinson, R. M., & Neagu, D. (2014). Interpreting random forest classification models using a feature contribution method. In *Integration of reusable systems* (pp. 193–218). Springer.

Papernot, N., & McDaniel, P. (2018). Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, .

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*, 3–14.

Peng, X., Li, Y., Tsang, I. W., Zhu, H., Lv, J., & Zhou, J. T. (2022). Xai beyond classification: Interpretable neural clustering. *Journal of Machine Learning Research*, *23*, 1–28.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*, 81–106.

Quinlan, J. R. (1987). Generating production rules from decision trees. In *International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 304–307). volume 87.

Rabiul Islam, S., Eberle, W., Khaled Ghafoor, S., & Ahmed, M. (2021). Explainable artificial intelligence approaches: A survey. *arXiv e-prints*, (pp. arXiv–2101).

Rai, A. (2020). Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*, 137–141.

Rajani, N. F., & Mooney, R. (2018a). Stacking with auxiliary features for visual question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2217–2226).

Rajani, N. F., & Mooney, R. J. (2018b). Ensembling visual explanations. In *Explainable and Interpretable Models in Computer Vision and Machine Learning* (pp. 155–172). Springer.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Nothing else matters: model-agnostic explanations by identifying prediction invariance. *arXiv preprint arXiv:1611.05817*, .

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 32.

Rieger, L., & Hansen, L. K. (2020). A simple defense against adversarial attacks on heatmap explanations. *arXiv preprint arXiv:2007.06381*, .

Rieger, L., Singh, C., Murdoch, W., & Yu, B. (2020). Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning* (pp. 8116–8126). PMLR.

Robnik-Šikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, *20*, 589–600.

Rosenbaum, L., Hinselmann, G., Jahn, A., & Zell, A. (2011). Interpreting linear support vector machine models with heat map molecule coloring. *Journal of Cheminformatics*, *3*, 1–12.

Roth, A. E. (1988). *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.

Rovnyak, S., Kretsinger, S., Thorp, J., & Brown, D. (1994). Decision trees for real-time transient stability prediction. *IEEE Transactions on Power Systems*, *9*, 1417–1426.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215.

Saeed, W., & Omlin, C. (2023). Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, *263*, 110273.

Samek, W., & Müller, K.-R. (2019). Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning* (pp. 5–22). Springer.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, .

Saranya, A., & Subhashini, R. (2023). A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision analytics journal*, (p. 100230).

Schmitz, G. P., Aldrich, C., & Gouws, F. S. (1999). Ann-dt: an algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks*, *10*, 1392–1401.

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, (pp. 1–59).

Setiono, R., & Leow, W. K. (2000). Fernn: An algorithm for fast extraction of rules from neural networks. *Applied Intelligence*, *12*, 15–25.

Seungjun, K. (2022). *Explainable AI (XAI) Methods Part 5— Global Surrogate Models*. https://towardsdatascience.com/explainable-ai-xai-methods-part-5-global-surrogate-models-9c228d27e13a.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning* (pp. 3145–3153). PMLR.

Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, .

Slack, D., Friedler, S. A., Scheidegger, C., & Roy, C. D. (2019). Assessing the local interpretability of machine learning models. *arXiv preprint arXiv:1902.03501*, .

Sollich, P. (1999). Probabilistic methods for support vector machines. In *NIPS* (pp. 349–355). volume 12.

Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine learning*, *46*, 21–52.

Song, C., Cheng, H.-P., Yang, H., Li, S., Wu, C., Wu, Q., Chen, Y., & Li, H. (2018). Mat: A multi-strength adversarial training method to mitigate adversarial attacks. In *2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* (pp. 476–481). IEEE.

Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social studies of science*, *48*, 25–56.

Strumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, *11*, 1–18.

Su, G., Wei, D., Varshney, K. R., & Malioutov, D. M. (2015). Interpretable two-level boolean rule learning for classification. *arXiv preprint arXiv:1511.07361*, .

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning* (pp. 3319–3328). PMLR.

Taha, I. A., & Ghosh, J. (1999). Symbolic interpretation of artificial neural networks. *IEEE Transactions on knowledge and data engineering*, *11*, 448–463.

Tallón-Ballesteros, A., & Chen, C. (2020). Explainable ai: Using shapley value to explain complex anomaly detection ml-based systems. *Machine learning and artificial intelligence*, *332*, 152.

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 303–310).

Tan, S., Sim, K. C., & Gales, M. (2015). Improving the interpretability of deep neural networks with stimulated learning. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 617–623). IEEE.

