Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

A REVIEW OF EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI): TOWARDS FINANCE

[1*]Visakh Chandran Melveetil, [2]Dr Saumendra Mohanty
[1]GradXs Researcher for PhD from IIC University of Technology, Cambodia*
[2]Professor & Area Chair (Business Analytics & AI) at Lloyd Business School, Research Guide IIC University of Technology in Cambodia and Liutebm University, School of International Programs, Lusaka, Zambia.
[1]Corresponding Author Email: mvisakhc@gmail.com*

ABSTRACT

Recently, Artificial Intelligence (AI) has evolved to become a critical capability of an Information System (IS). The development and application of AI have surged recently due to advances in computing and the exponential growth of system-captured data. As AI has found its way into various key domains like banking, law, healthcare, etc., it is essential to ensure that the decision-making process remains transparent. As AI technologies advance rapidly, we face exciting opportunities and new challenges. While AI research and development is in an astonishingly fast phase, concerns about potential threats posed by this are also increasing. Experts focus on developing effective governance strategies for AI systems to address these concerns. A critical aspect of this effort is ensuring that these AI systems are transparent and understandable to end users. Therefore, AI governance and explainability are crucial to establishing a secure and reliable AI-powered future. This study uses a structured literature review of over 70 research papers from 2018 to 2024 to provide an overview of significant outlets, academic discourse development, and key concepts and methodologies. With the ever-changing dynamics of AI due to the inclusion of modern AI technologies like Generative AI, Large Language models, etc, the challenges of explainability remain a genuine concern from the perspective of governance and regulatory compliance. The survey aims to fill this research gap by providing a comprehensive and up-to-date overview of relevant XAI approaches in Finance.

Keywords - Systematic literature review, Information system, Explainable AI, Artificial Intelligence, XAI, Finance, machine learning, deep learning

1. INTRODUCTION

Advancements in modern computing and exponential data growth have accelerated the development and adoption of AI in recent years. AI techniques are essential for society's overall progress, as they can execute mundane and complex tasks at high levels of accuracy and efficiency (Dwivedi et al., 2021). Maintaining transparency in AI decision-making is crucial as it becomes more prevalent in finance, law, healthcare, etc (Yigitcanlar et al., 2021).

The first generation of AI systems was easy to understand but had limitations as they needed help to solve complex real-world problems. However, with the advent of Deep Learning (DL) models, there has been a significant improvement in addressing such challenges. DL models can handle non-linear problems efficiently, making them more effective in solving day-to-day problems. However, they are often considered complex 'black box' models due to their intricacies in terms of layers and parameters (which constitute a model). Opposite to the black box is 'Transparency'. It involves comprehending the underlying mechanism through which an AI model operates (Gill, et al., 2022).

1.1. Relevance of AI Transparency

As we rely more on AI to make essential predictions in crucial environments, there is a greater need to improve its overall transparency. The primary risks associated with the black 'box model' (decision-making) are the lack of justification, challenges in legitimacy, or the ability to explain the decision-making process comprehensively (Li et al., 2022 & Rai, 2020). In specific fields, like precision healthcare, experts need access to more specific information from the model. This goes beyond just a basic binary prediction and can significantly assist in the diagnosis process. Similar challenges exist in other fields, such as transportation, banking, and finance (Zednik, 2021; Antoniadi et al., 2021).

Transparency in AI involves providing stakeholders access to information, decisions, and assumptions, thereby improving their understanding of AI processes or systems (Vinuesa et al., 2020). Due to the spread of AI in daily life, developing regulations concerning transparency is imperative. Thus, transparency will aid in promoting accountability and trust among the stakeholders of the AI system. To achieve higher levels of transparency, there is a need for tools and methods to enhance the interpretability and understanding of AI systems, and this is where Explainable AI (XAI) focuses (Yu et al., 2023).

1.2. Explainable AI (XAI) and AI Transparency

XAI is a collection of capabilities that an AI system uses to provide clear explanations of its decisions and actions to improve transparency, trust, and accountability. The importance of XAI has increased as AI gets adopted in healthcare, autonomous driving, and other sectors (Hu et al., 2021). XAI methods aim to provide comprehensible and reliable explanations for the outcomes generated by AI methods. The European General Data Protection Regulation (GDPR) and other regulations are pushing for more research in XAI. These regulations emphasize the need for ethical considerations, justifications, trust, and bias exploration in developing trustworthy XAI solutions (Payrovnaziri et al., 2020; Butz et al., 2022).

Various factors influence the importance of XAI, and they vary depending on the individuals involved.

- End-users require confidence in decisions through transparent processes and feedback.
- AI developers should clearly understand the boundaries of current models to validate and enhance their future versions.
- Product managers need access to clear decision-making explanations and should refine them when implementing algorithms in real-life situations.

(Zablocki et al., 2022; Dazeley et al., 2023)

1.3. XAI and Customer Personalization

When developing an XAI system, it's crucial to consider individuals from diverse cultural backgrounds and situations. This is because everyone interprets information differently based on their life experiences. The 'Contextual Utility Theory' can be used to tailor the model's explainability (Kou and Gui, 2020). For example, in healthcare, it's essential to explain AI-generated diagnoses or treatments in a way that's appropriate for each patient's medical knowledge level. Using simple language and visual aids can help patients with limited medical knowledge understand the reasoning behind recommendations, leading to greater engagement and comprehension. Healthcare professionals with a more in-depth understanding of medical concepts can receive more detailed explanations that delve into intricate technicalities and offer insights into how the model makes decisions (Ploug and Holm, 2020; Amann et al., 2020). Therefore, 'personalization' is critical when designing an XAI system, focusing on customizing outcomes for various user groups (Saeed and Omlin, 2023).

Also, XAI systems must be adaptable as AI models and data change over time. Proper data management ensures XAI systems' long-term usefulness and sustainability. XAI methods must be flexible enough to adapt to changes in data characteristics as data and its distributions evolve over time (Lopes et al., 2022; Sanneman et al., 2022).

1.4. Balanced or Unbalanced Approach in XAI

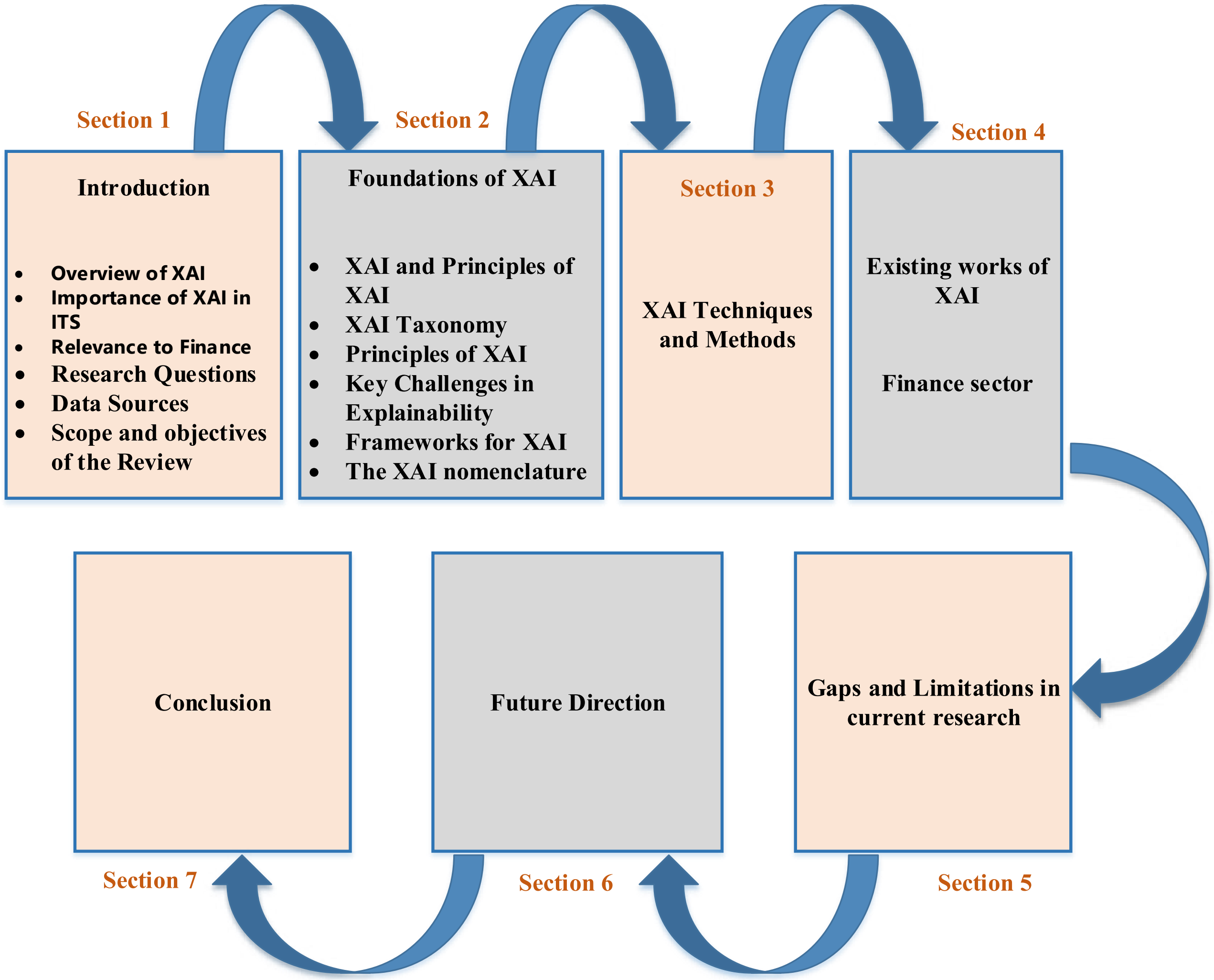
With the increasing popularity of XAI, researchers in academic and industrial circles are exploring various methodologies to balance interpretability and predictive accuracy. As research continues, it is becoming increasingly clear that while interpretable models are necessary, they may not fully capture the intricate data relationships present within a complex AI model. Conversely, while complex models tend to excel in accuracy, interpreting their behaviour can be challenging. As a result, obtaining the optimal balance between these factors is essential and greatly influenced by the needs of its application area. In industries like finance, transparency and interpretability are of utmost importance to comply with regulatory standards and address ethical concerns. In certain domains, such as medical imaging or signal identification, achieving high levels of precision is paramount. In such contexts, predictive accuracy takes priority over interpretability. This is because interpretability, while still important, becomes a secondary concern when the primary objective is to develop a model that accurately predicts outcomes (Weber et al., 2023; Páez 2019).

XAI is becoming increasingly vital in finance as it brings transparency to previously opaque areas such as credit scoring, risk assessment models, algorithmic trading strategies, etc. It thus addresses the demand for accountability and regulatory compliance by providing clear insights into the decision-making mechanisms of financial systems. Through XAI, stakeholders can better understand complex financial models, leading to greater trust, informed decision-making, risk reduction, and fortification of financial systems (Hoepner et al., 2021; Mill et al., 2023).

Integrating XAI ensures transparency and fairness in decision-making processes, particularly in public and non-profit organizations. It fosters accountability and trust among stakeholders by providing insight into decision-making. Furthermore, XAI is a vital component of legal and regulatory systems, as it enhances the comprehensibility of decisions, thereby promoting transparency and ethical governance. Ultimately, this reinforces compliance and the ethical underpinnings of governance practices (Langer et al., 2021; Larsson 2019).

2. SCOPE & OBJECTIVES OF STUDY

Our primary goal in this study is to thoroughly review XAI techniques, emphasizing their use in finance and governance domains. To do this, we will investigate and evaluate the various approaches and strategies used in XAI while considering governance and finance. We use well-defined taxonomies to organize our research, as shown in Figure 1.



Section 1

Introduction

- Overview of XAI
- Importance of XAI in ITS
- Relevance to Finance
- Research Questions
- Data Sources
- Scope and objectives of the Review

Section 2

Foundations of XAI

- XAI and Principles of XAI
- XAI Taxonomy
- Principles of XAI
- Key Challenges in Explainability
- Frameworks for XAI
- The XAI nomenclature

Section 3

XAI Techniques and Methods

Section 4

Existing works of XAI

Finance sector

Section 5

Gaps and Limitations in current research

Section 6

Future Direction

Section 7

Conclusion

Figure 1. Taxonomy of the Survey Paper

Furthermore, we focus on situations where these areas overlap, highlighting how XAI can effectively tackle obstacles in finance. Additionally, we thoroughly discuss the drawbacks of current methods, recognizing the need for improvements that adapt to the unique requirements of finance and governance applications. Our survey concludes by highlighting future directions in XAI research that are specifically relevant to finance and governance domains. We identify potential pathways for progress and innovation, supporting the continuous development of XAI in these critical industries.

To summarize, our survey contributions are outlined below:

- We conduct an extensive review of existing literature on XAI within the information systems of finance, providing a nuanced understanding of the applications, challenges, and advancements specific to these critical domains.
- Our review synthesizes key findings and methodologies from the literature, focusing on how XAI is employed in information systems to enhance transparency, accountability, and decision-making processes within the intricate landscapes of finance.
- We identify emerging trends within XAI that are particularly relevant to the dynamic environments of finance and governance information systems, offering insights into the evolving technologies and methodologies shaping the future of these domains.
- To enhance clarity and understanding, we develop a structured taxonomy or classification system tailored explicitly to XAI techniques applicable to information systems in finance. This taxonomy categorizes and organizes the diverse methodologies employed in these domains.

2.1. Data Sources

In our comprehensive survey of XAI in finance, we examined a total of 110+ papers. Leading the scenario are journals like "Frontiers in Artificial Intelligence,""Journal of Computer Science and Technology," and "Journal of Risk and Financial Management," each contributing significantly with eight papers. This analysis highlights the interdisciplinary nature of XAI applications in finance, emphasizing the key role played by these journals in disseminating impactful research. Additionally, a diverse distribution across various journals, with 71 papers from others, underscores the widespread influence of XAI research in shaping the financial realm. Figure 2 shows the research papers collected from various journals.

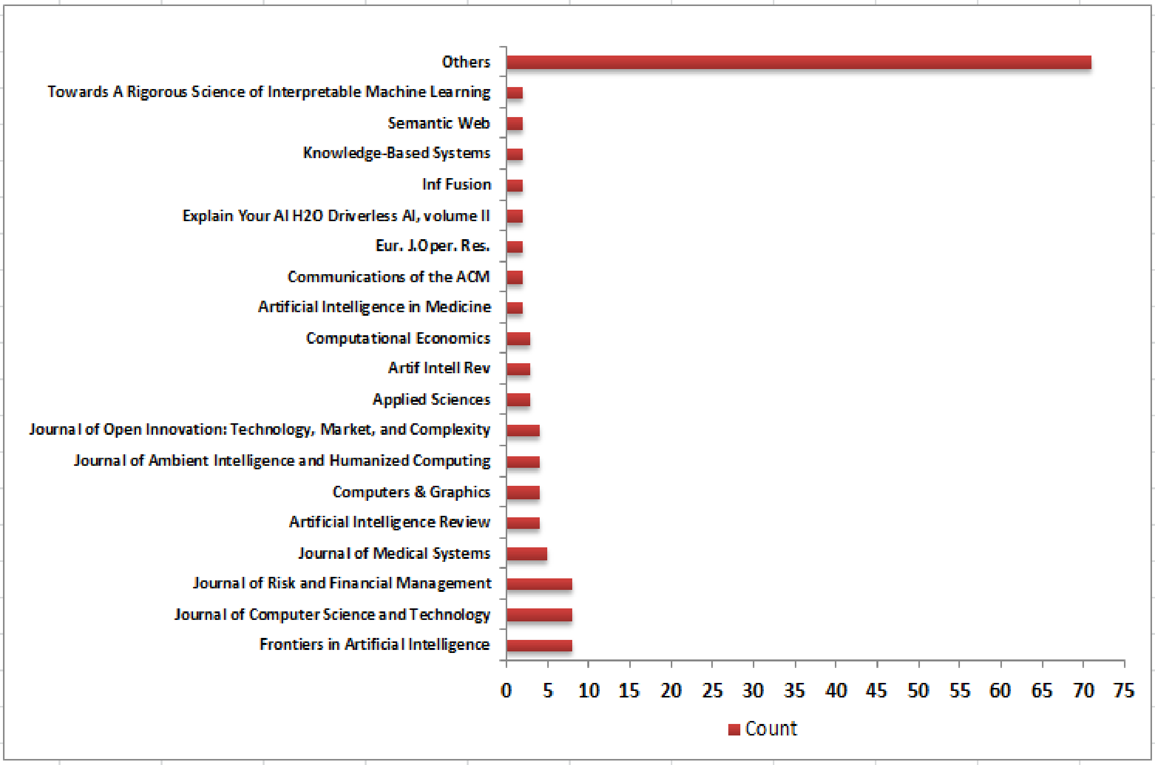


Figure 2. Research papers collected from various journals

Furthermore, we focus on situations where these areas overlap, highlighting how XAI can effectively tackle obstacles in finance and governance. Additionally, we thoroughly discuss the drawbacks of current methods, recognizing the need for improvements that adapt to the unique requirements of finance and governance applications. Our survey concludes by highlighting future directions in XAI research that are specifically relevant to finance and governance domains.

2.2. Structure of the paper

The organization of the paper is as follows: The introductory Section 1 provides an overview of XAI, outlines research questions, identifies data sources, and establishes the scope and objectives. Section 2 delves into the foundations of XAI, discussing principles, taxonomies, challenges, frameworks, and nomenclature. Section 3 explores XAI techniques and methods, while Section 4 reviews existing works in the finance sector. Section 5 critically evaluates gaps in current research, and Section 6 proposes future directions for XAI. The conclusion in Section 7 summarizes key insights and potential avenues for future exploration, making the paper a valuable resource for researchers, practitioners, and policymakers in the field of XAI and ITS.

3. LITERATURE REVIEW: EVALUATING THE THEORETICAL FRAMEWORKS

3.1 Explainable AI (XAI)

The Royal Society (2019) has emphasized the importance of understanding the functioning of AI decision-making systems as they become increasingly integrated into various real-world operations. The significance of interpretability in AI systems cannot be overstated, as it helps prevent bias, assures users of the system's performance, fulfills policy and regulatory standards, and aids developers in comprehending the system's behavior and identifying its vulnerabilities. However, Miller (2019) has noted that AI models face significant challenges regarding transparency and interpretability. Improving the explainability of AI systems is directly linked to enhanced user confidence in these systems, as users are more likely to trust AI-generated decisions or recommendations when they understand the influencing factors. Similarly, Zopounidis (1999) has highlighted that the clarity of decision interpretations often outweighs their complexity. Complex methods are less frequently utilized in practice, as their outputs can be difficult for financial decision-makers to comprehend.

3.2 XAI Taxonomy

Following are the core taxonomies that are used extensively within XAI.

1. AI Transparency: It is a term used to describe how easy it is for people who interact with AI systems or are impacted by them to understand how they work. It focuses on comprehending how the AI makes decisions or results.
2. AI Fairness: It means ensuring that AI systems treat all people and groups equally and without bias.
3. AI Trust or Trustworthy AI: It refers to people's confidence in AI systems. It's about believing these systems will work as expected, make fair and accurate decisions, and not cause harm.
4. AI Usability: It focuses on the effectiveness of an AI system in providing users with a reliable and safe environment to accomplish their tasks.
5. AI Reliability: The main emphasis is on the AI system's capacity to carry out its intended functions consistently and accurately, even in different conditions, without any unexpected failures or errors. This encompasses the system's capacity to deliver dependable and consistent results, thereby instilling user confidence.
6. Causality: The focus here is to understand the cause and effect of one or many inputs on each other and their effect on the output. Understanding the causality of patterns learned by AI models is essential for uncovering new insights or understandings. (Appelganc et al., 2022; Bajwa et al., 2021; Holzinger et al., 2019;)

To understand black-box AI models can be difficult due to their lack of interpretability. Recognizing the interpretability of a DL model allows us to find its inherent characteristics. This pertains to understanding the decision-making process of AI models. Intrinsic methods for modeling involve AI systems that can provide human-readable explanations of how a model function internally (Carvalho et al., 2019).