Tan, S., Soloviev, M., Hooker, G., & Wells, M. T. (2020). Tree space prototypes: Another look at making tree ensembles interpretable. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference* (pp. 23–34).

Thiagarajan, J. J., Kailkhura, B., Sattigeri, P., & Ramamurthy, K. N. (2016). Treeview: Peeking into deep neural networks via feature-space partitioning. *arXiv preprint arXiv:1611.07429*, .

Thrun, S. (1995). Extracting rules from artificial neural networks with distributed representations. *Advances in neural information processing systems*, (pp. 505–512).

Tolomei, G., Silvestri, F., Haines, A., & Lalmas, M. (2017). Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd*

ACM SIGKDD international conference on knowledge discovery and data mining (pp. 465–474).

Towell, G. G., & Shavlik, J. W. (1993). Extracting refined rules from knowledge-based neural networks. *Machine learning*, *13*, 71–101.

Trunk, A., Birkel, H., & Hartmann, E. (2020). On the current state of combining human and artificial intelligence for strategic organizational decision making. *Business Research*, *13*, 875–919.

Tsukimoto, H. (2000). Extracting rules from trained neural networks. *IEEE Transactions on Neural networks*, *11*, 377–389.

Usman, M., Jan, M. A., He, X., & Chen, J. (2019). A survey on representation learning efforts in cybersecurity domain. *ACM Computing Surveys (CSUR)*, *52*, 1–28.

Üstün, B., Melssen, W., & Buydens, L. (2007). Visualisation and interpretation of support vector regression models. *Analytica chimica acta*, *595*, 299–309.

Utgoff, P. E. (1989). Incremental induction of decision trees. *Machine learning*, *4*, 161–186.

Vaishak, B., & Ioannis, P. (2021). *Principles and Practice of Explainable Machine Learning*. `https://www.frontiersin.org/articles/10.3389/fdata.2021.688969/full`.

Vellido, A., Martín-Guerrero, J. D., & Lisboa, P. J. (2012). Making machine learning models interpretable. In *ESANN* (pp. 163–172). Citeseer volume 12.

Vilone, G., & Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, *3*, 615–661.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, *11*.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404.

Welling, S. H., Refsgaard, H. H., Brockhoff, P. B., & Clemmensen, L. H. (2016). Forest floor visualizations of random forests. *arXiv preprint arXiv:1605.09196*, .

Winter, E. (2002). The shapley value. *Handbook of game theory with economic applications*, *3*, 2025–2054.

Wisdom, S., Powers, T., Pitton, J., & Atlas, L. (2016). Interpretable recurrent neural networks using sequential sparse recovery. *arXiv preprint arXiv:1611.07252*, .

Wu, M., Hughes, M., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (2018). Beyond sparsity: Tree regularization of deep models for interpretability. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 32.

Xu, K., Park, D. H., Yi, C., & Sutton, C. (2018). Interpreting deep classifier by visual distillation of dark knowledge. *arXiv preprint arXiv:1803.04042*, .

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, .

Yu, F., Wei, C., Deng, P., Peng, T., & Hu, X. (2021). Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles. *Science Advances*, *7*, eabf4130.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.

Zeiler, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010). Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition* (pp. 2528–2535). IEEE.

Zeiler, M. D., Taylor, G. W., & Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision* (pp. 2018–2025). IEEE.

Zhang, P. (2016). An interval mean–average absolute deviation model for multiperiod portfolio selection with risk control and cardinality constraints. *Soft Computing*, *20*, 1203–1212.

Zhang, Q., Wu, Y. N., & Zhu, S.-C. (2018). Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8827–8836).

Zhang, Q., Yang, Y., Ma, H., & Wu, Y. N. (2019). Interpreting cnns via decision trees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6261–6270).

Zhang, Y., Su, H., Jia, T., & Chu, J. (2005). Rule extraction from trained support vector machines. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 61–70). Springer.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*, 1–38.

Zhou, Y., & Hooker, G. (2016). Interpreting models via single tree approximation. *arXiv preprint arXiv:1610.09036*, .

Zhou, Z.-H., Jiang, Y., & Chen, S.-F. (2003). Extracting symbolic rules from trained neural network ensembles. *Ai Communications*, *16*, 3–15.

Zilke, J. R., Mencía, E. L., & Janssen, F. (2016). Deepred–rule extraction from deep neural networks. In *International Conference on Discovery Science* (pp. 457–473). Springer.