3.3. Principles of XAI

In 2020, the National Institute of Standards and Technology (NIST) introduced four critical principles for XAI systems, it includes:

- Explanation Principle: For an AI system to be transparent, it must provide evidence or reasoning related to its outcome.
- Meaningful: The AI system's explanations must be meaningful and understandable to the end user.
- Explanation Accuracy: Information and explanations provided by the AI system must accurately reflect the system's process for arriving at its decisions.
- Knowledge Limits: The limitations of AI systems should be communicated to the users.

(Guidotti, et al., 2021). Figure 3 provides a visual summary of XAI principles.

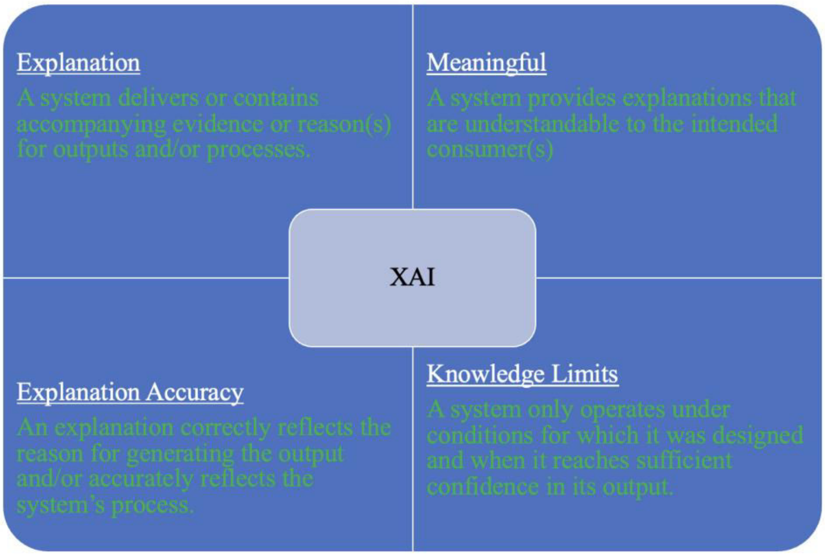

Explanation
A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.
Meaningful
A system provides explanations that are understandable to the intended consumer(s)
XAI
Explanation Accuracy
An explanation correctly reflects the reason for generating the output and/or accurately reflects the system's process.
Knowledge Limits
A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output.

Figure 3. NIST XAI Principles of XAI (Phillips, et al., 2020)

The methods that are involved in generating explanations for an AI model could vary differently depending on the following characteristics:

- Model-specific and Model-agnostic: Model-specific methods are built for a single model, whereas model-agnostic methods are aimed to improve the interpretability of 'any' ML models. Model-agnostic methods cannot access weights and structural information (Bonifazi et al., 2024).
- Regarding model interpretability, there are two categories of approaches: intrinsic and extrinsic. If a model can be easily comprehended without further analysis, it is deemed to have intrinsic interpretability. Conversely, if understanding the model necessitates post-training analysis methods, then it falls under the extrinsic category. Decision trees are a prime example of models with intrinsic interpretability, as they offer simple and transparent models. However, models that require additional analysis to achieve understandability are deemed extrinsic (Zolanvari, et al., 2021).
- While evaluating a system, choosing a Local or Global approach depends on the expected outcome. The global approach focuses on the tools to evaluate the whole of the model, whereas the local approach focuses on a specific instance – How a specific prediction is arrived at by the model (Setzu et al., 2021; Dennehy et al., 2023).

3.4. Key challenges in explainability

The following are the critical challenges with explainability:

- Transparency and Interpretability Constraints: AI algorithms in finance are often considered as 'black boxes', thereby making it difficult for stakeholders to understand and trust the outcomes (Dennehy et al., 2023; Munz et al., 2023)
- Assessment of Trustworthiness: The reliability and trustworthiness of AI model outputs are challenging to evaluate due to their inherent opacity. This is especially problematic in the

financial sector, where decisions carry significant weight and require high confidence (Dennehy et al., 2023).

- Addressing Bias and Data Limitations: XAI initiatives must address bias in AI models, data deficiencies (data quality and quantity), and expertise for effective integration in the business process. Organizations need robust data governance practices that identify and mitigate biases. Steps include evaluating data quality, identifying biases, and ensuring fairness throughout the AI pipeline via feature selection, monitoring, and testing (Mohammed & Shehu, 2023).
- Behavioural Bias Management in Financial Planning: XAI in finance faces the challenge of mitigating behavioural biases prevalent among financial planners, which can be worsened by AI technologies (Hasan, et al., 2022).
- Requirement for Human-Intelligible Explanations: There is a critical demand for AI and machine learning techniques to provide outputs with justifications that are understandable to human users. This requirement is particularly salient in credit risk assessments, Fraud detection, etc. (Ehsan & Riedi (2020)).
- Mitigating Unfairness in AI Systems: The lack of transparency in AI systems can create unfairness and discrimination. XAI aims to fix these issues by addressing and rectifying potential biases in opaque AI processes (Alikhademi et al., 2021).
- Speed of innovation in AI: The field of AI is advancing much faster than XAI, creating new challenges for XAI (Tang et al., 2020).

4. REVIEWING XAI TECHNIQUES AND METHODS

This review aims to thoroughly analyze XAI techniques based on existing research and explore their advantages, limitations, and challenges. Our research draws primarily on three extensive studies and surveys (Molnar (2021); Islam et al., (2021); Velez and Kim (2017)) with a particular emphasis on the survey conducted by Xiao Li et al. (2020), which we use to classify the various techniques. Two primary categories for XAI-based methods revealed in the survey are:

- Data-driven approaches generate insights and explanations from AI model training data. Complex, 'black box' models like Neural Networks employ this method.
- Conversely, knowledge-driven methods aim to incorporate domain knowledge, rules, and reasoning into AI systems. They are typically transparent and interpretable by design.

After training, black-box models are explained by post-hoc methods, including sending parameters to humans for interpretation.

Explainability approaches are categorized based on findings from multiple investigations. Based on the existing research, explained three categories of explainable systems (Velez and Kim (2017); Hall and Gill (2019); Islam et al., (2021); Jalayer et al., (2020)):

- Intrinsically Interpretable Techniques: Refers to the models and methods that are naturally easy to understand and explain.
- Model-agnostic techniques are methods used to interpret and understand a model regardless of its construction or associated complexity.
- Example-Based Methods: Methods that can explain the behavior of AI models by providing specific instances or examples.

(Velez and Kim (2017); Hall and Gill (2019); Islam et al (2021); Jalayer et al., (2020))

The aforementioned categories will be included in the knowledge-driven and data-driven approaches. Figure 4 displays the taxonomy of various XAI techniques.

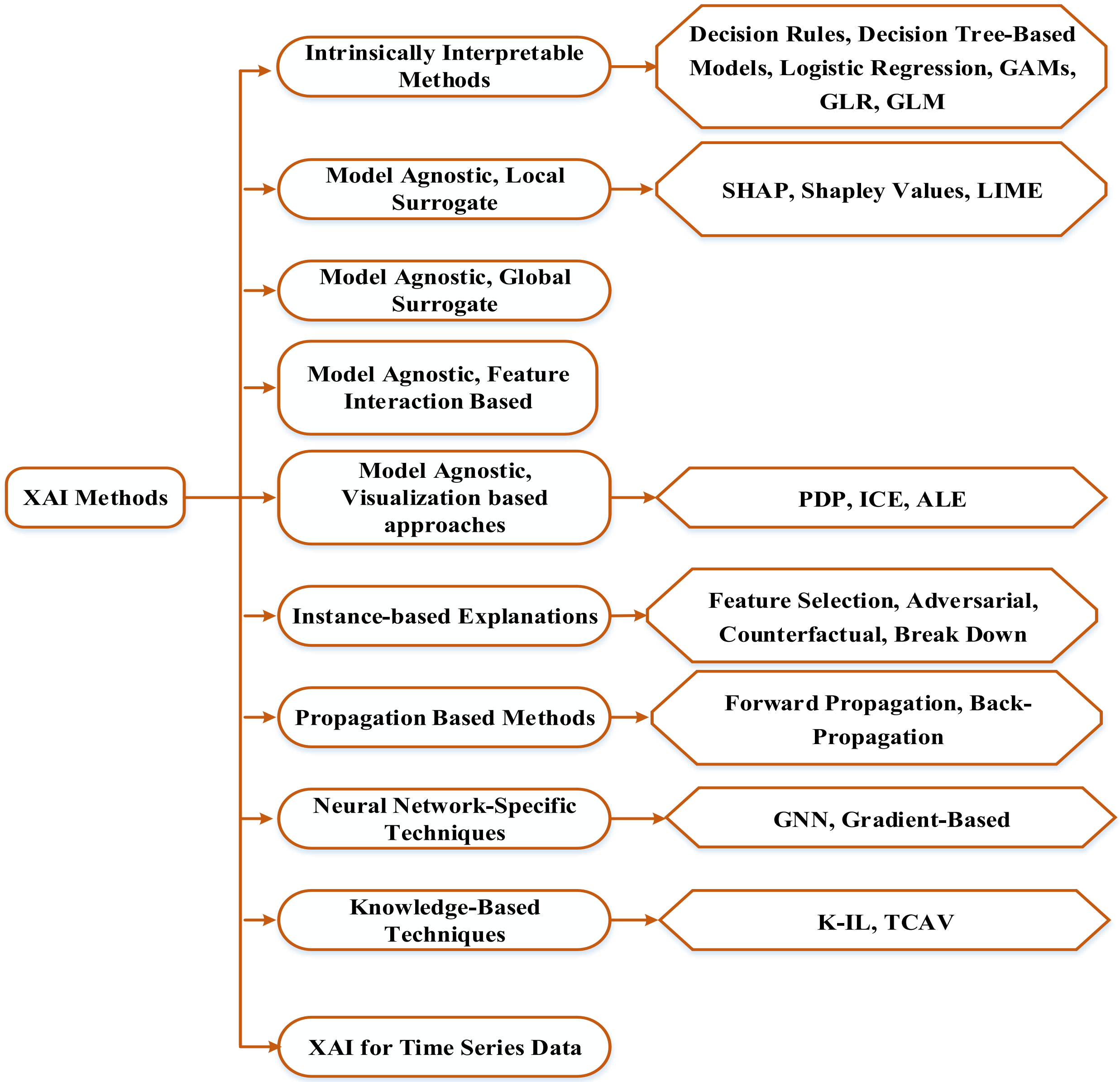

XAI Methods
Intrinsically Interpretable Methods
Decision Rules, Decision Tree-Based Models, Logistic Regression, GAMs, GLR, GLM
Model Agnostic, Local Surrogate
SHAP, Shapley Values, LIME
Model Agnostic, Global Surrogate
Model Agnostic, Feature Interaction Based
Model Agnostic, Visualization based approaches
PDP, ICE, ALE
Instance-based Explanations
Feature Selection, Adversarial, Counterfactual, Break Down
Propagation Based Methods
Forward Propagation, Back-Propagation
Neural Network-Specific Techniques
GNN, Gradient-Based
Knowledge-Based Techniques
K-IL, TCAV
XAI for Time Series Data

Figure 4. Taxonomy of XAI methods

4.1. Intrinsically Interpretable Methods

In the realm of XAI, the intrinsic interpretable models are crucial due to their transparency and simplicity. Linear regression is a prime example in this category because it can easily explain regression coefficients and linear relationships between variables. These models are widely applicable in medicine, sociology, sciences, finance, etc (Joyce et al., 2023).

Decision Rules follow a straightforward if-else statement that can be clearly articulated onto a decision tree for better clarity. Decision Tree-Based Models are machine learning algorithms that utilize decision trees for classification and regression tasks. These models segment the data by applying predetermined thresholds on the features, which makes them particularly effective in predicting interactive features that could be handled better by linear or logistic regressions. They are straightforward to interpret when the tree structure is simple enough. However, a dense tree can be challenging as minor changes in input can result in significant variations in the output, indicating a lack of smoothness or increased sensitivity to variations (Joyce et al., 2023).

Logistic regression is a classification technique that builds on linear regression. Unlike linear regression, the output of logistic regression is bounded between [0,1], which makes it suitable for predicting class probabilities (Biecek et al., 2021).

GAMs (Generalized Additive Models)are valuable for assessing linear model assumptions, particularly when analyzing output variable ‘y’ and feature ‘f ‘under the Gaussian Distribution without any feature interaction. When linear models are extended, they become more complex and less understandable due to increased interactions (Obster et al., 2023).

Generalized linear models (GLMs) extend traditional linear regression and can handle non-normal response variables. They have two main advantages over linear models: they maintain the weighted sum of features and can use non-Gaussian distributions (Yang et al., 2021).

It is worth noting that interpretable models can aid us in solving intricate problems, although they may not necessarily enhance our understanding of machine learning techniques. As we encounter more intricate problems with closely linked features, we must explore alternative methods for comprehending our systems. Within this domain, our attention will be directed toward techniques that employ existing data to offer interpretations of our models. These techniques can generally be categorized into three groups: visualization-based, model-agnostic, and example-based (Lundberg et al., 2022).

4.2. Model Agnostic, Local Surrogate

The interpretability technique outlined herein applies to any AI or ML model, regardless of its nature (model agnostic). Its primary objective is to clarify individual predictions (local) by constructing a more understandable (simpler) surrogate model that mimics the intricate model. These surrogate models are transparent and tailored to simulate the predictions issued by the black-box models (Neves et al., 2021).

SHAP (SHapley Additive exPlanations) is an AI methodology that employs game theory to explain model predictions. Each attribute in the predictive model is considered as a participant in a cooperative game, and Shapley values are used to determine its reasonable contribution to the final prediction. It ensures impartial assessments but may pose computational challenges and misinterpretations (Neves et al., 2021).

LIME, or Local Interpretable Model-agnostic Explanations, is a highly effective method for interpreting AI models, first introduced in reference by Neves et al., (2021). Unlike other techniques that rely on global surrogate models to explain overall behaviour, LIME concentrates on explaining individual predictions. It creates local surrogate models, which are simpler models that approximate the complex model's decisions but only for specific data subsets (Neves et al., 2021).

4.3. Model Agnostic, Global Surrogate

This interpretable approximation model predicts like a black-box model using the same dataset. Surrogate designs can explain black-box model behavior. The black-box model is solved by creating a new model. These methods are criticized for lacking interpretability since they do not provide an in-depth system overview. Global surrogate models may be developed using intrinsically interpretable models (Henckaerts et al., 2022).

4.4. Model Agnostic, Feature Interaction Based

This approach offers a versatile solution that can be applied to all predictive models, irrespective of their internal makeup, resulting in enhanced generalizability. The key goal is to scrutinize the influence of feature interactions on the predictive outcome. By grasping these interactions, we can deconstruct intricate models into more easily understandable components (Henckaerts et al., 2022).

4.5. Model Agnostic, Visualization-based Approaches

This approach values visual aids to enhance understanding and simplify the complexity of the model. These techniques prove particularly useful in situations where decision-making is intricate. However, the model's outputs are easily understandable without a deep understanding of AI mechanisms. Commonly used techniques include scatter plots, heatmaps, and other visualization methods. Partial Dependent Plots (PDP) and Individual Conditional Expectation (ICE) plots help machine learners examine target-predictor relationships. PDP allows us to explore target-predictor relationships while keeping other factors constant. ICE plots show how the projected result varies with specific characteristics while leaving others constant, graphically representing each occurrence in a single line (Duval, 2019).

4.6. Instance-based Explanations

The methods outlined before involve extensive analysis of systems, which can prove to be both computationally taxing and time-consuming, especially when dealing with large-scale setups. A more effective approach could be to provide explanations for individual cases rather than attempting to explain the entire system. This method doesn't delve into the intricate details of the entire model but instead focuses on its core components, answering questions like "What factors influenced the specific model's outcome?

Feature Selection: Instance-based feature selection maximizes information between chosen attributes and response parameters by selecting attributes per instance. This approach is primarily restricted to impromptu tactics, limiting the explanation generation system. To make it more accessible to non-experts, we must enhance the system's knowledge to produce more detailed explanations (Kong and Chaudhuri, 2021).

Adversarial: Instance-based adversarial explanations in machine learning refer to a specific type of explanation that focuses on how slight alterations to input data (instances) can lead to significant changes in the model's output (Kuppa & Le-Khac, 2021).

Counterfactual: A counterfactual explanation in the context of ML and data analysis provides an alternate scenario that helps to understand model decisions by considering "what if" questions. This concept stems from the idea of counterfactual thinking in philosophy and psychology, where one considers alternative realities based on different decisions or events (Crupi et al., 2022).

4.7. Propagation-Based Methods

This approach is particularly relevant for complex models rooted in DL. These approaches calculate and send information across a network or model. These approaches use forward propagation. These methods find output features using input perturbations. Machine learning and neural networks employ back-propagation techniques. Gradients are calculated through the network to update weights. The technique uses input characteristics to calculate the result. A well-known method, DEEPLift, uses back-propagation to interpret crucial model elements (Wang and Wang 2022).

4.8. Neural Network-Specific Technique

Investigating methods specific to Neural Networks (NN) is crucial as feature learning occurs in the hidden layers. We need tools to interpret NN, and gradient-based methods for interpretation are more computationally efficient than model-agnostic methods. Gradient-Based approaches: Grad-CAM (Class Activation Mapping) visualizes a neural network's gradient to provide a relevance score for every node based on its decision-making significance. Several studies have utilized Grad-CAM and reported various results (Ivanovs et al., 2021).

4.9. Knowledge-Based Techniques

The methods we have investigated so far only use metadata from the dataset to provide explanations. However, we can enhance the quality of explanations by incorporating relevant information, such as expert or domain knowledge. The data can aid in extracting the structure of the machine learning system to offer better insights. Therefore, Knowledge Infused Learning (K-IL) is a method that employs knowledge graphs to develop a thorough understanding of a system. According to a study, knowledge graphs are a way to create explanations (Rajabi and Kafaie, 2022). Quantitative Testing Concept Additive Vectors (QTCAV) is a method that provides explanations that are easily comprehensible to humans. This study concentrates on knowledge-based systems (Walia et al., 2022).

4.10. XAI for Time Series Data

We aim to understand how XAI affects time series data, as reported in the literature. Few studies include explainable artificial intelligence (XAI) methods for time series data.

4.10.1 Theory Driven Approach

This approach utilizes established theoretical frameworks and domain expertise to guide the explanation process (Wang et al., 2019).

4.10.2 Prototype-based Approach

The proposed prototype-based time series data categorization system is deep. An encoder-decoder prototype generated time series dataset descriptions (van der Waa et al., 2021).

4.11. Interactive Tool for Insights

XAI provides valuable insights, and involving humans in knowledge-based systems improves explainability. Interactive machine-learning platforms with human involvement are gaining attention, and real-time system behavior analysis tools are available (Spinner et al., 2019).

4.11.1 Explanatory Interactive Machine Learning (CAIPI)

This is a model-independent tool that facilitates active learning. It uses a question-based approach to learn from users. This tool can provide information on machine-learned knowledge via active learning (Teso and Kersting, 2019).

4.11.2 What-if Tool

Google introduced a tool called What-if, which aims to provide users with a better understanding of ML models. The tool allows users to analyze ML models interactively, but it has limited capability regarding model transparency. Consequently, despite its potential, the tool has significant limitations that must be addressed.

4.11.3 Contextual and Semantic Explanations (CaSE)

CaSE architecture is a tool for explaining text classification. The system uses the local explanation approach and semantic and contextual meanings. It consists of three main parts:

- Explanation algorithm (local)
- Classification algorithm
- Semantic approach (topic modelling in text mining)

(Al-Shedivat et al., 2020)

4.11.4 Human-in-the-Loop

This method highlights how involving a human can simplify a complex computational problem and suggests that involving humans can improve the system's explanations (Ding, 2018).

5. EXISTING WORKS OF XAI WITHIN FINANCE

This section explores research done on XAI in finance. XAI is gaining attention in finance for its ability to offer transparency, interpretability, and accountability in various applications. The following overview shows that XAI has significantly impacted the financial industry (Ding (2018)).

- Credit Scoring and Risk Assessments: Financial institutions use XAI techniques like SHAP and LIME to provide transparent and unbiased justifications for credit decisions while following regulatory guidelines.
- Algorithmic Trading and Portfolio Management: XAI is crucial in algorithmic trading and portfolio management as it helps financial analysts understand complex trading strategies. Techniques like SHAP and LIME can be used to comprehend how algorithmic models make decisions, improving portfolio management and addressing market transparency concerns.
- Fraud Detection and Prevention: XAI helps improve fraud detection and prevention in financial systems. Understanding the characteristics and patterns of fraud is essential for creating effective fraud detection models.
- Regulatory Compliance and Reporting: XAI helps with compliance in finance by offering precise and clear models that meet regulatory standards. It streamlines the reporting process and helps financial institutions to fulfil legal obligations.

Table 1 reviews the current state of research, key findings, methodologies employed, and potential areas for further exploration.

Table 1. Overview of reviewed Papers on XAI in Finance

S. No.	Authors	Key Findings	Limitations	Emerging Trends	Focus / Considerations
1	Khanal et al., (2023)	Identifies the main elements affecting fintech client happiness, specifically at F1 Soft of Nepal. Machine learning models are used.	Limited generalizability to other Fintech firms, dependence on data quality	PCA/XAI integration in customer satisfaction research, developing sophisticated model explanation techniques	Considerations for transparency and interpretability in XAI
2	Mishra et al., (2023)	Development of AI-based cyber security	Limited scalability to	AI is being used more to quickly	Ensuring fairness and

		models, Comparison with current methods	large-scale cyber threats, dependence on data quality	identify and reduce risks. Incorporating XAI techniques into cybersecurity systems.	transparency in AI-based cybersecurit.
3	Mishra et al., (2023)	Designed to manage and evaluate the risks of AI applications in finance.	Limited real-world implementation, potential interpretability trade-offs	Integration of SAFE AI with traditional financial models	Ethical considerations in transparent AI design
4	Hoffman et al., (2023)	Development of a scorecard method for evaluating machine-generated explanations in XAI	Limited generalizability to different XAI models, subjective scoring	Increased focus on XAI evaluation methodologies, integration with model interpretability	Considerations for bias and fairness in evaluation metrics
5	Hjelkrem et al., (2023)	SHAP has been shown to improve the transparency and interpretability of deep learning models for credit scoring by providing insights into their decision-making process.	Interpretability challenges in very complex deep learning architectures	Integration of SHAP with deep learning models in credit scoring	Considerations for transparency and regulatory compliance
6	Wand et al., (2023)	The correlation matrix of each state is mainly influenced by a few sector correlations, showing the impact of specific industrial sectors on market states.	High	Investigate sectoral dynamics in dynamic market states	Rigorous validation against diverse market scenarios
7	Torky et al., (2023)	The model showed it can automatically detect and understand factors that cause financial crises, offering transparency and interpretability.	Sensitivity to hyperparameters in optimization, interpretability challenges	Integration of XAI with optimization techniques in financial crisis analysis	Considerations for transparency, accountability, and regulatory compliance
8	Zhou et al., (2023)	The paper discussed the issue of black-box problem and lack of trust in AI predictions for fraud detection among	User-specific interpretability preferences, dependency on user feedback	Integration of user preferences in XAI for fraud detection	Considerations for user privacy and trust in AI decisions

		experts.			
9	Çelik et al., (2023)	The study shows how XAI can improve transparency and trust in predicting financial time series, especially in stocks.	Interpretability challenges in complex time series models	Integration of XAI for transparency and interpretability in time series models	Considerations for fairness, transparency, and ethical use of AI
10	Sai et al., (2023)	A new model was created to classify financial transactions as fraudulent or legitimate using various machine learning and deep neural network algorithms.	Interpretability challenges in deep neural networks and potential limitations in handling complex fraud patterns.	Using explainable AI to detect financial transaction fraud for transparency and accountability.	Considerations for fairness, transparency, and ethical use of AI
11	Predić et al., (2023)	The study proposes a machine learning system for predicting business purchases using neural networks, including XAI and LSTM networks.	Interpretability challenges in LSTM models and potential biases in training data.	Integration of XAI and LSTM in business purchase prediction for transparency and accountability	Considerations for fairness, transparency, and ethical use of AI
12	Biswas et al., (2023)	The study presents a fraud detection framework that uses SHAP to make machine learning models in credit card fraud detection more transparent and interpretable.	Challenges in handling imbalanced datasets and potential interpretability issues in highly complex fraud patterns.	Integration of SHAP in credit card fraud detection for transparency and accountability	Considerations for fairness, transparency, and ethical use of AI
13	Zhang et al., (2023)	The study shows that the DS-XGBoost model is effective in predicting financial risk.	Difficulties in understanding ensemble models and limitations in capturing intricate risk patterns.	Integration of DS-XGBoost in financial risk early warning for transparency and accountability	Considerations for fairness, transparency, and ethical use of AI
14	Zheng et al., (2023)	It creates a risk management tool for micro and small enterprises using user portrait theory and common objective indicators.	The study's findings may be influenced by the availability and quality of data specific to micro and small	Application of explainable ML in assessing default risk for micro and small enterprises	Considerations for fairness, transparency, and ethical use of ML in financial risk assessments

			enterprises.		
15	Liu et al., (2023)	The study uses four tree-based gradient boosting models to predict financial distress: extreme gradient boosting, random forest, decision tree, and logistic regression.	The interpretation may be influenced by the complexity of financial distress factors and the specific characteristics of the dataset used.	Interpretation of prediction results for financial distress using explainable ML in tree-based models	Considerations for fairness, transparency, and ethical use of explainable ML in financial decision-making
16	Lu &Calabrese (2023)	The paper examines how SMEs in the UK can access financing and emphasizes the need for fair financing to help these businesses grow.	The fairness measurement may be influenced by the specific criteria and data used in defining cohorts, and the study's findings may be context-dependent.	Application of the Cohort Shapley value for fairness measurement in SME financing	Considerations for fairness, transparency, and ethical use of fairness measurement techniques in SME financing
17	Freeborough et al., (2022)	Enhanced interpretability in RNNs for financial time series	Limited interpretability for extremely complex market behavior	Integration of attention mechanisms in RNNs for financial analytics	Fairness considerations
18	Li, et al., (2022)	Enhanced interpretability in financial risk detection using explainable case-based reasoning	Limited scalability in large datasets, dependence on case quality	Integration of case-based reasoning with XAI, focus on feature importance analysis	Considerations for transparency in case-based reasoning
19	De Lange et al., (2022)	Implementation of LIME and SHAP in enhancing interpretability in credit assessment in banks	Interpretability may vary based on the specific credit models	Integration of XAI in credit scoring models in banking	Considerations for fairness and transparency in credit assessment
20	Fritz-Morgenthal et al., (2022)	It emphasizes the need for clear understanding and transparency in models, data, and decision-making to effectively handle any remaining risks.	Challenges in operationalizing Responsible AI principles	Fusion of XAI with Responsible AI principles in risk assessment	Considerations for fairness, transparency, and ethical risk management

21	Owens et al., (2022)	XAI improves transparency and trust in the insurance industry by helping us understand decision-making in black-box machine learning models.	Interpretability challenges in complex insurance algorithms	Integration of XAI in underwriting and claims processing for transparency	Considerations for fairness, transparency, and ethical use of AI
22	Zhang et al., (2022)	Proposed a new XAI approach for financial distress prediction, addressing the need for accurate and explainable FDP models	Interpretability challenges in complex financial models	Adoption of XAI in financial distress prediction models	Considerations for fairness, transparency, and ethical use of AI
23	Zhang et al., (2022)	The paper discusses methods to improve the interpretability of AI artifacts in accounting and auditing. This is done to increase transparency and trust in AI applications used in auditing.	Interpretability challenges in complex auditing algorithms	Integration of XAI in risk assessment, fraud detection, and compliance checks	Considerations for fairness, transparency, and ethical use of AI
24	Lin et al., (2022)	Enhancement of model interpretability using Group SHAP in financial fraud detection	Challenges in handling large-scale, diverse fraud scenarios	Adoption of Group SHAP in financial fraud detection for group-level insights	Considerations for fairness, transparency, and ethical use of AI
25	Kuiper et al., (2022)	XAI is considered a way to increase transparency, fairness, and accountability in AI systems, particularly in finance.	The study's findings may be influenced by the sample size and regional variations in regulatory frameworks.	Integration of explainable AI in financial decision-making for enhanced transparency	Considerations for fairness, transparency, and ethical use of XAI in financial institutions
26	Lachuer et al., (2022)	In a bullish market, research shows that CSR has a negative relationship with overall financial market performance, but it does improve the financial performance of the most sustainable companies.	Challenges in capturing the full complexity of CSR and financial performance relationships, as well as potential variations across industries.	Integration of explainable AI in modeling CSR and financial performance for transparent and accountable decision-making	Considerations for fairness, transparency, and ethical use of AI in financial decision-making

27	Gramespacher et al., (2021)	ML techniques can be customized for credit evaluation requirements. Prioritizing an economic goal over accuracy can help increase profits. ML models that are simple and explainable can optimize the return target function without affecting accuracy or fidelity.	Sensitivity to market variations, reliance on historical data	Increasing use of XAI for optimizing economic target functions in loan portfolio management. Integration of XAI techniques into existing loan portfolio management systems.	Considerations for fair lending practices
28	Gramegna et al., (2021)	The study evaluates how well SHAP and LIME explain credit risk predictions in machine learning models. Both SHAP and LIME offer valuable insights into machine learning models for credit risk decision-making.	Dependency on model complexity and interpretability, dataset-specific results	Integration of SHAP and LIME in ensemble credit risk models	Considerations for fairness and bias in credit risk assessments
29	Moscato et al., (2021)	It compares various credit risk scoring models to predict loan repayment in a P2P platform.	Sensitivity to feature engineering, model interpretability trade-offs	Exploration of deep learning architectures in credit scoring models	Considerations for fairness and transparency in credit scoring
30	Dastile et al., (2021)	The study demonstrates the effectiveness of the proposed model in making accurate and explainable predictions for credit scoring	Interpretability challenges in complex deep learning architectures	Integration of SHAP and LIME in deep learning models for credit scoring	Considerations for fairness and transparency in credit assessment
31	Gite et al., (2021)	The study proposed using LSTM and Explainable AI (XAI) to predict future stock prices using visual representations.	Sensitivity to specific sentiment analysis algorithms and models	Integration of SHAP with sentiment analysis for stock prediction	Considerations for transparency and ethical use of sentiment data
32	Cirqueira et al., (2021)	The study uses a design science research methodology and an Information Systems theoretical lens to	Challenges in balancing interpretability and usability for diverse users	Integration of user-centric design in explainable fraud detection models	Considerations for user privacy and trust in AI explanations

		create and assess design principles that connect fraud expert tasks with explanation methods for XAI decision support in fraud detection.			
33	Preub et al., (2021)	Explored risks and limits connected with the application of AI models in financial modeling and its application in practice	High	Explore the impact of AI governance on risk management practices	Comprehensive literature review
34	Bussmann et al., (2021)	Integration of XAI techniques for interpretability in credit risk models	Interpretability may vary based on specific credit models	Adoption of XAI in credit scoring models for transparency and accountability	Considerations for fairness, transparency, and ethical risk management
35	Bussmann et al., (2020)	Implementation of explainable AI techniques (SHAP, LIME) in enhancing interpretability in Fintech risk management	Limited applicability to complex, proprietary models	Integration of model-agnostic methods in risk assessment	Considerations for transparency and regulatory compliance
36	Burgt et al., (2020)	Implemented XAI in banking enhancing transparency and trust among all stakeholders, including banks, regulators, and consumers.	Interpretability challenges in complex banking algorithms	Integration of XAI in risk assessment, fraud detection, and customer service	Considerations for fairness, transparency, and ethical use of AI
37	Demajo et al., (2020)	The paper presents a credit scoring model that is both accurate and interpretable, addressing the need for model interpretability due to regulations such as the GDPR and the Equal Credit Opportunity Act ECOA	Interpretability challenges in complex credit models	Integration of XAI in credit scoring for transparency and accountability	Considerations for fairness, transparency, and ethical use of AI
38	Ariza-Garzón et al., (2020)	The study highlights the importance of explainability in machine learning models for granting scoring in peer-to-	Interpretability challenges in complex lending models	Integration of XAI in lending models for transparency and accountability	Considerations for fairness, transparency, and ethical use of AI

		peer lending			
39	Rios et al., (2020)	The study demonstrates the effectiveness of LRP in providing intuitive, human-readable heat maps of input images, outpering traditional explainability concepts of LIME and SHAP for explainability	The approach may face challenges in handling highly complex and large-scale structured datasets.	Integration of Layer-Wise Relevance Propagation in structured data analysis for enhanced interpretability	Considerations for fairness, transparency, and ethical use of AI
40	Rao et al., (2020)	xFraud effectively and efficiently predicts the legitimacy of incoming transactions, advancing previous work by building upon a heterogeneous GNN to tackle transaction fraud detection.	The model's performance may be influenced by the quality and representativeness of the training data. Interpretability might be challenging for highly complex fraud patterns.	Integration of xFraud in fraud detection for transparency and accountability	Considerations for fairness, transparency, and ethical use of AI
41	Giudici et al., (2020)	Network-based credit risk models use financial network linkages, according to the report. Consideration of network structure improves risk assessment accuracy.	Complexity of the financial network affects model effectiveness. Quality and completeness of network data affect model performance.	Integration of network-based credit risk models for enhanced risk assessment	Considerations for fairness, transparency, and ethical use of network-based credit risk models
42	Pintelas et al., (2020)	A grey-box ensemble model that combines black-box accuracy with white-box interpretability is presented in the study. This hybrid technique balances model complexity and interpretability.	The model's performance may be influenced by the choice of base black-box models and the quality of interpretability methods.	Integration of grey-box ensemble models for balanced accuracy and interpretability	Considerations for fairness, transparency, and ethical use of grey-box models

6. GAP AND LIMITATIONS IN CURRENT RESEARCH

The finance industry has shown considerable interest in XAI because of various compliance or customer-centric drivers. The critical research gap encountered can be grouped as follows:

- Overcoming Challenges with Effective Solutions: The finance sector faces several challenges when implementing XAI. These challenges include the need for a widely accepted definition of explainability, balancing accuracy and explainability, and considering computational costs. Tackling these obstacles requires the collaboration of researchers, policymakers, and industry experts to establish definitive standards for explainability, devise privacy-preserving methods, and improve the effectiveness of explanations.
- Areas of application: While XAI has already proven helpful in financial domains such as risk management, portfolio optimization, stock market applications, and green supply chain finance, further research is still needed to explore its potential application in other areas of finance, such as credit scoring, fraud detection, and regulatory compliance.
- Proposed Research Agenda: It is imperative to develop a forward-thinking research agenda that elucidates the various applications of XAI within finance. This agenda should specifically address the existing challenges and identify potential areas for future research.
- Regulatory Compliance: It's important to focus on developing XAI methods that adhere to regulatory protocols and build trust in AI systems across the finance sector. Banking regulators are paying closer attention to XAI and require AI results and procedures to be understandable to bank personnel. We can cultivate trust and ensure regulatory compliance by advancing XAI research in this direction.

7. FUTURE DIRECTIONS

- The potential for XAI to drive progress in the financial industry is substantial, with critical avenues that will influence its advancement. Moving from the theoretical background to practical implementation is recommended when exploring a topic's real-world usage and application. Thus, the approach can notably enhance customer satisfaction metrics and enable the tailoring of models (or explanations) to deliver superior user experiences. Recently, there has been a surge in the creation of verifiable XAI solutions that provide visual explanations of the model. The evolution of Generative AI solutions imposes additional complexity on the explainability of the AI solutions (Bussmann et al., 2020; Bussmann et al., 2021; Burgt 2020; Demajo et al., 2020, Taeihagh 2021).
- In a previous study (Lee et al., 2018), a thorough examination of explainable recommendations was conducted, and the authors also explored potential avenues for advancing the field. Regarding the methodology adopted, it has been proposed that additional research is necessary to enhance the explainability of deep learning models used for recommendations.
- Experts recommend the use of knowledge-enhanced explainable recommendation systems that incorporate domain knowledge. One potential avenue for exploration is to combine graph embedding learning with recommendation models. Additionally, it is advised to incorporate heterogeneous information for explainability, such as utilizing multi-modal explanations, leveraging transfer learning across diverse information sources, and analyzing the impact of different information modalities on user acceptance of explanations.
- To provide efficient recommendations, it is essential to consider the context in which they are given and ensure that clear explanations are provided. A more comprehensive understanding of the recommendations can be achieved by combining a variety of explanations. By incorporating both symbolic reasoning and machine learning techniques, the quality of recommendations and explanations can be significantly improved, moving beyond collaborative filtering into collaborative reasoning. However, further research is necessary to enable machines to explain using natural language. Users will increasingly seek explanations for illogical recommendations as conversational recommendations continue to be refined through intelligent agent devices. Therefore, the underlying reasons behind recommendations must be thoroughly explored to maximize the system's effectiveness, transparency, and reliability. Hence, this is a continuous improvement activity.

- In today's fast-paced financial landscape, exploring effective techniques that can adapt to ever-evolving data structures is imperative. Striking a balance between data privacy and interpretability remains a top priority, and establishing robust frameworks that encourage responsible AI usage is critical. Looking at hybrid models that combine XAI with optimization methods holds great promise for achieving superior results. Robust testing of XAI methods across various economic scenarios, including stress testing, will provide valuable insights into model performance within dynamic financial environments. Also, the real-time adaptable methods are significantly important. A promising approach to enhancing interpretability is combining SHAP and LIME in hybrid models. Ongoing research investigates how these two methods can collaborate to explain intricate financial models.
- Implementing systems that facilitate ongoing improvement based on user feedback is crucial for increasing transparency and establishing credibility. Establishing frameworks for ethical implementation and compliance of AI is imperative to effectively address ethical dilemmas within the financial industry.

Adopting these guidelines can help the financial industry promote XAI's growth and widespread use, ensuring its longevity and efficacy in a dynamic environment. As referenced by the authors Gramespacher and Posth (2021), it is crucial to prioritize the development of user-friendly and intelligent interface modalities that align with users' objectives and requirements. For example, the system can gather user feedback to evaluate the quality of explanations provided. This feedback may include requests for more details, identification of redundant explanations, or suggestions for alternative explanations. Engaging in such communication can help improve future explanations.

8. CONCLUSION

XAI is a robust framework that promotes transparency and understanding in the decision-making process of traditional AI models, including DL and ML. In this systematic literature review, we analyzed over 70 primary studies out of 2500 related XAI articles from 2018 to 2024 to gain a structured understanding of the state-of-the-art XAI research in IS. Our exploration began with a detailed introduction emphasizing XAI's overarching significance in Information Technology Systems (ITS) and its relevance to the Finance sector. Through thoroughly examining existing works in this field, we uncovered vital gaps and limitations that must be addressed in future research. Our review culminates in a forward-looking exploration of the future directions XAI is poised to take. By weaving together these intricately connected sections, we present a comprehensive tapestry of XAI, providing insights, frameworks, and perspectives for the continued evolution of explainability in artificial intelligence.

REFERENCES

[1] Jalayer, M. Kahani, A. Beheshti, A. Pourmasoumi, and H. R. Motahari-Nezhad. Attention mechanism in predictive business process monitoring. In 24th IEEE International Enterprise Distributed Object Computing Conference, EDOC 2020, Eindhoven, The Netherlands, October 5-8, 2020, pages 181–186. IEEE, 2020.

[2] Al-Shedivat, M., Dubey, A., & Xing, E. (2020). Contextual explanation networks. The Journal of Machine Learning Research, 21(1), 7950-7993.

[3] Alikhademi, K., Richardson, B., Drobina, E., & Gilbert, J. E. (2021). Can explainable AI explain unfairness? A framework for evaluating explainable AI. arXiv preprint arXiv:2106.07483.

[4] Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q Consortium. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary

[5] Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision

[6] Appelganc, K., Rieger, T., Roesler, E., & Manzey, D. (2022). How much reliability is enough? A context-specific view on human interaction with (artificial) agents from different perspectives. Journal of Cognitive Engineering and Decision Making, 16(4), 207-221.

[7] Ariza-Garzón, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M. J. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. Ieee Access, 8, 64873-64890.

[8] Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: transforming the practice of medicine. Future healthcare journal, 8(2), e188.

[9] Biecek, P., Chlebus, M., Gajda, J., Gosiewska, A., Kozak, A., Ogonowski, D., ... & Wojewnik, P. (2021). Enabling machine learning algorithms for credit scoring--explainable artificial intelligence

(XAI) methods for clear understanding complex predictive models. arXiv preprint arXiv:2104.06735.

[10] Biswas, J., Mridha, A. A., Hossain, M. S., Trisha, A. S., Ahmed, M. S., & Hossain, M. I. (2023, July). Interpretable Credit Card Fraud Detection Using Machine Learning Leveraging SHAP. In 2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT) (pp. 1206-1211). IEEE.

[11] 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.

[12] Burgt, J. V. D. (2020). Explainable AI in banking. Journal of Digital Banking, 4(4), 344-350.

[13] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. Frontiers in Artificial Intelligence, 3, 26.

[14] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57, 203-216.

[15] Butz, R., Schulz, R., Hommersom, A., & van Eekelen, M. (2022). Investigating the understandability of XAI methods for enhanced user experience: When Bayesian network users became detectives. Artificial Intelligence in Medicine, 134, 102438.

[16] Molnar. Interpretable Machine Learning. Lulu.com, 2021.

[17] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832.

[18] Çelik, T. B., İcan, Ö., & Bulut, E. (2023). Extending machine learning prediction capabilities by explainable AI in financial time series prediction. Applied Soft Computing, 132, 109876.

[19] Cirqueira, D., Helfert, M., & Bezbradica, M. (2021, July). Towards design principles for user-centric explainable AI in fraud detection. In International Conference on Human-Computer Interaction (pp. 21-40). Cham: Springer International Publishing.

[20] Crupi, R., Castelnovo, A., Regoli, D., & San Miguel Gonzalez, B. (2022). Counterfactual explanations as interventions in latent space. Data Mining and Knowledge Discovery, 1-37.

[21] Dastile, X., & Celik, T. (2021). Making deep learning-based predictions for credit scoring explainable. IEEE Access, 9, 50426-50440.

[22] Dazeley, R., Vamplew, P., & Cruz, F. (2023). Explainable reinforcement learning for broad-xai: a conceptual framework and survey. Neural Computing and Applications, 1-24.

[23] De Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for Credit Assessment in Banks. Journal of Risk and Financial Management, 15(12), 556.

[24] Demajo, L. M., Vella, V., & Dingli, A. (2020). Explainable ai for interpretable credit scoring. arXiv preprint arXiv:2012.03749.

[25] Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y. K., Mäntymäki, M., & Pappas, I. O. (2023). Artificial intelligence (AI) and information systems: perspectives to responsible AI. Information Systems Frontiers, 25(1), 1-7.

[26] Ding, L. (2018). Human knowledge in constructing AI systems—Neural logic networks approach towards an explainable AI. Procedia computer science, 126, 1561-1570.

[27] Duval, A. (2019). Explainable artificial intelligence (XAI). MA4K9 Scholarly Report, Mathematics Institute, The University of Warwick, 1-53.

[28] Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 57, 101994.

[29] eber, L., Lapuschkin, S., Binder, A., & Samek, W. (2023). Beyond explaining: Opportunities and challenges of XAI-based model improvement. Information Fusion, 92, 154-176.

[30] Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable ai: Towards a reflective sociotechnical approach. In HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22 (pp. 449-466). Springer International Publishing.

[31] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608, 2 2017.

[32] Freeborough, W., & van Zyl, T. (2022). Investigating explainability methods in recurrent neural network architectures for financial time series data. Applied Sciences, 12(3), 1427.

[33] Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. (2022). Financial risk management and explainable, trustworthy, responsible AI. Frontiers in artificial intelligence, 5, 779799.

[34] Gill, S. S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., ... & Uhlig, S. (2022). AI for next generation computing: Emerging trends and future directions. Internet of Things, 19, 100514.

[35] Gite, S., Khatavkar, H., Kotecha, K., Srivastava, S., Maheshwari, P., & Pandey, N. (2021). Explainable stock prices prediction from financial news articles using sentiment analysis. PeerJ Computer Science, 7, e340.

[36] Giudici, P., & Raffinetti, E. (2023). SAFE artificial intelligence in finance. Finance Research Letters, 104088.

[37] Giudici, P., Hadji-Misheva, B., & Spelta, A. (2020). Network based credit risk models. Quality Engineering, 32(2), 199-211.

[38] Gramegna, A., & Giudici, P. (2021). SHAP and LIME: an evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence, 4, 752558.

[39] Gramespacher, T., & Posth, J. A. (2021). Employing explainable AI to optimize the return target function of a loan portfolio. Frontiers in Artificial Intelligence, 4, 693022.

[40] Guidotti, R., Monreale, A., Pedreschi, D., & Giannotti, F. (2021). Principles of explainable artificial intelligence. Explainable AI Within the Digital Transformation and Cyber Physical Systems: XAI Methods and Applications, 9-31

[41] Hasan, Z., Vaz, D., Athota, V. S., Désiré, S. S. M., & Pereira, V. (2022). Can Artificial Intelligence (AI) Manage Behavioural Biases Among Financial Planners?. Journal of Global Information Management (JGIM), 31(2), 1-18.

[42] Henckaerts, R., Antonio, K., & Côté, M. P. (2022). When stakes are high: Balancing accuracy and transparency with Model-Agnostic Interpretable Data-driven suRRogates. Expert Systems with Applications, 202, 117230.

[43] Hjelkrem, L. O., & Lange, P. E. D. (2023). Explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data. Journal of Risk and Financial Management, 16(4), 221.

[44] Hoepner, A. G., McMillan, D., Vivian, A., & Wese Simen, C. (2021). Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective. The European Journal of Finance, 27(1-2), 1-7.

[45] Hoffman, R. R., Jalaeian, M., Tate, C., Klein, G., & Mueller, S. T. (2023). Evaluating machine-generated explanations: a "Scorecard" method for XAI measurement science. Frontiers in Computer Science, 5, 1114806.

[46] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

[47] Hu, B., Tunison, P., Vasu, B., Menon, N., Collins, R., & Hoogs, A. (2021). XAITK: The explainable AI toolkit. Applied AI Letters, 2(4), e40.

[48] Ivanovs, M., Kadikis, R., & Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. Pattern Recognition Letters, 150, 228-234.

[49] Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. npj Digital Medicine, 6(1), 6.

[50] Khanal, M., Khadka, S. R., Subedi, H., Chaulagain, I. P., Regmi, L. N., & Bhandari, M. (2023). Explaining the Factors Affecting Customer Satisfaction at the Fintech Firm F1 Soft by Using PCA and XAI. FinTech, 2(1), 70-84.

[51] Kong, Z., & Chaudhuri, K. (2021). Understanding instance-based interpretability of variational auto-encoders. Advances in Neural Information Processing Systems, 34, 2400-2412.

[52] Kuiper, O., van den Berg, M., van der Burgt, J., & Leijnen, S. (2022). Exploring explainable AI in the financial sector: Perspectives of banks and supervisory authorities. In Artificial Intelligence and Machine Learning: 33rd Benelux Conference on Artificial Intelligence, BNAIC/Benelearn 2021, Esch-sur-Alzette, Luxembourg, November 10–12, 2021, Revised Selected Papers 33 (pp. 105-119). Springer International Publishing.

[53] Kuppa, A., & Le-Khac, N. A. (2021). Adversarial xai methods in cybersecurity. IEEE transactions on information forensics and security, 16, 4924-4938.

[54] Lachuer, J., & Jabeur, S. B. (2022). Explainable artificial intelligence modeling for corporate social responsibility and financial performance. Journal of Asset Management, 23(7), 619-630.

[55] Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... & 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.

[56] Larsson, S. (2019). The socio-legal relevance of artificial intelligence. Droit et société, 103(3), 573-593.

[57] Lee, K., Lee, H., Lee, H., Yoon, Y., Lee, E., & Rhee, W. (2018). Assuring explainability on demand response targeting via credit scoring. Energy, 161, 670-679.

[58] Li, W., Paraschiv, F., & Sermpinis, G. (2022). A data-driven explainable case-based reasoning approach for financial risk detection. Quantitative Finance, 22(12), 2257-2274.

[59] Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., ... & Dou, D. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. Knowledge and Information Systems, 64(12), 3197-3234.

[60] Lin, K., & Gao, Y. (2022). Model interpretability of financial fraud detection by group SHAP. Expert Systems with Applications, 210, 118354.

[61] Liu, J., Li, C., Ouyang, P., Liu, J., & Wu, C. (2023). Interpreting the prediction results of the tree-based gradient boosting models for financial distress prediction with an explainable machine learning approach. Journal of Forecasting, 42(5), 1112-1137.

[62] Lopes, P., Silva, E., Braga, C., Oliveira, T., & Rosado, L. (2022). XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. Applied Sciences, 12(19), 9423.

[63] Lu, X., & Calabrese, R. (2023). The Cohort Shapley value to measure fairness in financing small and medium enterprises in the UK. Finance Research Letters, 58, 104542.

[64] Lundberg, H., Mowla, N. I., Abedin, S. F., Thar, K., Mahmood, A., Gidlund, M., & Raza, S. (2022). Experimental Analysis of Trustworthy In-Vehicle Intrusion Detection System Using eXplainable Artificial Intelligence (XAI). IEEE Access, 10, 102831-102841.

[65] Mill, E. R., Garn, W., Ryman-Tubb, N. F., & Turner, C. (2023). Opportunities in Real Time Fraud Detection: An Explainable Artificial Intelligence (XAI) Research Agenda. International Journal of Advanced Computer Science and Applications, 14(5), 1172-1186.

[66] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 267, 1-38.

[67] Mishra, S. (2023). Exploring the Impact of AI-Based Cyber Security Financial Sector Management. Applied Sciences, 13(10), 5875.

[68] Mohammed, K., & Shehu, A. (2023). A REVIEW OF ARTIFICIAL INTELLIGENCE (AI) CHALLENGES AND FUTURE PROSPECTS OF EXPLAINABLE AI IN MAJOR FIELDS: A CASE STUDY OF NIGERIA. Open Journal of Physical Science (ISSN: 2734-2123), 4(1), 1-18.

[69] Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. Expert Systems with Applications, 165, 113986.

[70] Munz, P., Hennick, M., & Stewart, J. (2023). Maximizing AI reliability through anticipatory thinking and model risk audits. AI Magazine.

[71] Neves, I., Folgado, D., Santos, S., Barandas, M., Campagner, A., Ronzio, L., ... & Gamboa, H. (2021). Interpretable heartbeat classification using local model-agnostic explanations on ECGs. Computers in Biology and Medicine, 133, 104393.

[72] Obster, F., Brand, J., Ciolacu, M., & Humpe, A. (2023). Improving Boosted Generalized Additive Models with Random Forests: A Zoo Visitor Case Study for Smart Tourism. Procedia Computer Science, 217, 187-197.

[73] Owens, E., Sheehan, B., Mullins, M., Cunneen, M., Ressel, J., & Castignani, G. (2022). Explainable artificial intelligence (xai) in insurance. Risks, 10(12), 230.

[74] P. Hall and N. Gill. Explain Your AI H2O Driverless AI, volume II. O'Reilly Media, second edition, 2019.

[75] Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). Minds and Machines, 29(3), 441-459.

[76] Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., ... & He, Z. (2020). Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. Journal of the American Medical Informatics Association, 27(7), 1173-1185.

[77] Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2020). Four principles of explainable artificial intelligence. Gaithersburg, Maryland, 18.

[78] Pintelas, E., Livieris, I. E., & Pintelas, P. (2020). A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. Algorithms, 13(1), 17.

[79] Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. Artificial Intelligence in Medicine, 107, 101901.

[80] Predić, B., Ćirić, M., & Stoimenov, L. (2023). Business Purchase Prediction Based on XAI and LSTM Neural Networks. Electronics, 12(21), 4510.

[81] Preuß, B. (2021). Contemporary Approaches for AI Governance in Financial Institutions. Available at SSRN 3773581.

[82] Rai, A. (2020). Explainable AI: From black box to glass box. Journal of the Academy of Marketing Science, 48, 137-141.

[83] Rajabi, E., & Kafaie, S. (2022). Knowledge graphs and explainable ai in healthcare. Information, 13(10), 459.

[84] Rao, S. X., Zhang, S., Han, Z., Zhang, Z., Min, W., Chen, Z., ... & Zhang, C. (2020). xFraud: explainable fraud transaction detection. arXiv preprint arXiv:2011.12193.

[85] Rios, A., Gala, V., & Mckeever, S. (2020). Explaining deep learning models for structured data using layer-wise relevance propagation. arXiv preprint arXiv:2011.13429.

[86] S R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed. Explainable Artificial Intelligence Approaches: A Survey. arXiv preprint arXiv:2101.09429, 1 2021.

[87] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed. Explainable Artificial Intelligence Approaches: A Survey. arXiv preprint arXiv:2101.09429, 1 2021.

[88] Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems, 263, 110273.

[89] Sai, C. V., Das, D., Elmitwally, N., Elezaj, O., & Islam, M. B. Explainable Ai-Driven Financial Transaction Fraud Detection Using Machine Learning and Deep Neural Networks. Available at SSRN 4439980.

[90] Sanneman, L., & Shah, J. A. (2022). The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. International Journal of Human–Computer Interaction, 38(18-20), 1772-1788.

[91] Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). Glocalx-from local to global explanations of black box ai models. Artificial Intelligence, 294, 103457.

[92] Spinner, T., Schlegel, U., Schäfer, H., & El-Assady, M. (2019). explAIner: A visual analytics framework for interactive and explainable machine learning. IEEE transactions on visualization and computer graphics, 26(1), 1064-1074.

[93] Taeihagh, A. (2021). Governance of artificial intelligence. Policy and society, 40(2), 137-157.

[94] Tang, X., Li, X., Ding, Y., Song, M., & Bu, Y. (2020). The pace of artificial intelligence innovations: Speed, talent, and trial-and-error. Journal of Informetrics, 14(4), 101094.

[95] Teso, S., & Kersting, K. (2019, January). Explanatory interactive machine learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 239-245).

[96] The Royal Society (2019). Explainable AI: The Basics Policy Briefing. Issued: November 2019 DES6051 ISBN: 978-78252-433. The Royal Society. Available online at:https://royalsociety.org/topics-policy/projects/explainable-ai/?utm_source=report&utm_medium=print&utm_campaign=ai-interpretability/

[97] Torky, M., Gad, I., & Hassanien, A. E. (2023). Explainable AI Model for Recognizing Financial Crisis Roots Based on Pigeon Optimization and Gradient Boosting Model. International Journal of Computational Intelligence Systems, 16(1), 50.

[98] 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.

[99] Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., ... & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature communications, 11(1), 1-10.

[100] Walia, S., Kumar, K., Agarwal, S., & Kim, H. (2022). Using xai for deep learning-based image manipulation detection with shapley additive explanation. Symmetry, 14(8), 1611.

[101] Wand, T., Heßler, M., & Kamps, O. (2023). Identifying dominant industrial sectors in market states of the S&P 500 financial data. Journal of Statistical Mechanics: Theory and Experiment, 2023(4), 043402.

[102] Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019, May). Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-15).

[103] Wang, Y., & Wang, X. (2022). “Why Not Other Classes?”: Towards Class-Contrastive Back-Propagation Explanations. Advances in Neural Information Processing Systems, 35, 9085-9097.

[104] X.-H. Li, C. C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, C. Wang, Y. Gao, S. Zhang, X. Xue, and L. Chen. A Survey of Data-driven and Knowledge-aware eXplainable AI. IEEE Transactions on Knowledge and Data Engineering, pages 1–1, 3 2020.

[105] Yang, S. C. H., Folke, T., & Shafto, P. (2021). Abstraction, validation, and generalization for explainable artificial intelligence. Applied AI Letters, 2(4), e37.

[106] Yigitcanlar, T., Corchado, J. M., Mehmood, R., Li, R. Y. M., Mossberger, K., & Desouza, K. (2021). Responsible urban innovation with local government artificial intelligence (AI): A conceptual framework and research agenda. Journal of Open Innovation: Technology, Market, and Complexity, 7(1), 71.

[107] Yu, L., Li, Y., & Fan, F. (2023). Employees’ Appraisals and Trust of Artificial Intelligences’ Transparency and Opacity. Behavioral Sciences, 13(4), 344.

[108] Zablocki, É., Ben-Younes, H., Pérez, P., & Cord, M. (2022). Explainability of deep vision-based autonomous driving systems: Review and challenges. International Journal of Computer Vision, 130(10), 2425-2452.

[109] Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. Philosophy & technology, 34(2), 265-288.

[110] Zhang, C. A., Cho, S., & Vasarhelyi, M. (2022). Explainable artificial intelligence (xai) in auditing. International Journal of Accounting Information Systems, 46, 100572.

[111] Zhang, T., Zhu, W., Wu, Y., Wu, Z., Zhang, C., & Hu, X. (2023). An explainable financial risk early warning model based on the DS-XGBoost model. Finance Research Letters, 104045.

[112] Zhang, Z., Wu, C., Qu, S., & Chen, X. (2022). An explainable artificial intelligence approach for financial distress prediction. Information Processing & Management, 59(4), 102988.

[113] Zheng, C., Weng, F., Luo, Y., & Yang, C. (2023). Micro and small enterprises default risk portrait: evidence from explainable machine learning method. Journal of Ambient Intelligence and Humanized Computing, 1-11.

[114] Zhou, Y., Li, H., Xiao, Z., & Qiu, J. (2023). A user-centered explainable artificial intelligence approach for financial fraud detection. Finance Research Letters, 58, 104309.

[115] Zolanvari, M., Yang, Z., Khan, K., Jain, R., & Meskin, N. (2021). Trust xai: Model-agnostic explanations for ai with a case study on iiot security. IEEE internet of things journal.

[116] Zopounidis, C. (1999). Multicriteria decision aid in financial management. European Journal of Operational Research, 119(2), 404-415